A Cluster Based Analysis for Imbalanced Data using SMOTE and Cluster-Based Classification

Arbind Kumar Chaurasia[1], Sohit Agarwal[2],

[1]*Research Scholar(M.Tech-CSE),*[2]*Assistant Professor*

[1,2]*Department of Computer Science and Engineering*

Suresh Gyan Vihar University, Jaipur-302017

Abstract—**There is tremendous upturn in data repositories because of data generation by various organizations like government, cooperates, health caring in large amounts. Large amount of data is being produced, processed, collected, and analysed online. So there comes a requirement to transform this data into valuable information. This process of extracting the knowledge from large amount of data is referred as data mining. The proposed hybrid approach can be checked on different classifiers like Naïve Bayes, Random forest classifier etc. In proposed methodology we find that SMOTE algorithm which used K-nearest neighbour algorithm is limited to some minority class instances for creating synthetic samples, which sometimes leads to over fitting, so an effective oversampling approach can be developed.**

Keywords- Cluster, Classification, Imbalance data, Analysis, Prediction

I. Introduction

The majority of data in the original word are balanced. This happens if the distribution of the target class among different class levels is not equivalent. This classification of data is one of the toughest problems in machine learning and has become quite important recently. This has contributed to the development of most popular machine learning algorithms to maximize total accuracy, which is the percentage of precise predictions of any classifier. This results in a very low sensitivity and high accuracy to the positive class. The best approach is therefore not to concentrate on total precision but to optimize the sensitivities of the positive and negative groups separately. To overcome this problem, several methods have been developed:

Samples conform to the previous distribution of the minority and the majority the distribution of balanced classes in the training results. The techniques of sampling can be classified according to basic sampling and advanced methods. Primary sampling techniques include random minority class sampling (RSS), random minority class sampling (ROS) and the composite sampling of both.

But with random over-samples of minority data, it is possible that certain minority groups are somewhat enhanced, so that the model is trained in this case leads highly to overfitting. In contrast, random undersamples across the majority class lead to the loss of certain important information, since random data are deleted from the majority class.

SMOTE (Synthetic Minority Oversampling Technique) has been suggested by Chawla et al. (2002), as an advanced over-sampling method. It is intended by creating artificial examples within the minority class to enrich the minority class borders rather than replicating the existing examples in order to avoid overlap.

By combining the majority of subgroups with oversampling of the minor classes, multiple re-modeling for the imbalanced data set has been proposed in order to increase the categorization generalization and to prevent a combination of both sampling methods. Absolute knowledge loss or less stress can increase (Estabrooks et al., 2004). Any data set that can be configured using noise/boundary, redundant/atypical examples, leading to problems with data quality and valid descriptions, however, remove those cases that lead to poorly classified jobs.

Batista et al. (2004) proposed the use of Tomek link SMOTE to generate a series of synthetic samples for minority classes using SMOTE to over-sample minority classes, and the use of Tomek link for minority class subclasses where it would eliminate noise and border information. The motivation behind this method is similarly similar to the SMOTE To make connections. He also suggested SMOTE to his nearest neighbor. ENN deletes more cases than To make binding, so more detailed data cleaning is expected.

The Unilateral Selection (OSS) method for Kubat and Matwin (2018) is a sub-sampling method that applies CNns to a number of class noise and border instance after applying the tomake bindings tomek bindings. Confined situations can be called 'insecure' since the wrong side of their decision-making boundaries may contain little terms. The goal of CNN is to eradicate several of these far-reaching decisionsAnother method was proposed using the combination of SMOTE and OSS. The SMOTE and OSS methods provide a balancing system for data set distribution, so that the classification results are improved in terms of classification performance. The unilateral selection for subsampling follows the CNN1-NN classification rules, which do not provide sufficient subassembly and often lead to overfitting. Also, unilateral selection does not remove the externalities of the dataset by selecting (Prestanto et al., 2018). So, here we propose an approach that incorporates re-modeling techniques (oversampling and subampling) for the proper balance of the dataset and improves the nature and quality of the training data by eliminating inconsistent events such as noise /

limits / duplicates / advanced advanced data sets that lead to the workplace and all classes are correctly categorized.

II. LITERATURE REVIEW

Kubat and Matwin (1997): The goal of the proposed OSS subsampling technique is to remove some of these problematic (noise / borderline) examples from the majority class as a form of sub modelling. This suggests to them, it will reduce the examples of the majority class and therefore reduce the classification distinction.

According to Chawla et al. (2002): This paper shows that the combination of our (minor) minority class over-sampling with the minority class under-sampling can achieve higher performance classification (ROC space) than the subgroup alone.

