A Comparative Study of DataBoost.IM for Detect Root Cell of Liver Cancer

K. Lokanayaki , Dr.A.Malathi

Abstract: Ensemble learning is an active field in Data Mining research in last decade. Many literatures have been emerged and a terrific program has been made. It is important for data miners to achieve high classification accuracy of prediction models and especially true for imbalanced data. This paper has been designed to find out the best mining classification method DataBoost.IM for detect root cell of liver cancer with imbalanced dataset. The essential idea of DataBoost.IM is to use a weighted distribution for different minority class samples and majority class samples according to their level of difficulty in imbalanced learning. Finally, we proposed a new concept of analyzing ensemble learning DataBoost.IM algorithm which shows high efficiency compare than AdaBoost, DataBoost and AdaBoost1.

Index-Term - **DataBoost.IM , AdaBoost, AdaBoost1, Imbalanced Dataset**

I. INTRODUCTION

Many real world data mining applications involve learning from imbalanced data sets. When learning imbalanced data sets machine learning algorithms tend to produce high predictive accuracy over the majority class and minority class, but poor predictive accuracy over the minority class [3].

First Author namee: *,K.Lokanayaki, Assistant Professor, Department of Computer Application, Spurthy Group of Institutions Bangalore ,India*

Second Author name*,Dr.A.Malathi,* Assistant Professor, Department of Computer Science, Government Arts College, Coimbatore, India

Classification of imbalanced datasets is a common problem in many domains, such as, network intrusion detection, detecting fraudulent transactions, direct marketing Web mining, and medical diagnostics.For example, in medical databases when classifying the data from imbalanced dataset as abnormal or not[4] .

The nature of the application requires a fairly high detection rate of the minority class and allows for a small error ate in the majority class since the cost of misclassifying a cancerous patient as non-cancerous can be very high.

II. RELATED WORK:

Boosting was introduced by Schapire in 1990 as a method for boosting the performance of a weak learning algorithm. Recently expanded in Freund in 1996[7]. AdaBoost (Adaptive Boosting) was introduced by Freund & Schapire in 1995.It also called as AdaBoost.M1. The AdaBoost algorithm generates a set of classifiers and votes based on two algorithms differ substantially. It generates the classifiers sequentially and range the weights of the training instances provided as input to each inducer [10].

In particular, boosting is an ensemble method where the performance of weak classifiers is improved by focusing on seed examples which are difficult to classify. Boosting algorithms found a series of classifiers and the outputs of these classifiers are combined using weighted voting in the final prediction of the model [10].

In each step of the series, the training examples are re-weighted and selected based on the performance of earlier classifiers in the training series. This produces a

ISSN: 2278 – 7798

set of "easy" examples with low weights and a set of hard ones with high weights. During each of the iterations, boosting attempts to produce new classifiers that are better able to predict examples for which the previous classifier's performance is poor. This is achieved by concentrating on classifying the hard examples correctly. Recent studies have indicated that boosting algorithm is applicable to a broad spectrum of problems with great success [10, 11].

III. ENSEMBLE LEARNING FOR CLASS IMBALANCE:

An ensemble model is a combination of two or more models to avoid the drawbacks of individual models and to achieve high accuracy. The two models are combined by using high confidential wins scheme [14].Combining ensemble learning with sampling techniques i.e. oversampling, undersampling and SMOTE, to create balanced samples is a popular approach for classification with imbalanced data [16][17]. In [17],two new boosting undersampling methods are developed for create balanced samples. It also compared to 13 other sampling and boosting approaches in the literature.

In [15] trained on balanced and imbalanced bootstrap samples using base classifiers and also compared a bagging approach for fraud detection. The overall classification accuracy in trained on balanced samples using base classifiers to be more effective. In [16] bagging method in ensemble learning combined with EMO approach is used to evolve a population of binary classifiers for two UCI benchmark dataset.

In imbalanced datasets improve minority class accuracy using ensemble diversity but degrade majority class accuracy. However, the accuracies of both classes improve together when ensemble diversity is increased in balanced data sets. The [18] NCL approach used for minority class problem in imbalanced dataset. NCL approach used to maximize ensemble accuracy in minority class. It also combined with SPEA2 for evolve a diverse set of Pareto front classifiers using a GP approach for ensemble learning in class imbalance problem.

In order to ensure optimal classification accuracy for minority and majority class, DataBoost-IM algorithm was proposed in where synthetic data examples are generated for both minority and majority classes through the use of "seed" samples. This algorithm combines boosting, an ensemble-based learning algorithm, with data generation developed by Hongyu Guo and Herna L Viktor in 2004 [13].This algorithm to identify separately hard examples from generate synthetic examples for class. It also calculates the overall class distribution and the total weights of class's are rebalanced to improve the learning algorithms of majority class and minority class. It assigned an equal weight for each samples of original training set. It identified hard samples (seed samples).Generate set of synthetic samples and added to the original training set. Calculate the total weights of different classes are rebalanced Find error rate of classifier for a threshold value.

IV. EXPERIMENTAL DESIGN

The comparative analysis of the proposed work has been performed on Weka tool [15] using liver cell dataset. This dataset contains 1650 liver patient records with 10 attributes that are fine-needle aspiration biopsy (FNAB) specimen's tests. In this dataset the liver cell function tests are total high cellularity, acinar pattern, trabecular pattern, hyperchromasia, pleomorphism, irregularly granular chromatin, uniformly prominent nucleoli, multiple nucleoli, increased nuclear/cytoplasmic ratio, and atypical naked hepatocytic nuclei.

