ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 11, November 2015**

# An Approach for Class Imbalance Using Oversampling Technique

Rinku T Hadke, Asst.Prof. Prashant Khobragade

Scholar Student, Dept of CSE, Rajiv Gandhi College of Engineering And Research, Nagpur, India.

Assistant Professor, Dept of CSE, Rajiv Gandhi College of Engineering and Research, Nagpur, India.

**ABSTRACT**: In data mining, Class imbalance problem encounters when one class having majority of data than other classes. Most of the techniques are concentrating on the class that having majority of sample while avoiding the minority class sample. The minority class data are those data which occur less frequently but are important. There are broadly two techniques used to solve this problem those techniques are under-sampling and oversampling. Under-sampling is a technique which randomly removes the majority class sample. The problem associated with under-sampling is that it removes some important data present in majority class. Oversampling is a technique which replicates the minority class samples. The minority class is not represented well which leads to high misclassification error.

To overcome this issue, the probabilistic approach is used which will balance datasets and classify it accurately. Probabilistic oversampling approaches are used to synthetically generating and strategically selecting new minority class samples. The proposed approaches use the joint probability distribution of data attributes and Gibbs sampling to generate new minority class samples. Rare class problem leads to misclassification of minority class sample.

**KEYWORDS:** Imbalanced class distribution, probabilistic oversampling, approximating joint probability distribution, Gibbs sampling

## I. INTRODUCTION

In many real time applications large amount of data is generated with skewed distribution. A data set said to be highly skewed or class imbalanced if sample from one class is in higher number than other. In imbalance data set the class having more number of instances is called as major class while the one having relatively less number of Instances are called as minor class. Applications such as medical diagnosis prediction of rare but important disease is very important than regular treatment. Similar situations are observed in other areas, such as detecting fraud in banking operations, detecting network intrusions, managing risk and predicting failures of technical equipment. In such situation most of the classifier are biased towards the major classes and hence show very poor classification rates on minor classes. It is also possible that classifier predicts everything as major class and ignores the minor class.

The classification techniques usually assume a balanced class distribution (i.e. there data in the class is equally distributed). Usually, a classifier performs well when the classification technique is applied to a dataset evenly distributed among different classes. But many real applications face the imbalanced class distribution problem. In this situation, the classification task imposes difficulties when the classes present in the training data are imbalanced. The imbalanced class distribution problem occurs when one class is represented by a large number of examples (majority class) while the other is represented by only a few (minority class). In this case, a classifier usually tends to predict that samples have the majority class and completely ignore the minority class. This is known as the class imbalance problem. The idea of the class imbalance problem where a minority class is represented by only 1% of the training data and 99% for majority class.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 11, November 2015**

In Class Imbalance, Problem are raised when one class having more samples than other classes.A dataset is imbalanced if the classification categories are not approximately equally represented. The level of imbalance (ratio of size of the majority class to minority class) can be as huge as 1:99. It is noteworthy that class imbalance is emerging as an important issue in designing classifiers. The classical classifiers of balance datasets cannot deal with the class-imbalance problem because they pay more attention to the majority class. The main drawback associated with majority class is loss of important information.

***IMBALANCE DATA PROBLEM***

The classification techniques usually assume a balanced class distribution (i.e. there data in the class is equally distributed). Usually, a classifier performs well when the classification technique is applied to a dataset evenly distributed among different classes. But many real applications face the imbalanced class distribution problem. In this situation, the classification task imposes difficulties when the classes present in the training data are imbalanced. The imbalanced class distribution problem occurs when one class is represented by a large number of examples (majority class) while the other is represented by only a few (minority class). In this case, a classifier usually tends to predict that samples have the majority class and completely ignore the minority class. This is known as the class imbalance problem. Figure 1.1 illustrates the idea of the class imbalance problem where a minority class is represented by only 1% of the training data and 99% for majority class.

