IMBALANCED MULTICLASS DATA CLASSIFICATION USING ANT COLONY OPTIMIZATION ALGORITHM

Mrs. S. Lavanya[1], Dr. S. Palaniswami[2].
[1]Teaching Fellow, Department of Computer Science and Engineering,
Anna University Regional Campus, Coimbatore, Tamilnadu, India.
[2]Principal, Government College of Engineering, Bodinayakanur, Tamilnadu, India.

ABSTRACT

Class imbalance problems have drawn increasing interest lately because of its classification trouble caused by imbalanced class deliveries and poor prediction performance for minority class. This problem is particularly common in preparation and can be detected in various disciplines including fraud detection, anomaly detection, oil spillage detection, medical diagnosis, facial recognition. Many ensemble procedures only concentrated on two-class imbalance problems. There are numerous unresolved concerns in multiclass imbalanced problems. Using One-vs-One binarization technique for disintegrating the original multiclass data-set into binary classification problems. Then, each and every time these binary sub problems is imbalanced, applying undersampling step, using the ACOsampling algorithm in order to rebalance the data. Only taking out high frequency dataset from majority samples and mingling those with all minority samples to build the final balanced training set. Here taken different multiclass data such as thyroid, lung cancer and contraceptive. Finally evaluate the performance of each method on four benchmark skewed DNA microarray dataset by support vector machine (SVM) Classifier. It gives better accuracy, precision, f-measure and g-mean when comparing with RUS (Random Under Sampling) method.

Keywords: Multi-class classification, Binarization, SVM, Imbalance data, One-vs-One, Undersampling, Ant Colony Optimization.

1. INTRODUCTION

In the past decade, DNA micro array data has been one of the essential molecular biology technologies in post-genomic era. Using this, biologists, researcher and medical experts are permitted to notice the activity of thousands of genes in a cell concurrently. For now, DNA micro array dataset has been extensively applied to forecast gene expression and functions, examine gene regulatory activities, offer valuable evidence for drug discovery and also used classify for cancer and mining new subtypes of a specific type like thyroid, tumor etc [1]. Between these useful applications, the cancer classification has been concerned more attentions. For all that, it is well-known the micro array data normally has some appropriate structures, like high dimension, small sample set, huge noise and commonly, imbalanced class supplies. Skewed class distributions will underestimate enormously the prediction performance for all minority classes and pay for accurate evaluation for data classification performance, while some other features of microarray data will intensify this damage. So, it is compulsory to remedy this kind of bias by some effective strategies.

There are two main methods to deal with class imbalance problem: sampling-based strategy and cost sensitive learning. The Sampling, which cover oversampling and undersampling, deals with class imbalance by the way inserting samples for minority class samples or eliminate samples of majority class. At the same time cost-sensitive learning treats class imbalance through incurring different types of costs for different classes. Newly, some researchers are also concentrating on ensemble learning constructed on multiple different sampling method or weight- in datasets with the presenting best performance and generalization ability.

The implementation of binary classifier in the form of liner classifier produce such a problem, the first method relied on spreading binary classification problems to resolve the multiclass case unswervingly. This involved neural networks, support vector machines (SVM), decision trees, naive bayes, and k-nearest neighbours. The second method decomposes the multiclass problem into some binary classification tasks. Several attitudes are used for this decomposition: all-versus-all (OVA), one versus- one (OVO), error-correcting output coding, and generalized coding [2]. The third one relied on positioning the classes in a tree, fundamentally a binary tree, and applying a number of binary

ISSN: 2278 – 7798

classifiers at the nodes of the tree till a leaf node is extended [3].

