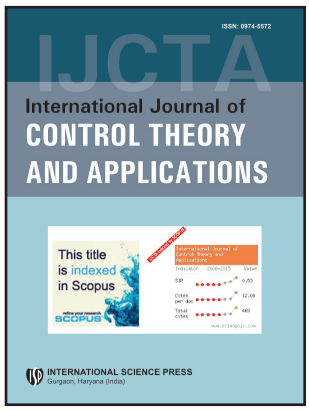

International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 12 • 2017

A Hybrid Ensemple Based Algorithm for Imbalanced Dataset Classification

K. Lokanayaki[a] and A. Malathi[b]

[a]*Assistant Professor, Department of Computer Science, St. Francis De Sales College, Bangalore*
[b]*Assistant Professor, PG and Research, Department of Computer Science, Government Arts College, Coimbatore*

Abstract: State-of-Art machine-learning methodologies are employed for classification imbalance in the last few years and have attained enormous popularity. Particularly, several ensemble learning and boost techniques have been introduced for the classification of imbalance problem. But these algorithms can be improved for developing high predictive accuracy in the classification for two class imbalanced dataset. In this research work, a novel algorithm combining an ensemble-based learning algorithm (DataBoost-IM) with Machine learning algorithm (SVM) is proposed for improving the predictive power of classifiers for imbalanced Liver cancer cell dataset containing two classes. In the DataBoost-IM with SVM algorithm identified accuracy of both the majority and minority classes from imbalanced liver cancer cell dataset during execution. By integrating DataBoost-IM with SVM, a better performance is achieved in comparison with DataBoost-IM and EasyEnsemple algorithms. The newly introduced DataBoost-IM with SVM classification algorithm is tested making use of the liver cancer cell dataset, providing an accuracy level of 87.3%, which is 85.4% and 86.2% more compared to the DataBoost-IM and EasyEnsemple algorithms.

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

1. EXISTING SYSTEM

The classification algorithm usually gives more importance for the classification for the imbalanced dataset. The process of addition of new sample to the already existing is referred to as over-sampling and the process of eliminating a sample is called as under-sampling. But in classification problems having imbalanced dataset, the minority class have more possibility to be misclassified compared to the majority class, owing to their design principles, which optimizes the overall classification accuracy obtained from the machine learning algorithms resulting in misclassification minority classes [1].

Many of the researchers have discovered under-sampled examples of the majority class [5] and the over-sampled examples of the minority class [3]. Specifically, several authors have introduced the majority class and the minority class for classification [2]. Few authors have assessed the boosting algorithms and ensemble learning algorithms for the classification of rare classes [6, 8, 9] and have integrated the boosting and synthetic data for

improving the prediction of the minority class [7]. Ensembles of classifiers comprise of a set of classifiers that are individually trained whose predictions are united for classifying the new instances [8, 9]. The boosting is an ensemble algorithm in which the performance of weak classifiers is enhanced by concentrating on hard examples that are hard to be classified. Boosting generates a series of classifiers and then the outputs of these classifiers are integrated making use of weighted voting in the last prediction of the model [10].

In every step of the data, the training examples get re-weighted and are chosen depending on the performance of the previous classifiers in the training data. This training data generates a set of "easy" examples having low weights and another set of hard examples having high weights. It is accomplishing by focusing on the correct classification of the hard examples. The studies carried out recently have shown that the boosting algorithm applies to a wide spectrum of problems successfully [10, 11]. Nonetheless, many recent studies conducted in the literature has been applied with success to cancer datasets samples.

In order to resolve this problem, the dataset extraction process can be widely categorized into three important sub-steps (a) Text Binarization using Niblack (N) and Markov Random Fields (NMRF) model, (b) Text extraction employing Hidden Markov Model (HMM) and (c) Text recognition by Hidden Markov Model with Partial (HMMP). At first, all the medical image reports are filtered by making use of the binarization method, so as to eliminate any noises, in case any noise exists in image. Thereafter the extraction of the text regions are performed making use of HMM [27]. After this, all of the connected Components are extracted showing the text portion in the image and every non-text character component is then removed. Once done, Optical Character Recognition (OCR) applying HMMP is utilized for recognizing the components extracted.