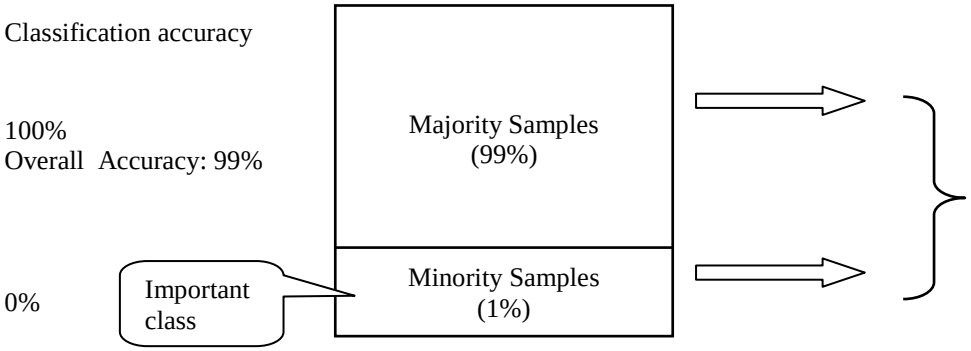


Fig 1:Distribution of Majority and Minority class

## II. RELATED WORK

In this paper [1] they introduce two probabilistic approaches, namely RACOG and wRACOG to synthetically generating and strategically selecting new minority class samples. The proposed approaches use the joint probability distribution of data attributes and Gibbs sampling to generate new minority class samples, While RACOG selects samples produced by the Gibbs sampler based on a predefined lag, wRACOG selects those samples that have the highest probability of being misclassified by the existing learning model.

This paper [2] define cost-free learning (CFL) formally in comparison with cost-sensitive learning (CSL). The main difference between them is that a CFL approach seeks optimal classification results without requiring any cost information, even in the class imbalance problem. Using the strategy can handle binary/multi-class classifications with/without abstaining. Significant features are observed from the new strategy. While the degree of class imbalance is changing, the proposed strategy is able to balance the errors and rejects accordingly and automatically.

This paper [3] presents a new hybrid sampling/boosting algorithm, called RUSBoost, for learning from skewed training data. This algorithm provides a simpler and faster alternative to SMOTEBoost, which is another algorithm that combines boosting and data sampling. This paper evaluates the performances of RUSBoost and SMOTEBoost, as well as their individual components (random under-sampling, synthetic minority oversam-pling technique, and AdaBoost).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 11, November 2015**

RUSBoost and SMOTEBoost both outperform the other procedures, and RUSBoost performs comparably to and often better than SMOTEBoost while being a simpler and faster technique.

In this paper [4] hey describe a method to solve the class imbalance problem by multi cluster-based majority under-sampling and random minority oversampling approach. Compared to under-sampling, cluster-based random undersampling can effectively avoid the important information loss of majority class and oversampling will helps to balance data i.e overcomes the drawback of under-sampling that it removes away many useful majority class samples.

This paper [5] gives data mining approaches that have been used in business purposes since its inception however, at present it is used successfully in new and emerging areas like education systems. In this paper, use of data mining approaches to predict students' final outcome, i.e., final grade in a particular course by overcoming the problem of imbalanced dataset. Several re-sampling techniques are given to balance the dataset so that to get better performance. Re-sampling techniques include SMOTE, RUS, ROS.

In this paper [6], when the class sizes are highly imbalanced, the standard algorithm tend to strongly favour the majority class and provide notably low detection of the minority class as a result.The method proposes an online fault detection algorithm based on incremental clustering. The algorithm accurately finds wafer faults even in severe class distribution skews and efficiently processes massive sensor data in terms of reductions in the required storage.

The class imbalance problem using under-sampling [7] has a drawback that it throws away important information in a majority class. To overcome this problem, this paper proposed a cluster based under-sampling method. This used a clustering algorithm that is performance guaranteed, named k-centers algorithm, which clusters the data in the majority class and selects a number of representative data in many proportions, and then combines them with all the data in the minority class as a training set.

In this paper [8], they improve the resampling strategy inside OOB (Oversampling based online bagging) and UOB (Under sampling based online bagging), and look into their performance in both static and dynamic data streams. They give the first comprehensive analysis of class imbalance in data streams, in terms of data distributions, imbalance rates and changes in class imbalance status. They find that UOB is better at recognizing minority-class examples in static data streams, and OOB is more robust against dynamic changes in class imbalance status.Then they propose two new ensemble methods that maintain both OOB and UOB with adaptive weights for final predictions, called WEOB1 and WEOB2. They are shown to possess the strength of OOB and UOB with good accuracy and robustness.

This paper [9] studies the challenges posed by the multiclass imbalance problems and investigates the generalization ability of some ensemble solutions, including their recently proposed algorithm AdaBoost.NC , with the aim of handling multiclass and imbalance effectively and directly.

