Volume 4, Issue 4, April 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Prediction for Classification of Highly Imbalanced Medical Dataset Using Databoost.IM with SVM

K.Lokanayaki
Assistant Professor
Spurthy Group of Institutions
Bangalore, India

Dr.A.Malathi
Assistant Professor
Govt. Arts College
Coimbatore, India

Abstract: Recently, Class imbalance problems have growing interest because of their classification difficulty caused by the imbalanced class distributions. In particular, many ensemble learning and machine learning methods have been proposed for classification of imbalance problem. However, these methods producing poor predictive accuracy of classification for two-class imbalanced dataset. In this paper, we propose a new approach that combines an ensemble-based learning algorithm (DataBoost.IM) with Machine learning algorithm (SVM) to improve the predictive power of classifiers for imbalanced Liver data sets consisting of two classes. In the DataBoost.IM–SVM method identified accuracy of both the majority and minority classes from imbalanced liver datasets during execution. This method was evaluated by F-measures, G-mean and overall accuracy, against imbalanced data sets. Our results compares with other existing algorithm for imbalanced Liver data set.

Keywords: Data mining, Imbalanced data sets, Ensembles of classifiers, SVM.

I. INTRODUCTION

The classification algorithm generally gives more important to classify for the imbalanced dataset. Process of adding new sample in existing is known as over-sampling and process of removing a sample known as under-sampling. For example in medical diagnosis in case of cancerous cell detection, misclassifying non-cancerous cells may leads to some additional clinical testing but misclassifying cancerous cells leads to very serious health risks. However in classification problems with imbalanced dataset, the minority class are more likely to be misclassified than the majority class, due to their design principles, optimize the overall classification accuracy produced by the machine learning algorithms which results in misclassification minority classes [1].

Most of the researchers found under-sampled examples of the majority class [5]; Ling and Li over-sampled examples of the minority class [3]. Especially, many authors proposed the majority class and the minority class for classification. Chawla et al. over-sampled the minority class and under-sampled the majority class [2]; Most of the learning algorithms aim to find a model with high prediction accuracy and a good generalization capability. Applying an algorithm alone is not good idea because size of data and class imbalance ratio is high and hence a new technique i.e. the combination of sampling method with algorithm is used [12].

Some authors evaluated boosting algorithms and ensemble learning algorithms to classify rare classes [6,8,9]; and Chawla et al. combined boosting and synthetic data to improve the prediction of the minority class [7] .Ensembles of classifiers consist of a set of individually trained classifiers whose predictions are combined to classify new instances [8, 9]. In particular, boosting is an ensemble method where the performance of weak classifiers is improved by focusing on hard examples which are difficult to classify. Boosting produces 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 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 performance is poor. It 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]. In the next section, we devoted a related work. In Section 3, we recall the basics of the DATABOOST.IM algorithm and SVM algorithm and also we introduce hybrid DataBoost.IM with SVM and give a detailed explanation of its novelties. An experimental evaluation is presented in Section 4. In Section 5 shown the result of DataBoost.IM with SVM significantly outperforms as well as other classifiers such as DataDoost.IM and EasyEnsemple in terms of classification accuracy. Section 6 concluding the paper.

II. RELATED WORKS

Many researches on the imbalanced data problem have been focused on several major groups of techniques. The popular method to solve imbalanced data problem balances the number of training examples among majority class and minority class. However, in oversampling techniques (minority), and undersampling (majority) techniques. These

© 2014, IJARCSSE All Rights Reserved

techniques have the problem of over generalization is largely attributed to the way in which synthetic samples are created (SMOTE algorithm) in oversampling [25] and the problem of removes the majority instances and find distance of farthest minority class from the decision boundaries (NearMiss-2,NearMiss-3) in undersampling[26]. Finally, the combination of preprocessing of instances with data cleaning two techniques (minority and majority) used to preprocessing of instances with classified the dataset (SMOTE with ENN and SMOTE with Tomek links) [9]. This technique is also present in a wrapper technique introduced in [28] that defines the best percentage to perform both undersampling and oversampling.

To improve classification accuracy for imbalanced dataset, proposed for Boosting method. Some authors proposed combine with boosting and SVM method for produced very effectively in the presence of imbalanced data [4]. Data sampling has received much attention in data mining related to class imbalance problem. Data sampling tries to overcome imbalanced class distributions problem by adding samples to or removing sampling from the data set [13]. This method improves the classification accuracy of minority class but, because of infinite data streams and continuous concept drifting, this method cannot suitable for skewed data stream classification. Most existing imbalance learning techniques are only designed for two- class problem.

