International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 3, Issue 9, September 2014 ISSN 2319 – 4847

STUDY OF MULTICLASS CLASSIFICATION FOR IMBALANCED BIOMEDICAL DATA

Mr. Roshan M. Pote[1], Prof. Mr. Shrikant P. Akarte[2]

[1]ME (CSE) ,Second Year,Department of CSE,Prof. Ram Meghe Institute Of Technology and Research, Badnera,Amravati
Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701.

[2] Assistant Professor, Department of CSE, Prof. Ram Meghe Institute Of Technology and Research, Badnera, Amravati.
Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701.

ABSTRACT

In this paper an approach is introduced for more than two class classification which can combine more than two SVM classifiers. By combining the results which are obtained from various binary SVM classifiers, Multi-class classification is performed. the features are extracted, then the SVM methods are applied to the extracted feature set which are the unbalance in the dataset because they didn't perform well[1]. This paper present an experimental results on multiple biomedical datasets show that the proposed solution can effectively cure the problem when the datasets are noisy and highly imbalanced.

Keywords: Multiclass Classification, Imbalanced Data, Support Vector Machine (SVM), Biomedical Data.

1. INTRODUCTION

One of the important data mining methods in biomedical research is classification. In the classification task, training examples are required to presage a target class of an unseen example. Nevertheless, training data sometimes have imbalanced class distribution. Inadequacy of absolute amount of examples of some classes for training a classifier is one of major reasons for the problem. In the field of biomedical, the issue of learning from these imbalanced data is highly important because it can invent useful knowledge to make important decision on the other hand it can also be extremely costly to misclassify these data. The remainder of this paper is organized as follows. Section III provides a description of Multiclass Classification, which Section IV describes the Research Work. Section V describes Methodological Analysis which contains four types of analysis, section VI describes Multiclass Classification Objective , section VII describes Multiclass Classification With Ramp Loss**,** section VIII presents Comparative Analysis, Finally section IX, concludes this paper.

2. LITERATURE REVIEW

Imbalanced data is a common and serious problem in many biomedical classification tasks. It creates lots of confusion on the training of classifiers and results in lower accuracy of minority classes prediction [2]. This is engendered due to the less availability or by limitations on data collection process such as high cost or privacy problems. The majority classes' overloads standard machine learning algorithms that it cannot bear the load and traditional classifiers making confusions between the decisions towards the majority class and try to optimize overall accuracy. To improve traditional methods or to develop new algorithms for solving the problem of class imbalance, many researches were done. Most of those studies are focused only on binary case or two classes. Only a few researches have been done for multiclass imbalance problem which are more common and complicated in structure in the real-world application.

3. MULTICLASS CLASSIFICATION

Multiclass or multinomial classification is the problem of classifying instances into more than two classes. While some classification algorithms naturally permit the use of more than two classes, others are by nature binary algorithms; these can, however, be turned into multinomial classifiers by a variety of strategies. Multiclass classification should not be confused with multi-label classification, where multiple labels are to be predicted for each instance [3].

A. Imbalanced Data

Errors of major class instances will control the total error. Thus, a classifier will certainly be biased toward the major class to minimize the total errors, as shown in the figure.

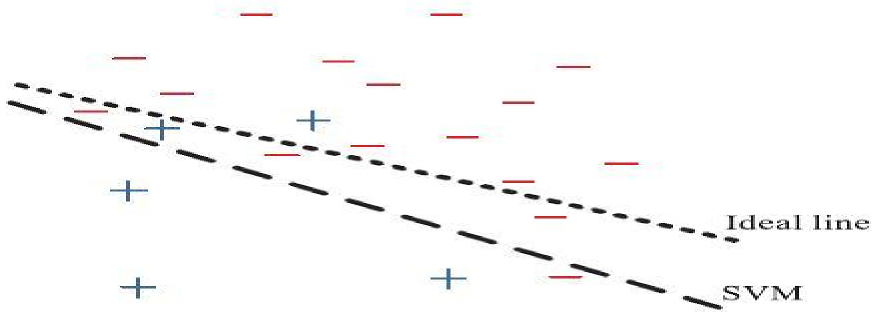


Fig.1: SVM on imbalanced dataset is biased toward the major class**.**

B. Noisy Data

An error of each point can take values ranging from 0 to +∞, therefore errors of a few bad or noisy points can critically compromise the overall errors which result in impairment of classifier's performance. From both cases that the summation of all errors is not suitable for use as objective function of the optimization problem [4].