In this study, this paper suggests presenting a novel undersampling method based on the technique of ant colony optimization (ACO), which is called ACO Sampling, to classify for skewed DNA microarray data [4]. ACOsampling is lead to find the corresponding optimal majority class sample subset. Considering the character of the classification tasks in this study, the overall accuracy is not an great measure as the fitness function, thus this paper proposes constructing it by three weighted indicative metrics, especially G-mean, F-measure and AUC, respectively. Following, many local optimal majority class sample subsets can be developed by iterative partitions, so the implication of each majority sample may be estimated according to its selection frequency, i.e., the higher selection frequency, the more evidence the equivalent sample set can provide. Next, a global optimum balanced sample subset can be established by combining the highly ranked sample set of majority class with all the examples of minority class. At last, this paper proposes constructing a SVM classifier upon the balanced training set for making future unlabelled samples.

The remainder of this paper is organized as follows. Section 2 reviews some work related with class imbalance problem. In Section 3, the idea and procedure of implementing ACOsampling method is described in detail. Experimental results and discussions are presented in Section 4. At last, concluding this paper in Section 5.

2. RELATED WORKS

Data preprocessing denote any type of processing performed on raw data to organize it for another processing technique. Commonly used as a preliminary data mining practice, data preprocessing changes the data into a specific format that will be easily and successfully processed for the persistence of user for example, in a neural network [5]. Utmost of classification algorithms are only focus on two-class imbalance problems. There are mysterious issues in multi-class imbalance problems, which happen in the real world applications. Consuming the method of binarization such as one-against-all (OAA) and one-against-one (OAO), can decrease the original multiclass imbalanced dataset into binary dataset [6].

Frequently called original training data set is splited into two classes such as testing set and training set. Then the training set can be splited into training set and validation set which are managed by the sampling methods [7]. The testing set is used to put on with the classifier to measure the performance [8].

The imbalanced dataset holds minority and majority class sample dataset. The Ant Colony Optimization (ACO) sampling algorithm is a undersampling technique used to mine the valuable and important dataset from the majority class samples [9]. And the Support Vector Machine (SVM) is greatest classifier used for multiclass imbalanced grouping [10].

It is well-known that in skewed appreciation tasks, overall accuracy (Acc) mostly gives bias estimation, thus some other exact evaluation metrics, such as G-mean, F-measure and area under the receiver operating characteristic curve (AUC), are needed to evaluation classification performance of a learner [11]. F-measure and G-mean may be observed as functions of the confusion matrix.

3. IMPLEMENTATION AND DESCRIPTION

Under sampling based on ant colony optimization

Ant colony optimization (ACO) algorithm, which is established by Colorni et al., is one great participant of swarm intelligence family.

ACO pretends the character of foraging by existent ant colony and in recent years, it has been effectively tested to solve various practical optimization problems, as well as path planning, protein folding travelling salesman problem (TSP), parameter optimization, etc. this paper offers intended an ACO algorithm to select thyroid-related marker genes in DNA micro array data. While in this study, this paper recommends transform it from feature space to sample space to examination an undersampling set which is noticed as the optimal subset projected on the given validation set.

In this ACO algorithm, many ants usually search pathways from nest to food. They select path ways rendering to the amounts of pheromone left in these path ways. The more pheromone is left, the more chance of corresponding pathway is selected. This paper propositions compute the prospect of choosing a pathway by:

$$Pij = \frac{\tau ij}{\sum_j^k \tau ij} \quad (1)$$

Where i denotes the ith site, i.e., the ith popular sample in original training set, j denotes

ISSN: 2278 – 7798

pathway, which may be allotted as 1 or 0 to denote whether choosing the equivalent sample or not. tij is pheromone concentration of the ith site in the jth pathway, pij and k are the probability of selection the jth pathway of the ith site and probable value of pathway j (0 or1), correspondingly. When an ant extents at the food source, the matching sample subset will be predictable by fitness function.

$$Fitness = \alpha \times F - measure + \beta \times G - mean + \gamma \times AUC$$
$$\text{s.t} : \alpha + \beta + \gamma = 1 \quad (2)$$

The fitness function is constitutive of three weighted metrics: G-mean, F-measure and AUC. When one sequence finishes, the pheromone of all pathways is modernised, the update function receives from the literature [38] and is defined as follows:

$$\tau ij(t+1) = \rho \times \tau ij(t) + \Delta \tau ij \quad (3)$$

Where r is the evaporation constant, which controls the decrement of pheromone, Dtij is increased pheromone of some excellent pathways. this paper proposes estimation pheromone e in the pathways of the finest 10% ants afterward each cycle and store these pathways in a set E.