This dataset contains 982 These attributes were examined in a series of 82 FNAB specimens from 982 liver cancer cell records and 668 non liver cancer cell records With the use of a step-wise logistic regression analysis, three features were identified as predictive of HCC: increased nuclear/cytoplasmic ratio ($P = 0.001$), trabecular pattern ($P = 0.002$), and atypical naked hepatocytic nuclei ($P = 0.03$). The experimental result has

ISSN: 2278 – 7798

been evaluated based on three major parameter as accuracy, mean square error and time which are shown in table2.

The experimental result has been evaluated based on three major parameters as accuracy, mean square error and time which are shown in table2.

Table2: Performance of various learning techniques

Learning Methods	**Accuracy (%)s**	**Mean Square Error**	**Time (sec)**
AdaBoost M1	72%	0.4558	0.03
Bagging	63%	0.4627	0.05
J48	64%	0.5037	0.02
LogitBoost	67%	0.4513	0.03
DataBoost.IM	74%	0.3527	0.65

From the above table it's clear that DataBoost.IM performs better in all four aspects to detect whether root cell or not of liver cancer. Since the time taken to detect root cell of liver cancer may increase which improves detection accuracy and drastically reduce mean square error. This paper presents a comparative analysis between various machine learning techniques such as AdaBoost M1, Bagging, J48, LogitBoost and DataBoost.IM to detect root cell or not of liver cancer. Each machine learning technique has their own merits to improve classification accuracy and to build pattern classification. From the above result it's clear that DataBoost.IM performs better than other existing machine learning techniques.

V. CONCLUSION:

This paper presents a comparative analysis between various machine learning techniques such as AdaBoost Ma,Bagging J48,LogitBoost and Databoost.IM.Each ensemble learning technique and machine has their own merits to improve classification accuracy and to build pattern classification. From the above result it's clear that DataBoost.IM performs better than other existing ensemble learning and machine learning techniques. In this paper has also analyzed Individual cell classification, since the dataset has only limited number of records which may reduce overall cell detection performance and the main aim of this paper is to examine individual cell detection completely.

REFERENCES:

[1] N. Japkowicz. Learning from imbalanced data sets: A comparison of various strategies, Learning from imbalanced data sets: The AAAI Workshop 10-15. Menlo Park, CA: AAAI Press. Technical Report WS-00-05, 2000.

[2] N. Chawla, K. Bowyer, L. Hall and W. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357, 2002.

[3] M.A. Maloof . Learning when data sets are Imbalanced and when costs are unequal and unknown, ICML-2003 Workshop on Learning from Imbalanced Data Sets II, 2003.

[4] Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan, P.: "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications". In: Proc. ACM SIGMOD Int. Conf. on Management of Data, Seattle, WA. (1998).

[5] M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: One-sided selection. Proceedings of the Fourteenth International Conference on Machine Learning San Francisco, CA, Morgan Kaufmann, 179- 186,1997

[6] M. Joshi, V. Kumar and R. Agarwal. Evaluating boosting algorithms to classify rare classes: comparison and improvements. Technical Report RC-22147, IBM Research Division, 2001.

[7] Drucker, H. & Cortes, C. (1996), Boosting decision trees, in `Advances in Neural Information processing Systems 8', pp. 479 485.

[8] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. the Proceedings of the Thirteenth International Conference on Machine Learning, Bari, Ita ly, 148-156, 1996

[9] Y. Freund and R.E.Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139, 1997.

[10] H.Schwenk and Y. Bengio. AdaBoosting Neural Networks Application to On-line Character Recognition, International Conference on Artificial Neural Networks (ICANN'97), Springer-Verlag, 969-972, 1997.

ISSN: 2278 – 7798

[11] T.G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning 40, 139-157, 2000.

[12] WEK.A (Waikato Environment for Knowledge Analysis) Website(www.cs.waikato.ac.nz/ml/weka).

[13] Y. Sun, M. S. Kamel, A. K.Wong, and Y.Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recog.*, vol. 40, no. 12,pp. 3358–3378, 2007.

[14] Elsayad, A. M. (2010). Predicting the severity of breast masses with ensemble of Bayesian classifiers. Journal of Computer Science, 6(5), 576-584.

[15] S. J. Stolfo , D. W. Fan , W. Lee , A. L. Prodromidis and P. K. Chan "Credit card fraud detection using meta-learning: Issues and initial results", *Proc. AAAI Workshop AI Approaches Fraud Detection Risk Manage.*, pp.83 -90 1997.

[16] A.clntyre and M. Heywood "Multiobjective competitive coevolution for efficient GP classifier problem decomposition", *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, pp.1930 -1937 2007

[17] X.-Y. Liu , J. Wu and Z.-H. Zhou "Exploratory undersampling for class-imbalance learning", *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 39, no. 2, pp.539 -550 2009

[18] S. Wang , K. Tang and X. Yao "Diversity exploration and negative correlation learning on imbalanced data sets", *Proc. Int. Joint Conf. Neural Netw.*, pp.3259 -3266 2009

ISSN: 2278 – 7798