2. RELATED WORKS

Data sampling has achieved great focus in data mining corresponding to class imbalance problem. Data sampling attempts to get over the imbalanced class distributions problem by the addition of samples or the removal of samples from the data set [13]. This technique enhances the classification accuracy of minority class but, due to infinite data streams and constant concept drifting, this technique is not appropriate for skewed data stream classification. Many of the available imbalance learning methods are only developed for two - class problem.

Ensemble classifier has also been developed for offering a probable solution to the class imbalance problem among researchers. In [16] proposed on the basis of the ensemble techniques SMOTE Boost and MSMOTE Boost for normalized synthetic example of over sampling. These techniques also computed the total number of examples present in the new dataset. The RUS Boost technique was designed to remove the examples from the majority class of under sampling and then got the total sum of weights in the new dataset [17].

New hybrid DataBoost. IM method was introduced for identifying the hard examples and then performs a rebalance process in both the classes of imbalanced dataset. This approach integrates the AdaBoost.M1 algorithm with a data generation technique [14]. At last, Easy Ensemble and Balance Cascade is introduced for hybrid Ensemble for the addition and removal of instances in a dataset. These approaches integrate both the bagging and boosting algorithms.

These algorithms are used for the classifier to work in parallel in a supervised manner. Easy Ensemble drive from Under Bagging and Balance Cascade are obtained from AdaBoost algorithm. Nonetheless, these learning techniques are greatly dependent on the original classification method and lack of generality. This relationship among these things is complex and task specific[15].

DataBoost-IM algorithm is a combination of boosting and an ensemble-based learning algorithm, along with data generation. This algorithm helps in identifying the hard examples and generates artificial examples for

class]16]. In this manner, focus has been on the improvement of the predictions accuracy of both the minority and majority classes making use of a novel approach of ensemble-based learning algorithm (DataBoost-IM) with Machine learning algorithm (SVM) for imbalanced dataset.

3. SUPPORT VECTOR MACHINE (SVM)

The Support Vector Machine (SVM) is a reliable classification machine learning algorithm proposed by Boser, Guyon, and Vapnik in 1992 [18]. It increases the predictive accuracy of a model with no over fitting of the binary classification issue. [19][20]. In this research work, SVM Soft Margin is utilized for providing a classification accuracy of majority class and minority class. In this algorithm every example of both classes [21] are split and computed by the objective function below:

$$\operatorname{argmin}\left\{\frac{1}{2}\|\mathrm{W}\|^2 + \mathrm{C}\sum_{i=0}^{n}\varepsilon_i \right. \tag{1}$$

subject to (for any $i = 1, ..., n$) $y_i(w \cdot x_i - b) \geq 1 - \beta \quad \varepsilon_i \geq 0,$

This rule in (1) in addition to the aim of reducing $\|\mathrm{W}\|$ can be resolved by making use of Lagrange multipliers as said above. Then the following problem has to be solved:

$$\operatorname{argminmax}\left\{\frac{1}{2}\|\mathrm{W}\|^2 + \mathrm{C}\sum_{i=0}^{n}\varepsilon_i - \sum_{i=0}^{n}\alpha_i[y_i(w\cdot x_i - b) - 1 + \varepsilon_i] - \sum_{i=0}^{n}\beta_i\varepsilon_i\right\}$$

with $$\alpha_i, \beta_i \geq 0 \tag{2}$$

Maximize (in α_i) $$\bar{\mathrm{L}}(\alpha) = \sum_{i=0}^{n}\alpha_i - \frac{1}{2}\sum_{i=0}^{n}\alpha_i\alpha_j y_i y_j k(x_i, x_j) \tag{3}$$

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

$$0 \leq \alpha_i \leq \mathrm{C} \text{ and } \sum_{i=0}^{n}\alpha_i y_i = 0$$

This constraint in (2) and (3) is utilized for finding the minority and majority classes. It is also utilized for reducing the impact of outliers on the classifier.

