2nd International Conference on Science, Technology and Management
University of Delhi (DU), Conference Centre, New Delhi (India)
27 September 2015, www.conferenceworld.in, www.icstmdu.com
978-81-931039-6-8

# AN EFFICIENT APPROACH FOR BALANCING IMBALANCE CLASS

**Ms. Barkha R. Hadke[1], Prof.Vikrant Chole[2]**

[1] *Department of Computer Science and Engineering,*

*G. H. Raisoni College of Engineering and Technology Nagpur, (India)*

**ABSTRACT**

*The classification of dataset Imbalanced problem always occurs where it makes difficult for balancing data. In this paper, we used the classifier for classification with the resampling techniques and mainly focused on the minority class in which the over-sampling method is used to balance data. This will be achieving by applying SMOTE technology on minority class and generates new synthetic samples between the nearby samples from this class, but SMOTE having the overgeneralize of the minority class region as it does not reflect on division of other neighbors from the majority classes. Thus, we introduce a new generalization of SMOTE, called RaSMOTE, which is a combination of rapidly convergence and SMOTE technique. Finally, this new method will be compared with existing method and the result shows that the new RaSMOTE method improves evaluation measure for the minority class as well as increases the accuracy of the prediction of datasets.*

***Keywords: Classification, Imbalance Problem, SMOTEboost, RACOG, Wracog***

## I. INTRODUCTION

Imbalanced data is a common problem in classification. This problem occurred when number of samples represents with more and less number of samples in classes. Many techniques have been developed to tackle the problem of imbalanced training sets in supervised learning. The Imbalance Ratio (IR), defined by the number of instances in the majority class divided by the number of instances in the minority class, expresses to which extent a dataset is imbalanced: a dataset with IR equal to 1 is perfectly balanced, the higher the IR, the more imbalanced the dataset.

In this paper, we consider the use of a well-known and widely used technique to balance the training set before the learning phase, the "Synthetic Minority Oversampling Technique" (SMOTE) methodology. [1]

Our main contribution is to introduce a new preprocessing method using SMOTE to generate synthetic examples. We propose the elimination of any synthetic example belong to the minority class. We carried out experiments in order to show the increasing accuracy in comparison with existing techniques i.e. RACOG and wRACOG [2], using different data-sets from the UCI repository with high imbalance ratios. The performance measured is based on the significance of the results which shows the analysis of G-Mean, F-Measure and classification result.

In this paper is includes as follows. In Sect. 2, Related work, we introduce the imbalanced data-set problem, discuss the evaluation metric used in this work, describe some preprocessing techniques for imbalanced data-sets, In Sect. 3 we present the algorithm called SMOTE and KNN Classification. In Sect. 4 we introduce the

experimental study, that is, the performance of datasets and the experimental analysis are improve as compared to the other techniques in order to validate prediction accuracy of our proposal. In Sect. 5 we draw some conclusions about the completed study.

## II. RELATED WORK

We consider the binary-class imbalanced data sets, where there is only one positive (minority) class and one negative (majority) class. the maximum data contains in the majority class and less data contains in minority class. Many techniques for dealing with class imbalance have arisen as a result of research and are grouped into two categories [5]: those at the level of the learning algorithm and those that modify data distribution (data level).

The machine learning algorithms classify imbalance data into majority and minority and it dominates the classification error in minority class. Consequently, the testing data in minority class are misclassified added often than those in the majority class. To handle the imbalanced data problem it used many different techniques such as data level, algorithmic level, cost sensitive level, feature selection level, and ensemble level.

Class imbalance problems are exits in transaction dataset, medical diagnosis, science and engineering problem and so on. Some authors introduce classifiers and technique to improve the accuracy of prediction in the classes.

## III. METHODOLOGY

### 3.1 KNN Algorithm

The purpose of the k Nearest Neighbors (KNN) algorithm is to use a database in which the data points are separated into several separate classes to predict the classification of a new sample point. K-Nearest Neighbors (KNN) classification divides data into a test set and a training set. In testing set, for each tuple k-nearest samples are found by using Euclidian distance and the classification is done on majority and minority of the classes.