This paper [10] gives Class imbalance problem which become greatest issue in data mining. Imbalance problem occur where one of the two classes having more sample than other classes. The most of algorithm are more focusing on classification of major sample while ignoring or misclassifying minority sample. The minority samples are those that rarely occur but very important. There are different methods available for classification of imbalance data set which is divided into three main categories, the algorithmic approach, data- pre-processing approach and feature selection approach. Each of this technique has their own advantages and disadvantages. In this paper systematic study of each approach is define which gives the right direction for research in class imbalance problem.

## III. PROPOSED SYSTEM

**Proposed Approach:**

The proposed approach is based on the idea of using the probability distribution of the minority class to generate new minority class training samples. This way can avoid the possibility of the synthetically generating training samples actually belonging to any other class in case of class overlap. In order to sample from the probability

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 11, November 2015**

distribution of the minority class, first need to estimate the distribution and then employ a mechanism to generate samples from that distribution. In the following, we apply Oversampling approach for strategically selecting minority class data sample that have the highest probability of being misclassified by existing learning model and generating new minority class sample to balance the imbalance distribution of data.

**Proposed Architecture:**

In this architecture create a classifier based on the input dataset. Run classifier and classify the dataset in binary class. In classification, decision tree classifier builds a model on training data and checks it with test data. The model is built with C4.5 decision tree classifier which deploys the classification task and gives correctly and incorrectly classified data. Then dataset is given to Clustering. In k-nn clustering, dataset is clustered in such a way that it clusters misclassified data. Select only those clusters which have minority samples that are misclassified. Then selected minority samples are given to probabilistic oversampling technique. Dataset which is obtained after by applying oversampling gives back to check whether the data set is balanced or not. The probabilistic oversampling technique handles the imbalanced datasets to get balanced datasets.

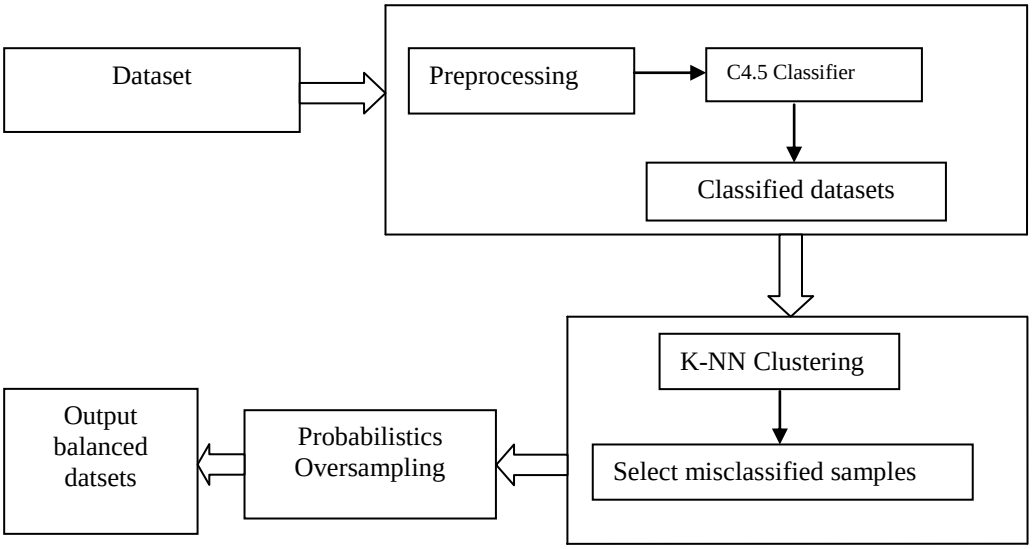


Fig 3: Proposed Architecture

## IV. RESULTS

It is expected that the proposed approach balance the datasets more accurately and reduce the misclassification rate of minority class samples. This approach improves the performance of classifiers in order to approve the accuracy of classifier.