Pseudo-code description of the undersampling algorithm based on ACO:

Input: Original training set: S,
Validation set:V.

Process:

For i=1:number of majority samples in S
for j=0:1
Assign initial pheromone ph_initial for pathwayij;
End for
End for
Set the optimal solution OPS=0;
For i=1:iteration times of ant colony
For j=1:size of ant colony *ant_n*
Acquire sampling set SSij by formula (1);
Train a classifier Cij for SSij;
Evaluate performance of Cij by V and formula (2);
End for
Find the optimal solution OPSi in the ith iteration;
If (OPS<OPSi)
OPS=OPSi;
End if
Update performance for each pathway by formula (3) and (4);
End for
Output: Undersampling training set S' which corresponds to OPS
Algorithm Source: ACO Sampling based under sampling method by Hualong Yu [1]

ACO Sampling Algorithm

By ACO algorithm specified overhead, an greatest undersampling subset may be mined as the last training set to build a classifier and identify future testing samples. Still, to monitor optimization process, this paper proposes to split the unique training set into two parts: training set and validation set, before ACO algorithm works. Generally, it can reason two severe problems for formed classifier: information loss and over fitting due to the assignation of validation set. In specific, when classification tasks are based on small sample set, these problems become more serious.

To solve this problem, this paper suggests a novel approach entitled as ACO Sampling to produce robust classifier by the combination of reduplicative partition of original sample set and ACO algorithm.

ISSN: 2278 – 7798

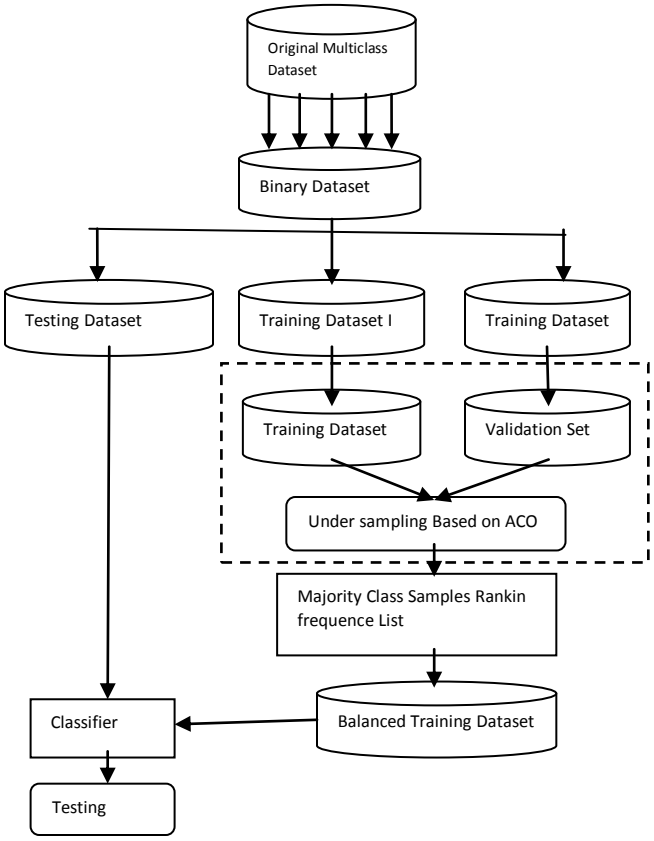


Fig. 1. The frame diagram of ACO Sampling strategy.

To solve this problem, this paper suggests a novel approach entitled as ACO Sampling to produce robust classifier by the combination of reduplicative partition of original sample set and ACO algorithm. The border diagram of ACO sampling strategy presents in Fig. 1.