### 3.2 SMOTE Algorithm

SMOTE [2] this technique introduced by Chawla et al., this technique used on minority classes, it is a very popular over-sampling method. Its main idea is to construct new minority class samples for selecting a near minority class neighbor randomly. The new synthetic minority samples are created as follows:

- For the continuous features:
  - Take the difference between a feature vector (minority class sample) and one of its k nearest neighbors (minority class samples).
  - Multiply this difference by a random number between 0 and 1.
  - Add this difference to the feature value of the original feature vector, thus creating a new feature vector
- For the nominal features
  - Take majority vote between the feature vector under consideration and its k-nearest neighbors for the nominal feature value. In the case of a tie, choose at random.
  - Assign that value to the new synthetic minority class sample.

Using this technique, a new minority class sample created in the neighborhood of the minority class sample and this consideration becomes the shortcoming of SMOTE that is known as overgeneralization [4[. This problem

978-81-931039-6-8

blindly generalizes the regions of the minority class without regard to the majority class. This strategy is particularly problematic in the case of skewed class distribution where the minority class is very sparse with respect to the majority class. In such a case SMOTE generation of synthetic examples may increase the occurrence of overlapping between classes. So, some adaptive strategies have been proposed to overcome this limitation.

## IV. EXPERIMENTAL STUDY: RaSMOTE

This paper presents a combine with SMOTE and KNN classifier based on ensemble methods and machine learning algorithm designed for imbalanced data classification.

In this paper, we propose a technique i.e. **Ra**pidly Convergence on **S**ynthetic **M**inority **T**echnique RaSMOTE algorithm that combines the rapidly convergence procedure with Synthetic Minority Oversampling Technique (SMOTE). Thus the SMOTE and rapidly convergence technique are improving the prediction accuracy of the minority classes to the entire data set.

Our goal is to build better model the minority class in the data set, by providing the classifiers. We want to improve the overall accuracy of the ensemble by focusing on the difficult minority (positive) class cases; the goal is to improve our True Positives (TP).

In this section we compare RaSMOTE to well known preprocessing algorithms:

1. **Input:** Training dataset, Testing dataset
2. Separate the samples into minority and majority from training dataset.
3. Apply KNN algorithm on minority class:
   Knnclassify(New_testData,Training,Train_Class_DATA, 5,'euclidean','nearest')
4. Applying SMOTE Algorithm to generate the synthetic samples to balance the training knowledge of different classes
5. To avoide the overgenerilazation apply RaSMOTE
   - All Minority generated by SMOTE
   - Adding original data and new synthetic samples we get New training dataset
   - Find the centroid by using k-means
   - Put the threshold value to calculate to half of size of majority class
   - Calculate the similar samples from synthetic samples
   - Repeat step 3
6. Exit

## V. EVALUATION MEASURES

In a class imbalance problem, the confusion matrix [7] (shown in Table 1) records the results of correctly and incorrectly recognized examples of each class. Accuracy is an important evaluation metric for accessing the classification performance and guiding the classifier modeling. Thus the result of RaSMOTE technique is evaluated by performance metrics which includes F-Measures, G-Mean and overall accuracy against imbalanced dataset.

**Table 1 Confusion Matrices for imbalance problem**

| | Positive Prediction | Negative Prediction |
|---|---|---|
| Positive class | True Positive (TP) | False Negative (FN) |
| Negative class | False Positive (FP) | True Negative (TN) |

When working in imbalanced domains, there are more appropriate metrics to be considered instead of accuracy. Specifically, we can obtain four metrics from Table I to measure the classification performance of both, positive and negative, classes independently.

*1) True positive rate:* is the percentage of positive instances correctly classified.

$$TP\text{rate} = TP/(TP + FN)$$

2) *True negative rate:* is the percentage of negative instances correctly classified.

$$TN\text{rate} = TN/(FP + TN)$$

3) *False positive rate:* is the percentage of negative instances misclassified.

$$FP\text{rate} = FP/(FP + TN)$$

4) *False negative rate:* is the percentage of positive instances misclassified.

$$FN\text{rate} = FN/(TP + FN)$$

Classification accuracy is not sufficient as a standard performance measure. ROC analysis [8] and metrics such as *precision*, *recall* and *F-value* [9, 10] have been used to understand the performance of the learning algorithm on the minority class. To improve the recall of learning algorithm without sacrificing the precision this is calculated as follows;

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

The *F-value* incorporates both *precision* and *recall*, and the "goodness" of a learning algorithm for the minority class can be measured by the *F-value*. While ROC curves represent the trade-off between values of TP and FP, the F-value basically incorporates the relative effects/costs of *recall* and *precision* into a single number.

