ISSN 2348 – 9928
IJAICT Volume 5, Issue 6, June 2018 Doi:01.0401/ijaict.2015.11.15 Published on 05 (6) 2018

IMBALANCED MULTICLASS DATA CLASSIFICATION USING ANT COLONY OPTIMIZATION ALGORITHM

Mr. L. Sriram
PG Scholar,
Computer Science and Engineering,
Anna University Regional Centre, Coimbatore, Tamilnadu, India

Mrs. S. Lavanya
Assistant Professor,
Computer Science and Engineering,
Anna University Regional Centre, Coimbatore, Tamilnadu, India

Mr. K. Vishnu.
PG Scholar,
Computer Science and Engineering,
Anna University Regional Centre, Coimbatore, Tamilnadu, India

Abstract— Class imbalance problems have drawn increasing interest lately because of its classification trouble caused by imbalanced class distributions and poor prediction performance for minority class. Many ensemble approaches only concentrated on two-class imbalance problems. There are many unresolved concerns in multiclass imbalanced problems. Using One-vs-One binarization technique for decomposing the original multiclass data-set into binary classification problems. Then, whenever each one of 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 mix those with all minority samples to build the final balanced training set. Lastly, evaluate the method on four benchmarks skewed DNA microarray dataset by support vector machine (SVM) Classifier.

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

I. INTRODUCTION

In the past decade, DNA micro array data has been one of the important molecular biology technologies in post-genomic era. Using this, biologists, scientist and medical experts are permitted to detect the activity of thousands of genes in a cell simultaneously. For now, DNA micro array dataset has been widely applied to predict gene expression and functions, examine gene regulatory activities, provide valuable information for drug discovery and also used classify for cancer and mining new subtypes of a specific type like thyroid, tumor etc [1]. Among these useful applications, the cancer classification has been attracted more attentions. For all that, it is well-known the micro array data generally has some appropriate features, like high dimension, small sample set, huge noise and generally, imbalanced class distributions. Skewed class distributions will underestimate extremely the prediction performance for all minority classes and afford accurate evaluation for data classification performance, while some other features of microarray data will intensify this damage. So, it is mandatory 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 concentrate 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 generate such a problem, the first method relied on extending binary classification problems to solve the multiclass case directly. This included neural networks, decision trees, support vector machines (SVM), naive bayes, and k-nearest neighbours. The second approach decomposes the multiclass problem into several binary classification tasks. Several approaches are used for this decomposition: one versus- all (OVO), all-versus-all (OVA), erorr-correcting output coding, and generalized coding [2]. The third one relied on arranging the classes in a tree, basically a binary tree, and utilizing a number of binary classifiers at the nodes of the tree till a leaf node is reached [3].

In this study, this paper proposes introducing a novel undersampling method based on the technique of ant colony optimization (ACO), which is called ACOSampling, 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

ISSN 2348 – 9928
Doi:01.0401/ijaict.2015.11.15 Published on 05 (6) 2018

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 F-measure, G-mean 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 information the corresponding sample set can provide. 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.

II. RELATED WORKS

Data preprocessing represent any type of processing executed on raw data to prepare it for another processing procedure. Generally used as a preliminary data mining practice, data preprocessing converts the data into a particular format that will be easily and effectively processed for the purpose of user for example, in a neural network [5]. Most of classification algorithms are only concentrate on two-class imbalance problems. There are inexplicable issues in multi-class imbalance problems, which exist in the real world applications. Using the method of binarization such as one-against-all (OAA) and one-against-one (OAO), can reduce the original multiclass imbalanced dataset into binary dataset [6].

Commonly called original training data set is splited into two category such as testing set and training set. Then the training set can be splited into training set and validation set which are processed by the sampling methods [7]. The testing set is used to apply with the classifier to measure the performance [8].

The imbalanced dataset contains minority and majority class sample dataset. The Ant Colony Optimization (ACO) Sampling algorithm is a undersampling method used to extract the valuable and important dataset from the majority class samples [9]. And the Support Vector Machine (SVM) is best classifier used for multiclass imbalanced classification [10].

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

III. IMPLEMENTATION AND DESCRIPTION

3.1 Under sampling based on ant colony optimization

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

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

In this ACO algorithm, many ants commonly search pathways from nest to food. They choose path ways according to the quantities of pheromone left in these path ways. The more pheromone is left, the more probability of corresponding pathway is selected. This paper proposes compute the probability of selecting a pathway by:

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

Where i represents the ith site, i.e., the ith majority sample in original training set, j represents pathway, which may be assigned as 1 or 0 to denote whether choosing the corresponding sample or not. tij is pheromone intensity of the ith site in the jth pathway, pij and k are the probability of picking the jth pathway of the ith site and possible value of pathway j (0 or1), respectively. When an ant reaches at the food source, the corresponding sample subset will be estimated 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: F-measure, G-mean and AUC. When one cycle finishes, the pheromone of all pathways is restructured, the update function inherits from the literature [38] and is described as follows:

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

ISSN 2348 – 9928
Doi:01.0401/ijaict.2015.11.15 Published on 05 (6) 2018

Where ρ is the evaporation coefficient, which controls the decrement of pheromone, Dtijis increased pheromone of some excellent pathways. In this paper, this paper proposes totalling pheromone e in the pathways of the finest 10% ants after each cycle and store these pathways in a set E

3.2 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

3.3 ACOSampling Algorithm

By ACO algorithm stated above, an best undersampling subset may be extracted as the last training set to build a classifier and recognize future testing samples. Still, to guide optimization procedure, this paper proposes to split original training set into two parts: training set and validation set, before ACO algorithm works. Normally, it can cause two severe problems for created classifier: information loss and over fitting due to the engagement of validation set. In particular, when classification tasks are based on small sample set, these problems become more serious

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

3.4 Pseudo-code description of ACOSampling 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;

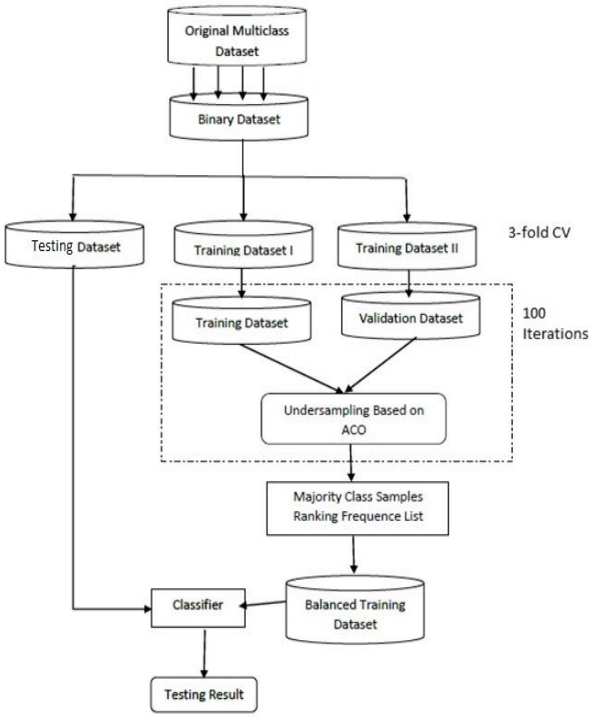

Fig 1 : The frame diagram of ACO Sampling strategy.

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

Combine some highly ranked majority samples and all minority samples to construct a balanced training set: BS.

Output: Final training set:BS

In particular, of the four SVM variations considered in this correspondence, the novel granular SVMs–repetitive

undersampling algorithm (GSVM-RU) is the best in terms of both effectiveness and efficiency.

IV. RESULTS AND DISCUSSION

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

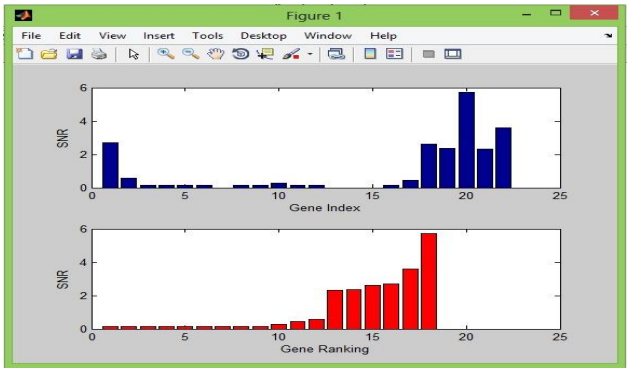

Fig 2 : shows that selecting few dataset which are strongly related to classification task based on SNR value distribution for gene index and gene ranking.

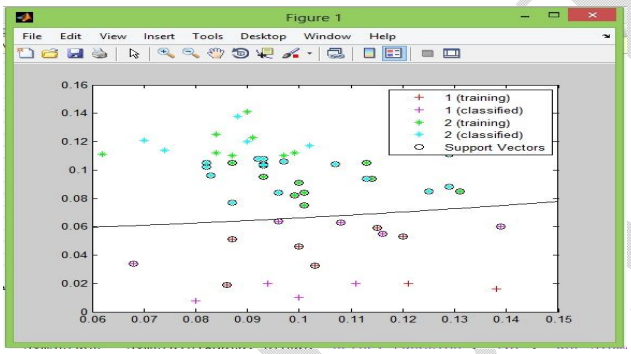

Fig 3 : shows the one of binary class datasets are trained and classified using SVM Classifier

V. 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 pair wise 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. Future work is constructing Support Vector Machine (SVM) classifier evaluate the multiclass DNA microarray dataset.

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] 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).

[9] 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.

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

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

[12] 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.

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

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

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

[16] 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.

[17] 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.

[18] 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.