Pseudo-code description of ACO Sampling strategy:

Input: Initial training set: IS

Process:

For i=1:100 (iteration times)

Divide randomly IS into two sets: training set S and validation set V;

Run undersampling algorithm based on ACO to acquire S';

Record manority class sample index of S' into RECi;

End for

Compute emerging times for each majority example based on all records REC1-REC100 and give the corresponding frequence list;

Rank all samples in descendent order according to the frequence list;

Combine some extremely ordered majority samples and all minority samples to build a balanced training set: BS.

Output: Final training set:BS

In particular, of the four SVM differences considered in this correspondence, the novel granular SVMs–repetitive undersampling algorithm (GSVM-RU) is the best in terms of both efficiency and effectiveness.

4. RESULTS AND DISCUSSION

In this section, the proposed system results have been discussed.

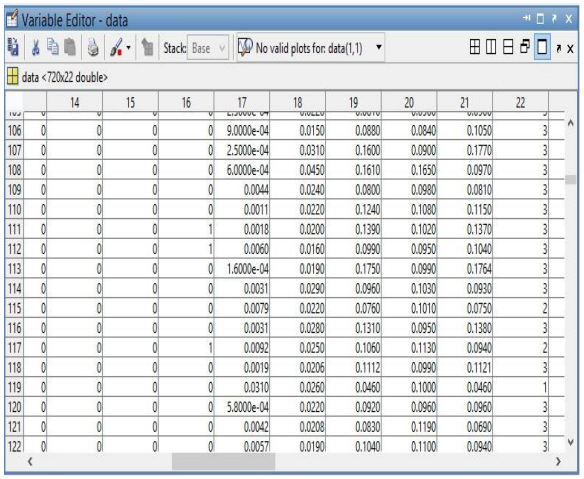


Fig. 2 shows the original multiclass dataset loaded into matlab software.

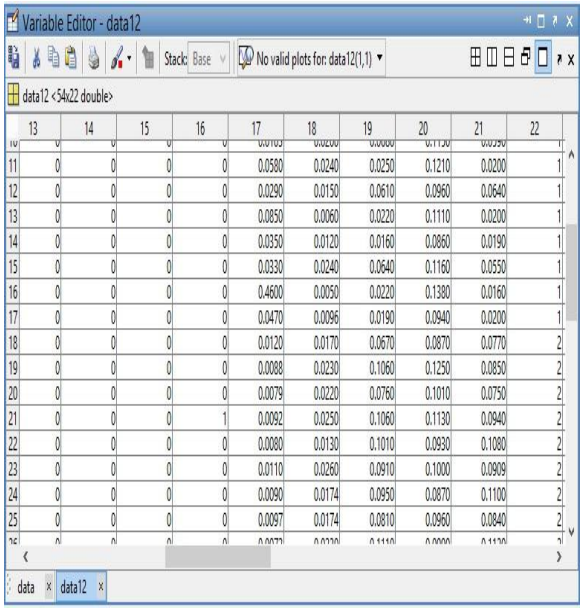


Fig. 3 shows the multiclass datasets are splitted into binary class based on binarization.

ISSN: 2278 – 7798

Performance (%)	**Sampling Method**	
	RUS	**ACO**
Lung Cancer		
Accuracy	0.8919	0.9459
F-measure	0.9333	0.9655
G-mean	0.7454	0.8819
Precision	0.8750	0.9333
Thyroid		
Accuracy	0.8421	0.9757
F-measure	0.4215	0.5455
G-mean	0.8123	0.8574
Precision	0.3921	0.4286
New-Thyroid		
Accuracy	0.9464	0.9636
F-measure	0.9725	0.9688
G-mean	0	0.9626
Precision	0.9464	0.9688
Contraceptive		
Accuracy	0.6655	0.7000
F-measure	0.7992	0.7307
G-mean	0	0.7320
Precision	0.6655	0.9077
Balance		
Accuracy	0.6765	0.5882
F-measure	0.4407	0.5227
G-mean	0.6330	0.6844
Precision	0.3611	0.3538

Table 1 showing that accuracy, f-measure, g-mean and precision value of two undersampling method such as Random Under Sampling (RUS) and Ant Colony Optimization (ACO) algorithm.