4. RESEARCH WORK

The problems of imbalanced data classification have been studied by many researchers to improve the performance of classification models. Binary classification was the main area of interest of their work. General ideas such as feature selection [5], sampling-based approach, and cost sensitive learning can easily be extended to multiclass problem, but it is rather difficult for algorithm-specific techniques [6, 7]. The related work is presented shows promising results in improving multiclass imbalance classification.

5. METHODOLOGICAL ANALYSIS

A. Sampling

This is the common approach to deal with class imbalance. To deal with the class imbalance the original class frequencies are changed at a preprocessing step.

B. Cost-Sensitive Learning

It considers the costs of misclassification data in one class to another class, and then attempts to minimize total costs instead of total errors. A widely used approach in applying cost sensitive approach to SVM [8] is to assign a higher penalty error for a minority class in the optimization problem. Zhou and Liu advocated a rescaling technique [9] to solve this problem. In converting a confusion matrix € $_{I,j}$ to class penalty factors for multiclass SVM model. The weights of each class w_r will be rescaled simultaneously according to their misclassification costs. Solving the relations in Eq. (1) will get relative optimal weights of imbalanced data for multiclass SVM.

$$\frac{\omega_1}{\omega_2}=\frac{\epsilon_{1.2}}{\epsilon_{2.1}}\,,\quad \frac{\omega_1}{\omega_3}=\frac{\epsilon_{1.3}}{\epsilon_{3.1}},\quad \dots\dots,\frac{\omega_1}{\omega_m}=\frac{\epsilon_{1.m}}{\epsilon_{m.1}}$$

$$\frac{\omega_2}{\omega_3}=\frac{\epsilon_{2.3}}{\epsilon_{3.2}},\quad \dots\dots,\frac{\omega_2}{\omega_m}=\frac{\epsilon_{2.m}}{\epsilon_{m.2}}\qquad (1)$$

$$\dots\dots,\quad \dots\dots$$

$$\dots\dots,\frac{\omega_m-1}{\omega_m}=\frac{\epsilon_{m-1.m}}{\epsilon_{m.m-1}}$$

With the optimal weight for each class, the objective of multiclass SVM in formula can be re written as

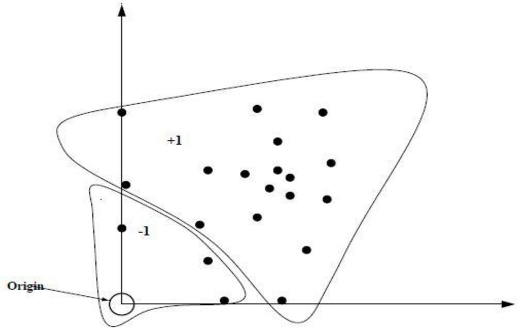
$$\min_{W}\frac{1}{2}\|W\|^2+\ C\sum_{i=1}^{n}\omega_{yi}\,.\,\xi i\qquad (2)$$

C. One Class SVM

Instead of differentiating all examples one class SVM only acknowledges examples from one class. When examples from the target classes are rare or difficult to obtain, It is useful. Sch¨olkopf et al. [10] proposed a maximum margin based classifier which is an adaptation of the Support Vector Machine algorithm for the case of one-class classification. This classifier separates the training data from the origin by the mean of a separating hyper plane$\langle w, z\rangle - \rho$, where w is the normal vector of the hyperplane and ρ its bias (it represents the distance from the hyper plane to the origin). Then, to separate the data set from the origin, one needs to solve the following quadratic programming problem.

$$\min_{\square} \frac{1}{2}||W||^2 + \frac{1}{\upsilon\iota}\sum_{i=1}^{n} \xi i - \rho$$

In order to apply one-class SVM for multiclass problem, multiple models of one-class SVM will be trained together



One-Class SVM Classifier. The origin is the only original member of the second class.

Figure 2: One class SVM

D. Boosting Classifier

Another approach to solve class-imbalanced problem for 2-class datasets is boosting classifier. It iteratively improves the classifiers' performance by either updating misclassification cost [11] or changing the distribution ratio [7] of each class. This approach contains the main drawback that it slows the speed, it might require a classifier to train datasets multiple times until the result is met or all data points are classified. Moreover, another method is used. It is a fast and effective technique, called ConCave-Convex Procedure (CCCP) [12], to solve the optimized problem.

6. MULTICLASS CLASSIFICATION OBJECTIVE

In this section, a new objective function is introduced that is more suitable for imbalanced data classification. The popular measures which have been used to compare the performance of learning models for imbalanced data.