Ensemble classifier is also developing for produce possible solution to the class imbalance problem among researchers. In [16] introduced based on ensemble methods SMOTEBoost and MSMOTEBoost for normalized synthetic example of oversampling. These methods also calculated total number of examples in the new dataset. The RUSBoost method developed for removes examples from the majority class of undersampling and found total sum of weights in new dataset [17]. New hybrid DataBoost.IM approach invented for identifies hard examples and then carries out a rebalance process in both classes of imbalanced dataset. This approach combines AdaBoost.M1 algorithm with a data generation method [14]. Finally introduced EasyEnsemble and BalanceCascade of hybrid Ensemble for adding instances and removing instances in a dataset. These approaches combine both bagging and boosting algorithms. These algorithms used for classifier in parallel and and works in a supervised manner. EasyEnsemble drive from UnderBagging and BalanceCascade derive from AdaBoost algorithm. However, these learning methods highly depend on the original classification method and lack of generality. This relationship among these things is complex and task- and method specific. In this way, we focus on improving the predictions of both the minority and majority classes using a new approach ensemble-based learning algorithm (DataBoost.IM) with Machine learning algorithm (SVM) for imbalanced dataset.

III. DATABOOST.IM ALGORITHM

This algorithm combines boosting, an ensemble-based learning algorithm, with data generation developed by Hongyu Guo and Herna L Viktor in 2004 [16].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 classes are rebalanced to improve the learning algorithms of majority class and minority class.

Algorithm Databoost-IM

Input: Sequence of *m* examples $(x1 , y1),...,(xm , ym)$ with labels $yi \in Y =\{ 1,...,k \}$

Integer ***T*** specifying number of iterations

Initialize $D_1(i) =1 / m$ for all i .

Do for $t = 1, 2,\ldots$ ***T***

1. Identify hard examples from the original data set for different classes
2. Generate synthetic data to balance the training knowledge of different classes
3. Add synthetic data to the original training set to form a new training data set
4. Update and balance the total weights of the different classes in the new training data set
5. Get back a hypothesis $h_t : X \rightarrow Y$.
6. Calculate the error of $h_t : \epsilon_t = \Sigma D_t(i)$ if $\epsilon t > 1/2$, then set $\boldsymbol{T} = t - 1$ and abort loop.
7. Set $\beta_t = \epsilon_t / (1 - \epsilon_t)$.
8. Update distribution

Output the final hypothesis

The above algorithm each example of the original training set is assigned an equal weight. The original training set is used to train the first classifier of the DataBoost-IM ensembles. Secondly, the hard examples (so-called seed examples) are identified and for each of these seed examples, a set of synthetic examples is generated. During the third stage of the algorithm, the synthetic examples are added to the original training set and the class distribution and the total weights of different classes are rebalanced. The second and third stages of the DataBoost-IM algorithm are re-executed until reaching a user specified number of iterations or the current component classifier's error rate is worse than a threshold value.

IV. SUPPORT VECTOR MACHINE (SVM)

The Support Vector Machine (SVM) is a machine learning algorithm introduced by Boser, Guyon, and Vapnik in 1992 for classification [18].It also developed binary classification problem for producing a high-accuracy classifier on imbalanced data [19][20]. In this paper,we are using SVM Soft Margin method for produce classification accuracy of majority class and minority class.In this method splits all the example of both class [21]. This function calculated by the following objective function

© 2014, IJARCSSE All Rights Reserved

$$\arg\min_{\mathbf{w},\xi,b}\left\{\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{n}\xi_i\right\}$$

_____ (1)

subject to (for any $i=1,\ldots n$)

$$y_i(\mathbf{w}\cdot\mathbf{x_i}-b)\geq 1-\xi_i,\qquad \xi_i\geq 0$$

This constraint in (1) along with the objective of minimizing $\|\mathbf{w}\|$ can be solved using Lagrange multipliers as done above. One has then to solve the following problem:

$$\arg\min_{\mathbf{w},\xi,b}\max_{\boldsymbol{\alpha},\boldsymbol{\beta}}\left\{\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i[y_i(\mathbf{w}\cdot\mathbf{x_i}-b)-1+\xi_i]-\sum_{i=1}^{n}\beta_i\xi_i\right\}$$

with $\alpha_i,\beta_i\geq 0$.