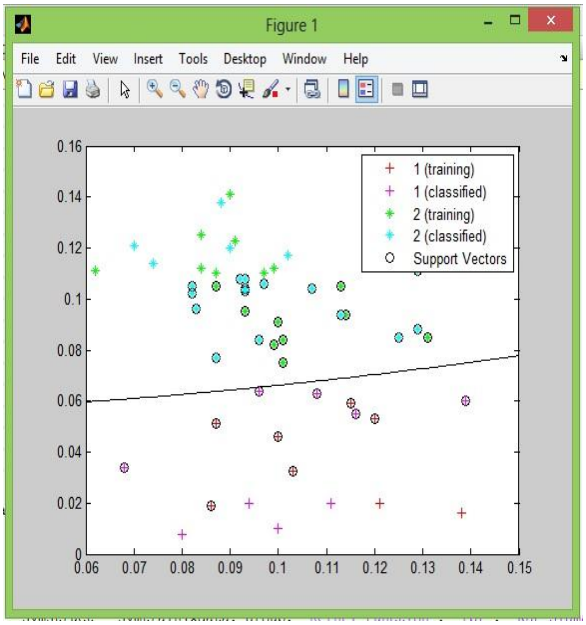
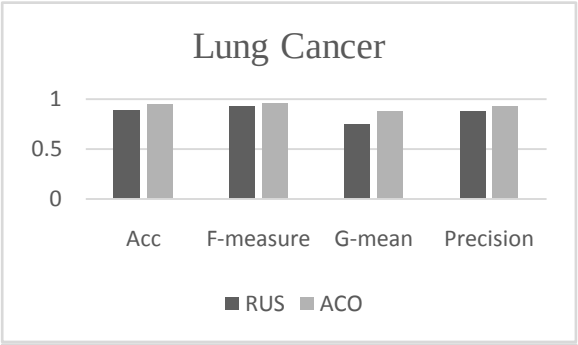
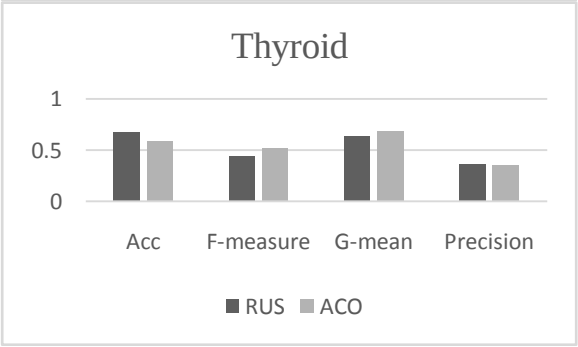


Fig. 4 shows the one of binary class datasets are trained and classified using SVM Classifier