Thus, *F-value* may be defined as follows:

$$F\text{-}value = \frac{(1+\beta^2)\cdot Recall \cdot Precision}{\beta^2 \cdot Recall + Precision}$$

## VI. EVALUATION RESULTS

In our implementation, we used imbalance four dataset to test the performance of the proposed method. For this experiment, we use UCI Machine Learning Repository [6] which follows a 5-fold cross validation strategy: we divide the data in 5-folds and classify the instances of each fold as training data and testing data.

We focus on the minority class which contains less number of samples. The result of evaluating the performance of the RaSMOTE algorithm, in comparison with the SMOTE and previous Techniques, it OVERCOMES THE shortage of oversampling and improves the classification precision on the basis of maximizing data balance. The

comparison results of different dataset training algorithm over all dataset are shown in following table 2. Following table shows the Accuracy of Minority class by using KNN which shows the comparison between the before applying KNN classifier and after applying RaSMOTE algorithm:

**Table 2 Comparison between Accuracy of KNN Classifiers**

| DATASET | Before KNN | After KNN |
|---|---|---|
| Glass | 71.4286 | 76.7442 |
| Vehicle | 29.5455 | 36.3636 |
| Yeast | 66.667 | 81.8182 |
| Ecoli | 68.75 | 81.25 |

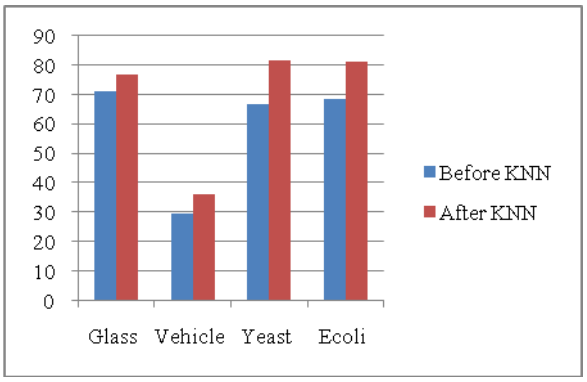


**Fig 1: Comparison between Accuracy of KNN**

## VII. CONCLUSION

The proposed RaSMOTE algorithm is based on the integration of the Rapidly Convergence and SMOTE algorithm within the KNN Classifiers. Experimental results from several imbalanced datasets indicate that the proposed RaSMOTE algorithm can result in better prediction of minority classes than SMOTE and previous Techniques. Data sets used in our experiments contained different degrees of imbalance and different sizes, thus providing a diverse test data. While convergence improves the predictive accuracy of classifiers by focusing on minority samples that belong to all the classes, the RaSMOTE combines the power of SMOTE in vastly improving the recall with rapidly convergence in improving the precision and with this algorithm it also improves the performance of a classifier only on the minority class examples.

## REFERENCES

[1] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) SMOTE: Synthetic minority over-sampling technique. J Artif Intell Res 16:321–357

[2] Barnan Das, Narayanan C. Krishnan And Diane J. Cook, Fellow, "RACOG AND WRACOG: Two Probabilistic Oversampling Techniques" IEEE Transaction On Knowledge And Data Engineering - Draft 1

[3] Feng Hu, Hang Li, Huabin Lou, Jin Dai, "Parallel Oversampling Algorithm Based on NRSBoundary-SMOTE", Journal of Information & Computational Science 11:13 (2014) 4655–4665 September 1, 2014.

[4] Tomasz Maciejewski and Jerzy Stefanowski, "Local Neighbourhood Extension of SMOTE for Mining Imbalanced Data", 978-1-4244-9925-0/11/$26.00 ©2011 IEEE.

[5] Chawla NV, Japkowicz N, Kolcz A (2004) Editorial: special issue on learning from imbalanced data sets. SIGKDD Explor 6(1):1–6

[6] Blake, C., Mrez,: UCI Repository of Machine learning Database. Department of information and Computer Science, University of California, Irvine, CA, USA(1998)

[7] G. M. Weiss and F. Provost, "Learning when training data are costly: The effect of class distribution on tree induction," *J. Artif. Intell. Res.*, vol. 19, pp. 315–354, 2003.