## V. CONCLUSION AND FUTURE WORK

Thus this paper represents imbalanced classproblem and also gives an idea of classification of the imbalanced dataset. Also, this paper has given a survey of the problems of the imbalanced dataset. In this paper, several types of imbalance data handling techniques and solutions of the imbalance dataset are given. And finally we conclude that, the solution for solving the imbalanced dataset problem is the data level process. Because, the data level process provides better results by using the oversampling algorithm for pre-processing and for balancing proposed approach is useful. Thus

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 11, November 2015**

this paper might be useful for the researchers to know about the imbalance dataset problems and also its solutions.The result shows that clustered based minorityoversampling can improve the performance of classifiers for imbalanced datasets.

## REFERENCES

[1] Barman Das, Narayanan C. Krishnan, And Diane J. Cook, "Racog And Waco Two Probabilistic Oversampling Techniques" Ieee Transactions On Knowledge And Data Engineering, Vol. 27, No. 1, January 2015, Pp 222-232.
[2] Xiaowan Zhang And Bao-Gang Hu, "A New Strategy Of Cost-Free Learning In The Class Imbalance Problem" Ieee Transactions On Knowledge And Data Engineering, Vol. 26, No. 12, December 2014, Pp 2872-2881.
[3] Chris Seiffert, Taghi M. Khoshgoftaar, Member, Ieee, Jason Van Hulse, Member, Ieee, 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, Pp 185-194.
[4] Rushi Longadge, Snehlata S. Dongre, Latesh Malik "Multi-Cluster Based Approach For Skewed Data In Data Mining" E-Issn: 2278-0661, P-Issn: 2278-8727volume 12, Issue 6 (Jul. - Aug. 2013), Pp 66-73.
[5]Raisul Islam Rashu, Naheena Haq, Rashedur M Rahman "Data Mining Approaches To Predict Final Grade By Overcoming Class Imbalance Problem" 2014 17th International Conference On Computer And Information Technology (Iccit), Pp 215-222.
[6]Jueun Kwak, Taehyung Lee And Chang Ouk Kim "An Incremental Clustering-Based Fault Detection Algorithmfor Class-Imbalanced Process Data" Ieee Transactions On Semiconductor Manufacturing, Vol. 28, No. 3, Pp 212-220.
[7]Wattana Jindaluang And Varin Chouvatut,Sanpawat Kantabutra "Under-Sampling By Algorithm With Performance Guaranteed For Class-Imbalance Problem"2014 International Computer Science And Engineering Conference (Icsec), Pp 812-823.
[8]Shuo Wang, Member, Ieee, Leandro L. Minku, Member, Ieee, And Xin Yao, Fellow, Ieee"Resampling-Based Ensemble Methods For Online Class Imbalance Learning"Ieee Transactions On Knowledge And Data Engineering,Vol. 27,No. 5,May 2015
[9] Shuo Wang, Member, Xin Yao, "Multiclass Imbalance Problems: Analysis And Potential Solutions" Ieee Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 42, No. 4, August 2012, Pp 1356-1359.
[10] Mr.Rushi Longadge, Ms. Snehlata S. Dongre, Dr. Latesh Malik "Class Imbalance Problem In Data Mining: Review" International Journal Of Computer Science And Network (Ijcsn) Volume 2, Issue 1, February 2013.
[11] N. V. Chawla, L. O. Hall, K. W. Bowyer, And W. P. Kegelmeyer, "Smote: Synthetic Minority Oversampling Technique," J. Artif. Intell. Res., Vol. 16, Pp. 321–357, 2002.
[12]S. Hu, Y. Liang, L. Ma, And Y. He, "Msmote: Improving Classification Performance When Training Data Is Imbalanced," In Proc. 2nd Int.Workshop Comput. Sci. Eng., 2009, Vol. 2, Pp. 13–17.
[13] S. Kotsiantis, P. Pintelas, Mixture Of Expert Agents For Handling Imbalanced Data Sets, Annals Of Mathematics, Computing & Teleinformatics, Vol 1, No 1 (46-55), 2003.
[14] M. Kubat And S. Matwin, "Addressing The Curse Of Imbalanced Training Sets: One-Sided Selection," Proc. 14th Int'l Conf. Machine Learning, Pp. 179-186, 1997.
[15] Data Mining Concepts And Technique-Jiawei Han & Micheline Kamber Harcourt(Book)

## BIOGRAPHY

**Asst.ProfPrashantKhobragade** Assistant Professor in the CSEDepartment, Rajiv Gandhi College of Engineering and Research, Nagpur, India. He has received Master of Technology (Mtech.) degree from RTMNU, India. His research interests are Data miningetc.