A. Metrics for imbalanced classification

The most widely used measure in comparing the performance of classifier models is Accuracy (Acc), but, it is normally not relevant for an imbalanced data set. Alternative performance measures are available, such as G-mean, F-measure, and volume under Receiver Operating Characteristic (ROC).

B. G-mean

To solve the problems of Acc in imbalanced dataset G-mean has been introduced. It is the geometric mean of all class accuracies, so the G-mean for m class can be calculated by Equation.

$$G - mean = \left(\prod_{i=1}^{m} acc_i\right)^{1/m}$$

C. F-measure

Evaluation of the text classification system is done widely by using F-measure. It is defined as a harmonic mean of Precision. It is defined as a harmonic mean of Precision (π) and Recall (ρ) in binary class problem. To extend F-measure to multiclass, two types of average, micro-average and macro-average are commonly used [13].

$$F - measure = \frac{2.\pi.\rho}{\pi + \rho}$$

D. Volume under ROC (multiclass AUC)

The Area Under the ROC Curve AUC metric is one of the most popular measures in comparing binary classifiers because there is no necessitance to extend F-measure to multiclass. There are two types of average; micro-average and macro-average are commonly used [13].

7. MULTICLASS CLASSIFICATION WITH RAMP LOSS

SVM with 0-1 loss function was tried to solve by replacing it with other smooth functions like sigmoid function [14, 15], logistic function [16], polynomial function [15], and hyperbolic tangent function [17]. While giving accurate result these functions, still suffer from being computationally- overpriced when solved as an optimization problem due to its non-convex nature. Alternatively, a short hinge loss or ramp loss function which is a non-smooth but continuous function has been proved to be accurate and efficient for SVM problem [18-20].

Algorithm For Multiclass SVM with Ramp Loss

```
1: Find ratios of each class: $ratio_r$
2: Calculate $C_r$ for each class:
   $C_r = C/ratio_r$
3: initialize $\alpha$
4: repeat
5: compute $\xi_i = 1 + max_{r \overset{m}{\neq} y_i} \sum_{j=1}^{n} \alpha_{j,r} K(x_i,x_j) - \sum_{i=1}^{n} \alpha_{j,r} K(x_i,x_j)$
6: $min_{\alpha} \frac{1}{2} \sum_{i=1}^{n} \sum_{i=1}^{n} K(x_i,x_j) - \sum_{i=1}^{n} \alpha_{i,y_i}$
s.t.    $\alpha_{i,r} \leq \begin{cases} C_{yi} & if\ r = y_i\ and\ \xi_i \leq z \\ 0 & otherwise \end{cases}$
7: until $\alpha$ convergs
8: $f(x) = arg\ max_{r \epsilon Y} \sum_{i=1}^{n} \alpha_{i,r} K(x_i,x)$
```

A. Ramp loss function

The ramp loss or truncated hinge loss will be in range [0,1] when error ξ is less than a constant ramp parameter, and z will be equal to 1 when ξ is greater than z.

$$\min_{w} \frac{1}{2} \|W\|^2 + \sum_{i=1}^{n} C_{yi} \xi i - \sum_{i=1}^{n} C_{yi}\ H_z\ (\xi i)$$

8. COMPARATIVE ANALYSIS

In biomedical data the multiclass imbalance problem is commonly found in many classification systems. It Become worse in many application due to noisy and imbalanced data. For that purpose many alternative methods were studied to find the better solution for the management of imbalanced data. It is found that the workings of researchers were focused at only single approach, because it was difficult to combine the solution in one way. Some worked on sampling in which original class frequencies are changed at a preprocessing step. In cost sensitive learning it considers the costs of misclassification data in one class to another class, and then attempts to minimize total costs instead of total errors. Where as in one class SVM only admit the existence examples from one class. But boosting classifier used for 2-class dataset to solve class-imbalance problem. The objectives of multiclass classifiers are studied in which different parameters are optimized such as matrices for imbalanced class, G-mean, F-measure & volume under ROC. The effective method which was studies was multiclass classification with ramp loss in which the algorithm is proposed to manage the data imbalance problem more effectively than the other technique which was studied.

9. CONCLUSION

This paper presents the different techniques used for management of data imbalance, such as Sampling, cost sensitive learning, One Class SVM, Boosting Classifier. Also the different parameters like matrices for imbalanced classes, G-mean, F-measure & volume under ROC and the most effective method classification with ramp loss for multiclass classification is studied and algorithm for Multiclass SVM with Ramp Loss is analyzed.