[8] F. Provost, T. Fawcett, Robust Classification for Imprecise Environments, *Machine Learning*, vol. 42/3, pp. 203-231, 2001.

[9] M. Buckland, F. Gey, The Relationship Between Recall and Precision, *Journal of the American Society for Information Science*, 45(1):12--19, 1994.

[10] M. Joshi, V. Kumar, R. Agarwal, Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements, *First IEEE International Conference on Data Mining*, San Jose, CA, 2001.

[11] Juanli Hu, Jiabin Deng, Mingxiang Sui, "A New Approach For Decision Tree Based On Principal Component Analysis", Proceedings Of Conference On Computational Intelligence And Software Engineering, Page No:1-4, 2009.

[12] K.P.N.V.Satyasree, Dr. J. V. R. Murthy, "An Exhaustive Literature Review On Class Imbalance Problem" Ijettcs,Volume 2, Issue 3, May – June 2013

[13] N. V. Chawla, K.W. Bowyer, L. O. Hall, Andw. P. Kegelmeyer, "Smote synthetic Minority Over-Sampling Technique," J. Artif. Intell. Res., Vol. 16,Pp. 321–357, 2002.

[14] A Review On Ensembles For The Class Imbalance Problem: Bagging-, Boosting-, And Hybrid-Based Approaches, IEEE Transactions On Systems, Man, And Cybernetics

[15] R. LazaEt Al., "Evaluating The Effect Of Unbalanced Data In Biomedical Document Classification," Journal Of Integrative Bioinformatics, Vol. 8, No. 3, Pp. 177, 2011 Sep, 2011.

[16] C. Drummond, R. C. Holte. C "Decision Tree, Class Imbalance, And Cost Sensitivity: Why Under-Sampling Beats Over-Sampling, In: Workshop On Learning From Imbalanced Data Sets" Ii, International Conference On Machine Learning, 2003.

[17] Y. Freund and R. E. Schapire. "A Decision-Theoretic Generalization Of On-Line Learning And An Application To Boosting". Journal Of Computer And System Science, 55(1):119-139, 1997.

[18] N. V. Chawla, A. Lazarevic, L. O. Hall, And K. W. Bowyer. "Smoteboost: Improving Prediction Of The Minority Class In Boosting. In Knowledge Discovery In Databases" Pkdd 2003, Pp. 107–119, 2003.

[19] Dr.D.Ramyachitr, P.Manikandan "Imbalanced Dataset Classification And Solutions: A Review" International Journal Of Computing And Business Research (Ijcbr),Volume 5 Issue 4 July 2014

[20] Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. "Rusboost: A hybrid approach to alleviating class imbalance", IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 2010, 40(1):185–197.

[21] Dr. Ali Mirza Mahmood "An Overview of Class Imbalance Learning in Knowledge Discovery" Associate Professor, DMS SVH College of Engineering, Machilipatnam. Krishna University, Machilipatnam, Andhra Pradesh, India.

[22] Barnan Das, Narayanan C. Krishnan And Diane J. Cook, Fellow,"RACOG AND WRACOG: Two Probabilistic Oversampling Techniques" IEEE Transaction On Knowledge And Data Engineering - Draft 1

[23] Yun Zhai, Haifeng Sui, Changsheng Zhang "A New Over-sample Method Based on Distribution Density" JOURNAL OF COMPUTERS, VOL. 9, NO. 2, FEBRUARY 2014

[24] Yoav Freund Robert E. Schapire "Experiments with a New Boosting Algorithm" Machine Learning: Proceedings of the Thirteenth International Conference, 1996.

[25] E. Ramentol1, n. Verbiest, r. Bello, y. Caballero1, c. Cornelis and f. Herrera "Smote-Frst: A New Resampling Method Using Fuzzy Rough Set Theory" March 20, 2012

[26] Haibo He, Member, Ieee, And Edwardo A. Garcia "Learning From Imbalanced Data" IEEE Transactions On Knowledge And Data Engineering, Vol. 21, No. 9, September 2009

[27] Sunita Beniwal*, Jitender Arora "Classification and Feature Selection Techniques in Data Mining" International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 6, August – 2012, ISSN: 2278-0181