4. OVERVIEW OF FRAMEWORK

The framework is illustrated in Figure 1.

Dataset Description

It has become increasingly tedious to maintain and obtain the information in biomedical images in computer-based patient report systems. In this case also the fixed textual information like patient reports or patient lab reports are full with X-ray images, MRI scans, CT scans, and video streams. Efficient filtering for digital medical images[12]. It is tried to have a more human-like style of logical thinking in programming computers.

Fuzzy logic is exploited when answers do not have a distinct true Mahmud et. al., [12] also taken Bangla multi font characters recognized isolate and continuous printed characters segmentation. In [15] feed-forward neural network is utilized for the classification of recognition data. After the document binarization a top-downs segmentation approach is used. The first lines of the documents are found, then the words get extracted and at last word and label word are matched, then the index value are stored to database [25].

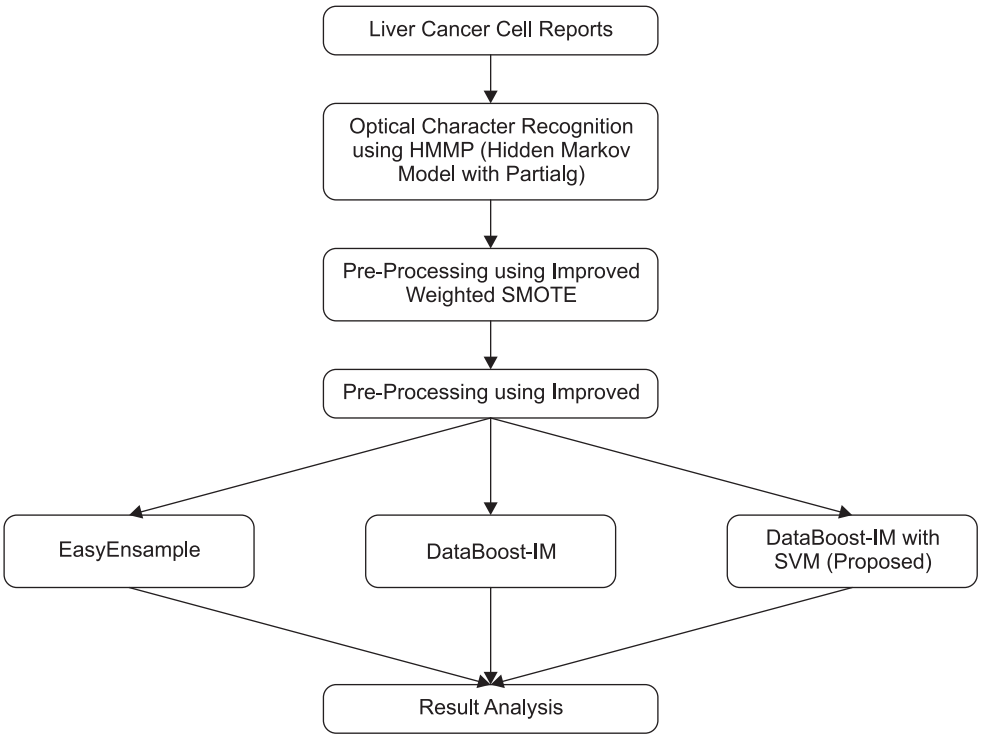

Figure 1: Overview of Framework

This search process needs a word models, a probable word lexicon or dictionary, and a statistical language model [25]. A report-based matching algorithm can do the recognition of the textual information from medical report images. It is utilized for converting the pixel information to text. At last, text will be stored onto database. In the system presented, HMM with Partial (HMMP) matching algorithm built for word level is employed.

5. PROPOSED ALGORITHM

1. Data Pre-Processing Using Improved Weighted SMOTE

The imbalanced dataset is a basic subject, which has developed in Machine Learning [22]. Imbalanced datasets is imposed with different problems. The initial problem is about a measure of performance. For the purpose of overcoming this issue, evaluation metrics are used to guide the learning process towards the necessary solution. The second problem is deficiency in data. When a class may have an extreme small amount of samples, then it becomes increasingly difficult to develop accurate decision boundaries between classes. The third problem is noise. Noisy data have a serious effect on minority classes compared to majority classes. Class imbalance is a problem faced often in Bioinformatics datasets. Unluckily, the minority class generally is also the class of interest.