According to Garcia et al. (2010): The problem of unbalanced learning is related to the performance of the data presented in the data presented and the algorithms present in the presence of severe classification bias.

Liu and others. (2013): Sub-Sampling is a popular method for tackling class disequilibrium issues that only uses a subset of the majority class and thus is very efficient. The big disadvantage is that many class examples are ignored.

Ganganawar (2012): In this paper we present a short overview of existing solutions to proposed problems of class imbalance at either data or algorithms. One common practice to handle imbalanced data problems is artificial recuperating it by over- and/or under-sampling, which some scientists have proved to be well-integrated with class imbalanced data set, modified support for vector machines, set-based minority-class rule methods.

According to Lopez et al. (2017): The paper offers a detailed overview of the main issues involved with the use of the internal characteristics of such classified data. It would help to develop the existing models: minor inconsistencies, a lack of focus of training results, duplication of groups, recognition of noise-related results, importance of instances of restrictions and sets adjustment, data for training and testing. We study algorithms on data like features including a view on the actions of certain experimental instances and various approaches and praise.

Agarwal etc. (2015): The paper suggested the SCUT hybrid sampling approach for balancing the amount of examples of the training in this multi-class national setting. With the production of synthetic examples, our SCUT numerical accepts minority instances and utilizes cluster analysis for the classification of the major samples.

Cao and Zhai (2015): The classification of two numbers of unbalanced data in paper was proposed with a hybrid sampling method. SMOTE is used to generate standardized points for the minority groups, and then the subsampling procedure has been used to remove much of the low-grade samples. Thus comparatively balanced data sets are created and the new data set can be addressed by using SVM.

Prestanto and others. (2018): This study will explain the imbalance class in the multiclass EDM dataset management method using a combination of SMOTE and OSS. The class SMOTE and OSS method provide a balancing system for data set distribution, so that classification results improve classification performance.

III. PROPOSED WORK

In this study, the first step is to gather and then split the unbalanced data collection that we want to define into a testing data set and a test dataset. The model is fitted first to a testing data set, an example set that matches the model parameters. The model is conditioned using a supervised learning approach in a training data set. The test data set is, finally, a data set used to evaluate neutrally how the model fits in with the training data set. The educational curriculum is then split into two subcategories, the minority and the majority.

The main goal of balancing classes is to increase the frequency of minority classes and reduce the frequency of the majority. This is done to get approximately the same number of instances for both classes. Therefore, there is a re-modeling strategy for the equilibrium class that states the data-level approach. There are two methods: over-sampling and under-sampling. When a data class is the minority class described below in the data sample, an oversampling technique will be used to extend the instances of the minority class. And for those classes that are represented as a majority, a subampling will be used to balance the minority class. Two subsets are created once the majority class is sub-sampled and the minority class is over-sampled. We combine them together to create a new balance training set.

We used the classifier to train the new balance training set, and we primarily used the test set created to evaluate the performance of the classifier.we will evaluate the proposed work of publicly available datasets from UCI repositories, KEEL repositories, NASA datasets, etc.

IV. PERFORMANCE EVALUATION:

The proposed approach will be analysed based on the performance measures. Following are the metrics that can be considered such as:

Confusion Matrix: confusion, It's considered an error unit. It is a specific table design that allows the performance of an algorithm, normally a controlled learning one, to be visualized.

TABLE 1: PERFORMANCE EVALUATION OF CONFUSION MATRIX

Actual	Predicted	
	Negative Class	Positive Class
Positive Class	False Negative (FN)	True Positive (TP)
Negative Class	True Negative (TN)	False Positive (FP)

A. True Optimistic (TP): Results are positive and are expected to be positive.

B. False Negative (FN): False negatives (FN) are good, but negative.

C. True Talks (TN): The conclusion is pessimistic and optimistic.

D. False Positive (FP): The observation is positive, but negative.

Precision: The accuracy is referred to as the significant fraction of the identified instances. The exact number of true positives in a class for a class function is the total (i.e. the number of properly identified items in a positive class) separated by an overall number of elements marked as positive elements (i.e. the sum of genuinely positive and false positive elements, which are items incorrectly marked as class objects).

Precision = TP/ (TP+FP).

Recall: Remembering the relevant instances is called the fraction. In a classification process, recall is classified by a total number of elements currently of the positive class (i.e. the aggregate of true positive and false negatives, subjects not identified as positive but predisposed to belong to the positive class). Recall is classifying as the number of true positives.