_____ (2)

. Maximize (in α_i)

$$\tilde{L}(\alpha)=\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j)$$

_____ (3)

subject to (for any $i=1,\ldots,n$)

$$0\leq\alpha_i\leq C,\qquad \text{and}\qquad \sum_{i=1}^{n}\alpha_i y_i=0.$$

This constraint in (2) and (3) used to find minority and majority classes. It also used to reduce the effect of outliers on the classifier.

V. PROPOSED ALGORITHM DATABOOST.IM WITH SVM

This paper presents a combine with DataBoost.IM and SVM based on Ensemble Boosting method and machine learning algorithm designed for imbalanced data classification. The proposed method to overcome the shortages of over-sampling and under-sampling and improves classification precision on the basis of maximizing data balance. DataBoost algorithm [16] according to the ratio of imbalanced samples, and integrates code generation of sub-classifiers into a classifier. Boost and code generation method can be used in conjunction with many other learning algorithms to improve their performance. In this way, the proposed method uses the minority class information, and also finds the information of the majority class.

Suppose that an imbalanced dataset contains m *ex*amples from the majority class and n labels from the minority class where *n>m*. First, the DataBoost.IM-SVM method divides training data set into *m* equivalent subsets, where *m* is greater than or equal to i. Then we add examples, which the results are different in two-class, to candidate data set. It is difficult to decide the category of these examples. So, these examples probably include abundant information. Last, we integrate two selected subsets into new training datasets, train and get a classifier using SVM method. Experiments of this paper show the DataBoost.IM-SVM method can get comprehensive classification information when the value of *m.*

Based on description above, the proposed DataBoost.IM-SVM method is described as follows:

Algorithm DataBoost.IM-SVM

Input: Sequence of *m* examples *(x1 , y1),...,(xm , ym)* with labels *yi ϵY ={ 1,...,k }*

Integer ***T*** specifying number of iterations

Initialize $D_1(i)=1/m$ for all *i* .

Do for $t = 1, 2,\ldots$ ***T***

1. Identify hard examples from the original data set for different classes
2. Generate synthetic data to balance the training knowledge of different classes
3. Add synthetic data to the original training set to form a new training data set
4. Update and balance the total weights of the different classes in the new training data set
5. Get back a hypothesis $h_t : X \rightarrow Y$.
6. Calculate the error of h_t: $\epsilon_t = \Sigma D_t(i)$ if $\epsilon t > 1/2$, then set $\boldsymbol{T} = t - 1$ and abort loop.
7. Set $\beta_t = \epsilon_t / (1` - \epsilon_t)$.
8. Update distribution
9. Calculate using (2)
10. Calculate using (3)
11. Implement (1) and (2) in function (1)
12. Repeat

Until less than termination condition

Output the final hypothesis

© 2014, IJARCSSE All Rights Reserved

VI. EVALUATION MEASURES

Accuracy is an important evaluation metric for assessing the classification performance and guiding the classifier modeling. In this section, we present the results obtained by the experiments carried out in this research. DataBoost.IM–SVM method was evaluated by F-measures, *G*-mean and overall accuracy, against imbalanced data sets. Our experiments with other DataBoost.IM [14], EasyEnsample[15] existing algorithm for imbalanced Liver data set.

TABLE I

True positive rate *TPrate* = TP/(TP+FN)
TPrate = % of positive cases classified correctly which belong to the positive class.
False positive rate *FPrate* = FP/(FP+TN)
FPrate = % of negative cases misclassified which belong to the positive class.

Several measures have been developed to deal with the classification problem with the class imbalance, including F-measure, G-mean, and AUC [22]. Given the number of true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs), we can obtain the confusion matrix presented in Table I after a classification process. We can also define several common measures. The TP rate and FP rate defined by Table II.

TABLE II

	Predicted Negative	Predicted Positive
Actual Negative	*TN* (the number of True Negatives)	*FP*(the number of False Positives)
Actual Positive	*FN* (the number of False Negatives)	*TP*(the number of True Positives)

Based on these measures, other measures have been presented, such as F-measure and G-mean. F-measure is often used in the fields of information retrieval and machine learning for measuring search, document classification, and query classification performance. *F*-measure considers both the precision *P* and the recall *R* to compute the score. It can be interpreted as a weighted average of the precision and recall as follows:

$$\text{F-Measure} = 2.\frac{\text{precision.recall}}{\text{precision}+\text{recall}}$$

Another criteria used to evaluate a classifier's performance on skew data is the G-mean. The G-mean is defined as

$$\text{G-Mean}=\sqrt{Positive\ Accuracy \times Negative\ Accuracy}$$

G-mean is defined by two parameters called Positive Accuracy (sensitivity) and Negative Accuracy specificity). Positive Accuracy shows the performance of the positive class and Negative Accuracy shows the performance of the negative class. G-mean measures the balanced performance of a learning algorithm between these two classes.