ISSN: 2278 – 7798

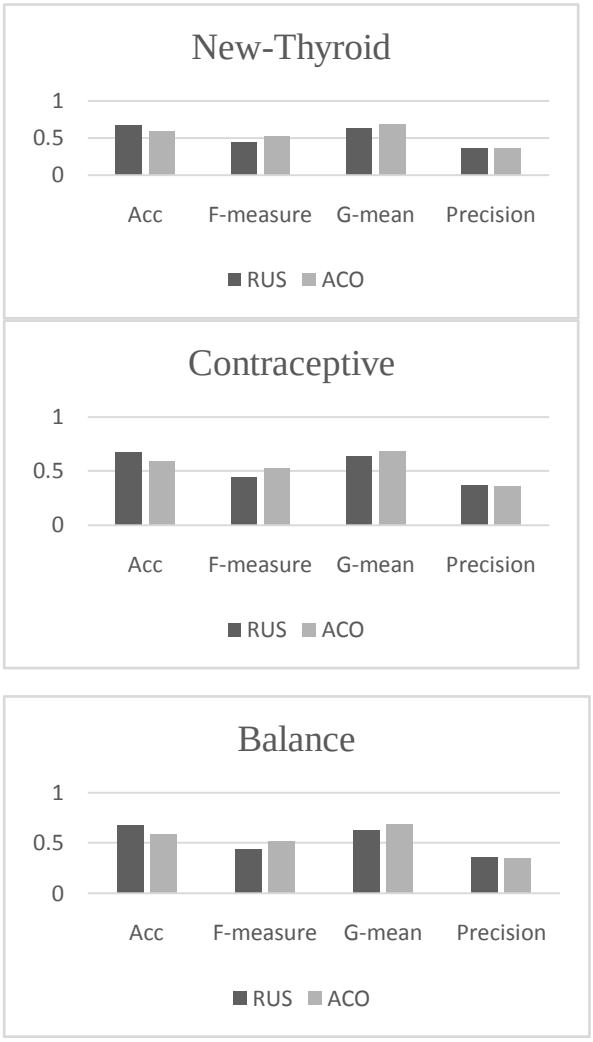


Fig. 5 shows that graphical representation of performance of ACO and RUS methods by using the dataset of lung cancer, thyroid, contraceptive and balance

5. CONCLUSION

Classifying the multiclass imbalanced by the binarization technique such as One-Against-One (OAO) for class decomposing original multiclass imbalanced dataset into binary dataset problem used to make support by the classifier support vector machine (SVM). This pairwise learning method split the multiclass dataset into supportable binary class dataset. And then constructing Ant Colony Optimization Sampling algorithm of swarm intelligence approach which works better for imbalanced classification of multiclass dataset. Using Support Vector Machine (SVM) classifier evaluate the DNA microarray dataset in ACOSampling method and also by the RUS (Random Under Sampling) method. The ACO Sampling algorithm gives better performance comparing with the Radom Under Sampling method.

6. REFERENCES

[1] Alberto Fern´andez, Mara Jos´e Del Jesus, and Francisco Herrera, "Multi-class Imbalanced Data-Sets with Linguistic Fuzzy Rule Based Classification Systems Based on Pairwise Learning", Fuzzy Sets and Systems 159(18), 2378–2398 (2008).

[2] Alberto Fernández, Victoria López, 2013, "Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches", Knowledge-Based Systems 42 97–110.

[3] D.H. Wolpert, W.G. Macready, 1997, "No free lunch theorems for optimization", IEEE Trans. Evol. Comput. 1(1) 67–82.

[4] David Martens, Manu De Backer, 2007, "Classification with Ant Colony Optimization", IEEE Vol. 11, No. 5.

[5] M. Wasikowski, X.W. Chen, 2010, "Combating the small sample class imbalance problem using feature selection", IEEE Trans. Knowl. Data Eng. 22(10) 1388–1400.

[6] Mahendra Sahare, Hitesh Gupta, 2012, "A Review of Multi-Class Classification for Imbalanced Data," ISSN (online): 2277-7970, Volume-2 Number-3.

[7] Minlong Lin, Ke Tang, 2013, "Dynamic Sampling Approach to Training Neural Networks for Multiclass Imbalance Classification," IEEE, Vol. 24, NO. 4.

[8] Piyaphol Phoungphol, Yanqing Zhang, Yichuan Zhao, 2012, "Robust Multiclass Classification for Learning from Imbalanced Biomedical Data", ISSN 1007-0214 02/10 pp619-628.

[9] Q. Shen, Z. Mei, B.X. Ye, 2009, "Simultaneous genes and training sample sselection by modified particle swarm optimization for gene expression data classification", Comput. Biol. Med. 39(7) 646–649.

[10] Qiang Yu a, HuajinTang b,c,n, KayChenTan a, HaoyongYu, 2009, "SVMs Modeling for Highly Imbalanced Classification" IEEE Part B: Cybernetics, Vol. 39, No. 1, February.

[11] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, et al., 2002, "Prediction of central nervous system embryonal tumour outcome based on gene expression", Nature 415 (6870) 436–442.

ISSN: 2278 – 7798