Recall = TP/ (TP + FN)

V. RESULT

Sensitivity is a significant parameter in the evaluation of the imbalanced dataset classifier efficiency. As with specificity, from the total number of cases present we will calculate the number of positive type instances correctly categorized into our data collection. For imbalances in the data set minority class instances we are most interested in the right definition of minority class instances than in the imbalanced data collection.

Via the use of data resampling, we can create the classification pattern on the training dataset, and then measure the sensitivity of how many minority class instances are properly categorized by the model.

The first column of the table 1 shows the datasets, first row of the table shows data sampling techniques and the entities under are the sensitivity evaluator of the SVM classifier given the sampling technique and the corresponding dataset.

Table 2 and Figure 1 shows the comparison of sensitivity value obtained for all the datasets for existing method (SMOTE, SMOTE-RUS, SMOTE-TOMEK Links and SMOTE-OSS) and proposed method in the form of table and bar graph. From the results, it is observed that the proposed method shown higher sensitivity value as compared to methods SMOTE (Chawla et al., 2002), SMOTE with Random Undersampling (Agrawal et al., 2015), SMOTE with Tomek Links (Batista et al., 2004) and SMOTE with One Sided Selection (Pristyanto et al., 2018). Thus, our proposed method has better performance than existing methods as after removal of all the noisy, borderline, outliers and redundant instances from majority class and oversampling of minority class will lead to the proper classification of instances and will classify minority class instances more properly.

TABLE 2 PERFORMANCE COMPARISON IN TERMS OF SENSITIVITY (%)

Dataset	**SMOTE (%)**	**SMOTE-RUS (%)**	**SMOTE-Tomek Links (%)**	**SMOTE-OSS (%)**	**Proposed Hybrid Approach (%)**
CM1	50	50	50	50	**83**
PC1	67	67	71	75	**79**
PC2	33	36	40	45	**67**
PC4	79	75	79	79	**82**
MC2	64	73	55	64	**91**
MW1	46	54	62	62	**77**
KC3	67	44	56	56	**78**
Haberman	32	32	32	32	**44**

Pima	69	69	69	66	**75**
Breast Cancer	22	25	22	38	**87**
Satimage	85	85	85	85	**88**
Optdigits	94	94	94	94	**95**

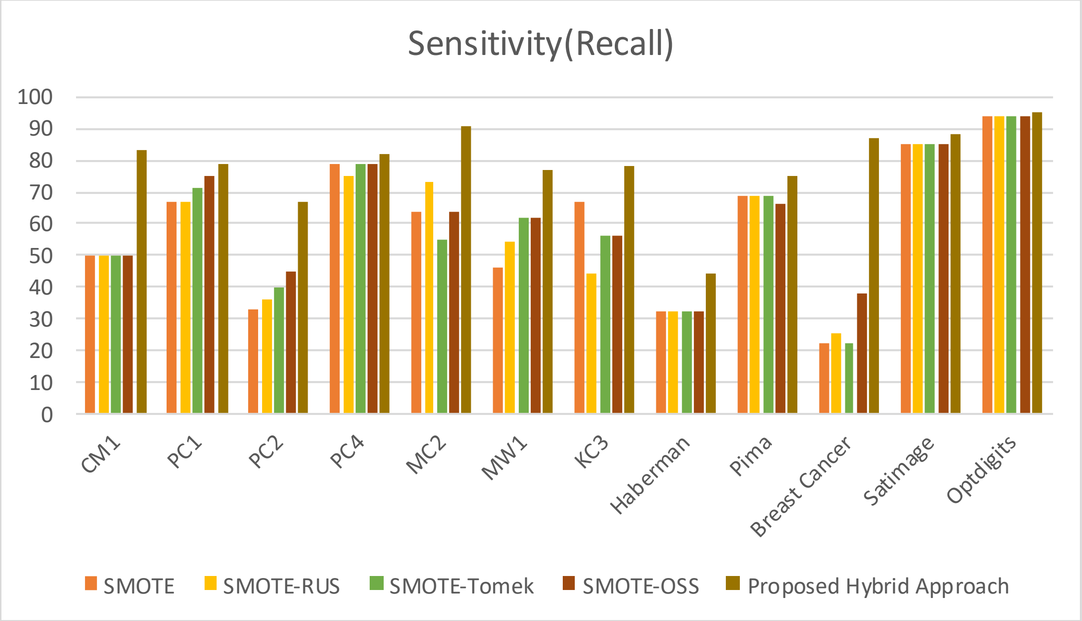

Fig 1: Bar Graph of Sensitivity

Sensitivity is an important parameter for evaluating the performance of any classifier. But with class imbalance it is not declared as good parameter because only classifying majority class instances and not correctly classifying minority class instances which are of main interest than also it gives higher accuracy.