VII. EVALUATION RESULTS

In our experiments, we used imbalanced liver datasets to test the performance of the proposed method. This analyzes done on UCI Machine Learning Repository [27]. We take the minority class as the target class and all the other categories as majority class. The results of evaluating the performance of the DataBoost.IM–SVM algorithm, in comparison with the DataBoost.IM [14], EasyEnsample[15] which has become a de facto standard against which new algorithms are being judged.

© 2014, IJARCSSE All Rights Reserved

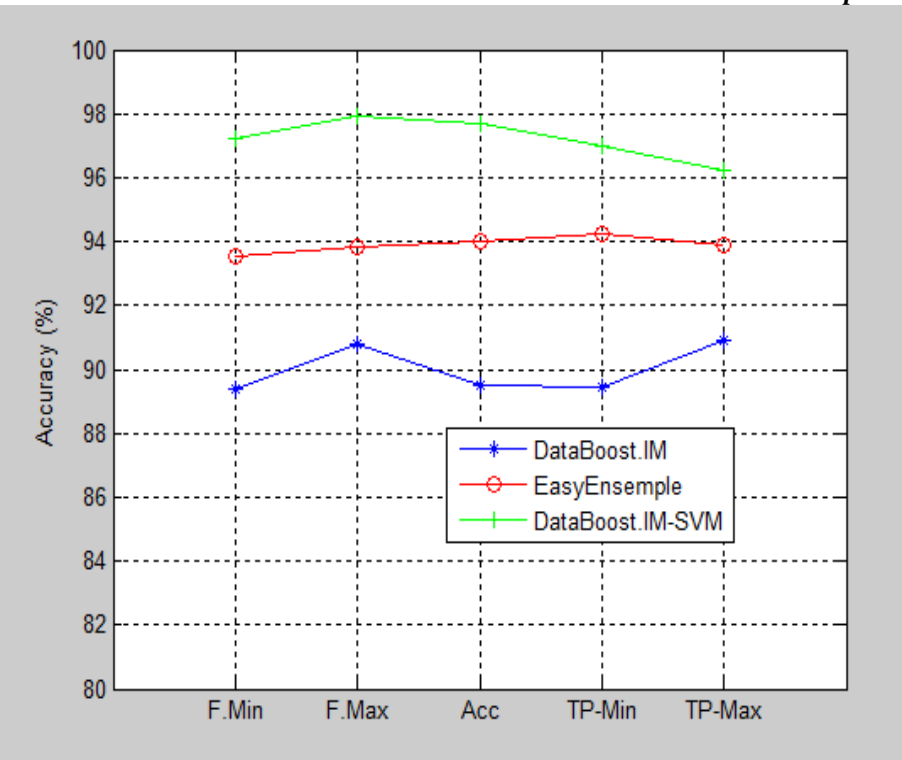

Fig. 1. F-measure and accuracy of the compared methods

Figure 1 shows the average F-measure values and accuracy of the compared methods. The results show that DataBoost.IM–SVM method has higher F-measure and accuracy than other compared methods on liver imbalanced datasets.