One among the techniques for improving this condition is data sampling. To get over this issue, an Improved Weighted SMOTE algorithm was introduced. Initially, Improved Weighted SMOTE [23-24] fixes different sampling rates for dissimilar minority class samples. A novel setting of missing data imputation is done, which is the assignment of the missing data in data sets with heterogeneous attributes by providing both continuous and discrete data accordingly.

2. DataBoost-IM with SVM

This paper introduces DataBoost-IM with SVM based on Ensemble Boosting algorithm combined with machine learning algorithm developed for imbalanced data classification. The algorithm proposed is utilized for getting

over the setbacks of over-sampling and under-sampling and enhances the classification precision based on the maximization of the data balance. DataBoost algorithm [16] according to the ratio of imbalanced samples, and combines the code generation of the sub-classifiers into a classifier. Boost and code generation technique can be utilized along with several other learning algorithms to boost their performance. In this manner, the newly introduced algorithm makes use of the minority class information, and also discovers the information of the majority class.

Consider an imbalanced dataset containing m *ex*amples from the majority class and n labels from the minority class where $n > m$. At first, the DataBoost-IM with SVM algorithm partitions the training data set into m equivalent subsets, where m is greater than or equivalent to i. Then the examples are added, which the results are different in two-class, to candidate data set. It is hard to determine the group of these examples. Therefore, these examples possibly consist of enormous information. Finally, the two chosen subsets are integrated into new training datasets, tested and a classifier is got making use of SVM algorithm. Experiments conducted in this work indicate the DataBoost-IM with SVM algorithm can provide meaningful classification information if the value of m.

On the basis of description mentioned above, the new DataBoost-IM with SVM algorithm is explained as below:

Algorithm: DataBoost-IM with SVM

Input: Sequence of m examples $(x_1, y_1), ..., (x_m, y_m)$ with labels $y_i \in Y = \{1, ..., k\}$

Integer **T** specifying number of iterations

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

Do for: $t = 1, 2, ..., \mathbf{T}$

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 $h_t: \in_t = \Sigma D_t(i)$ if $\in_t > 1/2$, then set $\mathbf{T} = t - 1$ and abort loop.
7. Set $\beta_t = \in_t /(1 - \in_t)$.
8. Update distribution
9. Calculate using (2)
10. Calculate using (3)
11. Implement (2) and (3) in function (1)
12. Repeat

Until less than termination condition

Output the final hypothesis

The DataBoost-IM algorithm allocates equal weight for every example in the actual training set. The actual training set is utilized for training the first classifier of the DataBoost-IM ensembles. The identification of the

hard examples are done and a set of artificial examples get generated. Then the artificial examples are added to the actual training set and the class distribution and the total weights of various classes are rebalanced. Then the re-execution of the second and third stages of the DataBoost-IM algorithm are done till a user desired number of iterations is reached or the present component classifier's error rate is worse compared to a threshold value [16].

Support vector machine employs a nonlinear mapping for transforming the actual training data into a higher dimension. In this new dimension, it looks out for the linear optimal separating hyper plane. With the aid of a suitable non linear mapping to an adequately high dimension, data from two classes can always be isolated by a hyper plane. The SVM discovers this hyper plane making use of support vectors and margins.

6. EXPERIMENTAL RESULT AND ANALYSIS

Accuracy is an essential evaluation metric for the assessment of the classification performance and directing the classifier modeling. The experiment has been carried out making use of MATLAB tool for this combining approach. In order to assess the performance, various ensample algorithms such as DataBoost-IM[14], and EasyEnsemple [26] with a new algorithm DataBoost-IM with SVM a classification method are utilized in the liver cancer cell dataset.