REFERENCES

[1] Weiss G M. Mining with rarity: A unifying framework. SIGKDD Explor Newsl, 2004, 6: 7-19.
[2] Chawla N V, Japkowicz N. Editorial: Special issue on learning from imbalanced datasets. SIGKDD Explorations, 2004, 6: 1-6.
[3] Ferri C, Hernndez-orallo J, Salido M. Volume under the ROC surface for multi-class problems. exact computation and evaluation of approximations. In: Proc. of 14th European Conference on Machine Learning. Cavtat- Dubrovnik, Croatia, 2003: 108-120.
[4] Yang X Y, Liu J, Zhang M Q, Niu K. A new multi-class SVM algorithm based on one-class SVM. In: Proceedings of the 7th International Conference on Computational Science. Beijing, China, 2007: 677-684.
[5] Wasikowski M, Chen X W. Combating the small sample class imbalance problem using feature selection. Knowledge and Data Engineering, IEEE Transactions on, 2010, 22: 1388-1400.
[6] Chen X, Gerlach B, Casasent D. Pruning support vectors for imbalanced data classification. In: Proceedings ofThe International Joint Conference on Neural Networks. Montrcal, Canada, 2005: 1883-1888.
[7] Tang Y, Zhang Y Q, Chawla N, Krasser S. SVMs modeling for highly imbalanced classification. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 2009, 39: 281-288.ine, 2001, 22: 127-136.
[8] Bach F R, Heckerman D, Horvitz E. Considering cost asymmetry in learning classifiers. J. Mach. Learn. Res.,2006, 7: 1713-1741.
[9] Zhou Z H, Liu X Y. On multi-class cost-sensitive learning. Computational Intelligence, 2010, 26: 232-257.

[10] B. Sch¨olkopf, J. Platt, J. Shawe-Taylor, A. J. Smola, R. Williamson, Estimating the support of a high-dimentional distribution, Neural Computation 13 (2001) 1443–1471.

[11] Sun Y, Kamel M S, Wong A K, Wang Y. Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition, 2007, 40: 3358-3378.

[12] Yuille A L, Rangarajan A. The concave-convex procedure (CCCP). Advances in Neural Information Processing Systems, 2002, 15: 915-936.

[13] O¨ zgu¨r A, O¨ zgu¨r L, Gu¨ngo¨r T. Text categorization with class-based and corpus-based keyword selection. In: Proceedings of the 20th International Conference on Computer and Information Sciences. Istanbul, Turkey, 2005: 606-615.

[14] Ratnagiri M, Rabiner L, Juang B H. Multi-class classification using a new sigmoid loss function for minimum classification error (MCE). In: Proceedings of the 9th International Conference on Machine Learning and Applications (ICMLA). Washington DC, USA, 2010: 84- 89.

[15] Perez-Cruz F, Navia-Vazquez A, Figueiras-Vidal A, Artes- Rodriguez A. Empirical risk minimization for support vector classifiers. Neural Networks, IEEE Transactions on, 2003, 14: 296-303.

[16] Liu Y, Shen X. Multicategory -Learning. Journal of the American Statistical Association, 2006, 101: 500-509.

[17] Perez-Cruz F, Navia-Vazquez A, Alarcon-Diana P, Artes- Rodriguez A. Support vector classifier with hyperbolic tangent penalty function. In: Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Istanbul, Turkey, 2000: 3458-3461

[18] Collobert R, Sinz F, Weston J, Bottou L. Trading convexity for scalability. In: Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, USA, 2006: 201-208.

[19] Wu Y, Liu Y. Robust Truncated hinge-loss support vector machines. JASA, 2007, 102: 974-983.

[20] Phoungphol P, Zhang Y, Zhao Y. Multiclass SVM with ramp loss for imbalanced data classification. In: Proceedings of the 2012 IEEE International Conference on Granular Computing. Hangzhou, China, 2012: 160-165.

AUTHOR

Mr. Roshan M. Pote, ME (CSE) ,Second Year,Department of CSE,Prof. Ram Meghe Institute Of Technology and Research, Badnera, Amravati. Sant Gadgebaba Amravati University, Amravati, Maharashtra, India – 444701.

Prof. Mr. Shrikant P. Akarte, [2] Assistant Professor,Department of CSE, Prof. Ram Meghe Institute Of Technology and Research, Badnera,Amravati. Sant Gadgebaba Amravati University, Amarvati, Maharashtra, India – 444701.