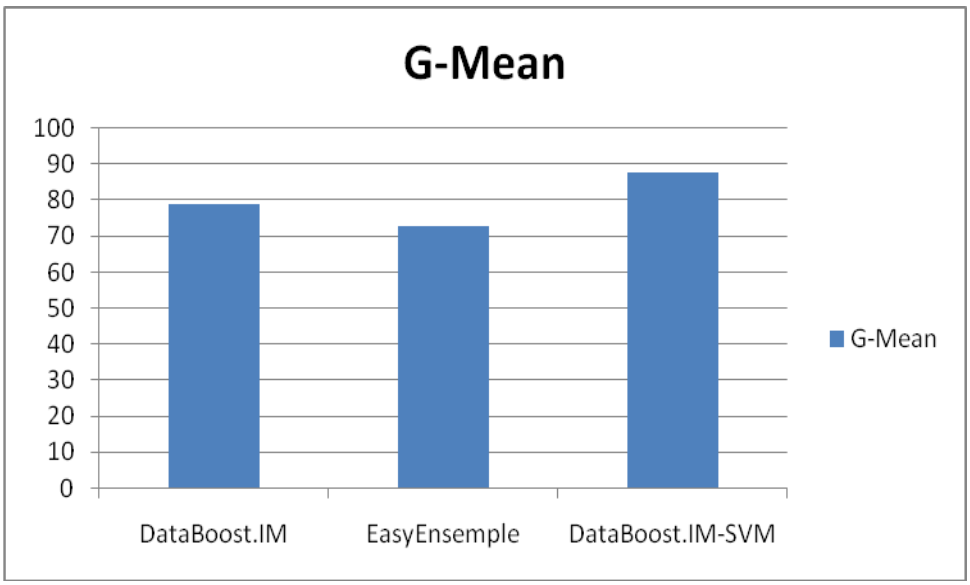

Fig. 2. *G-mean* of the compared methods

The average *G*-mean values of the compared methods are summarized in figure 2. The results show that DataBoost.IM–SVM has higher *G*-mean than other compared methods on imbalanced liver datasets.

VIII. CONCLUSION

Experimental results on imbalanced Liver dataset demonstrate that proposed DataBoost.IM-SVM performs better than other approaches of using component classifiers such as: DataBoost.IM and Easy Ensemble. Besides these, it is found that DataBoost.IM-SVM demonstrates good performance on imbalanced classification problems. The results indicate the DataBoost.IM-SVM approach performs well against imbalanced data sets. The DataBoost.IM-SVM algorithm achieved comparable and slightly better predictions in terms of the G-mean and F-Measures metrics, against both the minority and majority classes. Our future research work will be used in the frame of multi class learning problems and cost based learning problems.

© 2014, IJARCSSE All Rights Reserved

REFERENCE

[1] Yuchun Tang, Yan-Qing Zhang, Nitesh V. Chawla, and Sven Krasser "Correspondence SVMs Modeling for Highly Imbalanced Classification" IEEE Transactions On Systems,Man, And Cybernetics—Part B: Cybernetics, Vol. 39, No. 1, February 2009.

[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]. Nitesh V. Chawla, Nathalie Japkowicz, Aleksander Ko lcz "Special Issue on Learning from Imbalanced Data Sets" Volume 6, Issue 1 - Page 1-6.

[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] N. Chawla, A. Lazarevic, L. Hall and K. Bowyer. SMOTEBoost: improving prediction of the minority class in boosting. 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia , 107-119, 2003.

[8] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. the Proceedings of the Thirteenth International Conference on Machine Learning, Bari, Italy, 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.

[11] T.G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting,

[12]Peng Liu, Lijun Cai, Yong Wang, Longbo Zhang "Classifying Skewed Data Streams Based on Reusing Data" International Conference on Computer Application and System Modeling (ICCASM 2010).

[13] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance"IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 40, No. 1, January 2010.

[14] C. Seiffert, T. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "Rusboost: A hybrid approach to alleviating class imbalance," *IEEE Trans. Syst.,Man, Cybern. A, Syst., Humans*, vol. 40, no. 1, pp. 185–197, Jan. 2010.

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

[16] 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.

[17] J. Van Hulse, T. Khoshgoftaar, and A. Napolitano, "An empirical comparison of repetitive undersampling techniques," in *Proc. IEEE Int. Conf.Inf. Reuse Integr.*, 2009, pp. 29–34.

[18] Vapnik, V. N. (1995). The nature of statistical learning theory. New York:Springer.

[19] Platt C J. Fast training of support vector machines using sequential minimal optimization Source. In Advances in kernel method:support vector learning. 1999: 185-208.

[20] Abe S. Support Vector Machines for Pattern Classification. London: Springer-Verlag; 2006.

[21] Cortes, Corinna; Vladimir Vapnik (1995). "Support-Vector Networks". *Machine Learning* **20**: 273–297.

[22] S.Wang and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models," in *IEEE Symp. Comput. Intell. Data Mining*, 2009, pp. 324–331

[23] Bradley, A. P. (1997). The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition, 30(6),* 1145–1159.

[24] Swets, J. **(1988).** Measuring the Accuracy of Diagnostic Systems. *Science,***240: 1285-1293.**

[25] Wang, B. X., Japkowicz, N., 2004. Imbalanced data set learning with syn-thetic samples. In: Proceedings of the IRIS Machine Learning Workshop.

[26] Tang, Y., Zhang, Y.-Q., Chawla, N. V., Kresser, S., 2009. SVMs modeling for highly imbalanced classification. IEEE Transactions on Systems, Man and Cybernetics, Part B 39 (1), 281–288.

[27] Blake, C., Merz, C.: UCI Repository of Machine Learning Databases. Department of Information and Computer Sciences, University of California, Irvine, CA, USA (1998),

© 2014, IJARCSSE All Rights Reserved