Text based medical reports comprises of the information of liver cancer cell dataset. These medical reports will be transformed into liver cancer cell dataset samples. This dataset has 16500 liver patient records along with ten attributes that are fine-needle aspiration biopsy (FNAB) specimen's tests. The liver cell function tests are associated with 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 with this dataset. Among these dataset samples 70 % of the samples are utilized for the purpose of training and 30 % of the dataset is applied for testing. The experimental results have evaluated the performance of classifiers as indicated in Figure 2 and it is shown that the new DataBoost-IM with SVM algorithm provides greater accuracy.

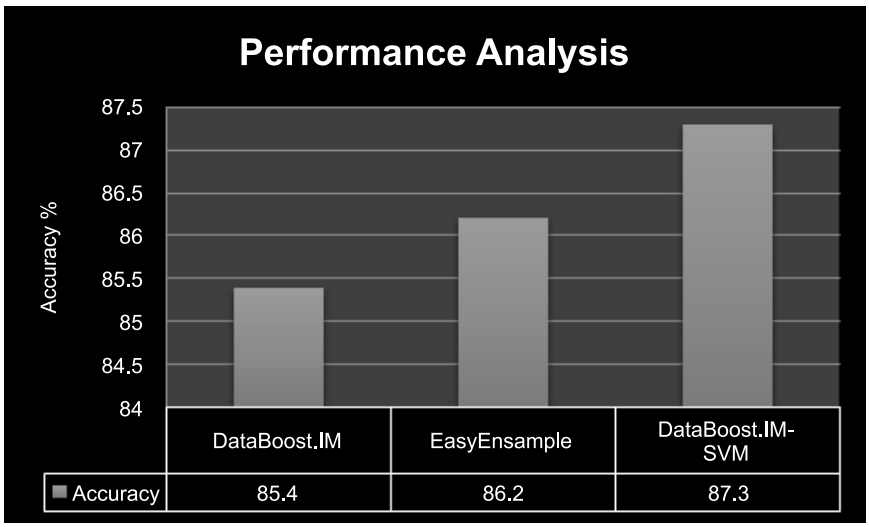

Figure 2: Performance of Proposed Algorithm

7. CONCLUSION

This paper proposed DataBoost-IM with SVM for the imbalanced liver cell dataset and provides the efficient extraction of the hard samples from the minority and majority classes. Furthermore, classified whether cancer

cells or not using Support Vector Machine (SVM). The experimental results got from the new algorithm indicate that greater prediction accuracy is obtained by employing DataBoost-IM with SVM when compared to the other available algorithms. The future research work will be utilized in the frame of multi class learning challenges and cost based learning issues.

REFERENCES

[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] J.U. Mahmud, M.F. Raihan and C.M. Rahman, "A Complete OCR System for continuous Bengali Character", TENCON 2003, Conference on Convergent Technologies for Asia-Pacific Region, 15-17 Oct. 2003.

[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] A.O.M. Asaduzzaman et. al., "Printed bangla text recognition using artificial neural network with heuristic method," Proceedings of International Conference on Computer and Information Technology, 2002, pp. 27-28.

[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 under sampling 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 CJ. 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] Chawla NV., Japowicz N, and Kolcz A. "Editorial: Special issue on learning from imbalanced data sets." SIGKDD Explorations 6, No. 1 (2004): 1-6.

[23] "Challenges and Surveys in Key Management and Authentication Scheme for Wireless Sensor Networks" in Abstract of Emerging Trends in Scientific Research 2014-2015.

[24] http://econpapers.repec.org/article/pkpabetsr/ Impact Factor: 0.119.

[25] "Biologically Inspired Intelligent Robots Using Artificial Muscles", International Journal of pharma and bio sciences, Impact Factor = 5.121(scopus indexed).

[26] Victor Marti, Horst Bunke, "Using a Statistical Language Model to Improve the Performance of an HMM-Based Cursive Handwriting Recognition System" International Journal of Pattern Recognition and Artificial Intelligence, Aug-2001, 15(01), 65-90.

[27] 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.

[28] K. Lokanayaki, A. Malathi, "Recognition Text for Liver Cell Lab Reports Based on Hybrid HMM Approach", International Journal Of Engineering And Computer Science (IJECS), Volume 4, Issue 6, June 2015, ISSN : 2319-7242.