VI. CONCLUSION

The cluster-based subsampling (CBE) strategy aims to solve class imbalance problems by leaving the majority of instances in the overlapping areas of training information. This was achieved by grouping the training dataset into some K groups and bypassing all instances of the majority class satisfying $0 < r < 1$ and r incomplete / noise instances, unnecessary examples, and the presence of outliers, such that data models have the potential to influence general estimation. For example, sampling techniques such as RSS are another data quality problem. Classification, time of sampling. These problems are inherent in most real-world data sets, as they may be incorrectly created or raised due to the nature of the application domain. It could be argued that these problematic instances can be easily deleted during the data cleansing process, but removing instances blindly from the dataset may worsen the class imbalance problem, depending on the nature of the available information.

References

[1.] Fernandez, A., Lopez, V., Galar, M., Jesus, M. J. and Herrera, F. 2013. Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-Based Systems* **42**: 97–110.

[2.] Ganganwar, V. 2012. An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering* **2**: 2250-2459.

[3.] García, V., Marqués, A. I. and Sánchez, J. S. 2012. Improving Risk Predictions by Preprocessing Imbalanced Credit Data. *In: Proceedings of 19th International Conference on Neutral Information Processing* held at Doha during November 12-15, 2012, pp. 68-75.

[4.] García, V., Sánchez, J. S. and Mollineda, R. A. 2012. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems* **25**: 13–21.

[5.] García, V., Sánchez, J.S., Mollineda, Alejo, R., R., and Sotoca, M. 2007. The class imbalance problem in pattern classification and learning. *In: Proceedings of fourth national conference on data mining and machine learning* held at Spain during September 11-14, 2007, pp. 283-291

[6.] Ghanem, A., Venkatesh, S., and West, G. 2010. Multi-Class Pattern Classification in Imbalanced Data. *In: Proceedings of 20th International Conference on Pattern Recognition* held at Istanbul during August 23-26, 2010, pp. 2881-2884.

[7.] Gray, D., Bowes, D., Davey, N., Sun, Y. and Christianson B. 2011. Further Thoughts on Precision. *In: Proceedings of 15th International Conference on Evaluation and Assessment in Software Engineering* held at Durham during April 11-12, 2011, pp. 129-133.

[8.] Gray, D., Bowes, D., Davey, N., Sun, Y. and Christianson B. 2012. Reflections on the NASA MDP data sets. *IET Software* **6**: 549-558.

[9.] Guo, X., Yin, Y., Dong, C., Yang, G. and Zhou, G. 2008. On the Class Imbalance Problem. *In: Proceedings of 2008 Fourth International Conference on Natural Computation* held at Jinan during October 18-20, 2008, pp. 192-201.

[10.] Han, H., Wang, W. Y. and Mao, B. H. 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. *In: Proceedings of International Conference on Intelligent Computing* held at Hefei during August 23-26, 2005, pp. 878-887.

[11.] Hart, P. 1968. The condensed nearest neighbour rule. *IEEE Transactions on Information Theory* **14**: 515-516.

[12.] Hawkins, D.M. 1980. Introduction. Identification of Outliers, London pp. 1-12.

[13.] He, H., Bai, Y., Garcia, E. A. and Li, S. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *In: Proceedings of IEEE International Joint Conference on Neural Networks* held at Hong Kong during June 1-8, 2008, pp. 1322-1328.

[14.] He, H. and Garcia E. A. 2009. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering* **21**: 1263-1284.

[15.] Holte R. C., Acker L. E. and Porter B. W. 1989. Concept learning and the problem of small disjuncts. *In: Proceedings of the 11th international joint conference on Artificial intelligence* held at San Francisco during August 20 - 25, 1989, pp. 813-818.

[16.] Hsu, C.W., Chang, C. C. and Lin, C. J. 2003. A Practical Guide to Support Vector Classification. Technical Report submitted to Department of Computer Science, National Taiwan University Taipei City, Taiwan

[17.] Huang, G., Song, S., Gupta, J.N.D. and Wu, C. 2014. Semi-Supervised and Unsupervised Extreme Learning Machines. IEEE Transactions on Cybernetics 44: 2405 – 2417.

[18.] Japkowicz, N. 2000. The Class Imbalance Problem: Significance and Strategies. In Proceedings of the 2000 International Conference on Artificial Intelligence (IC-AI’2000) held at Las Vegas during June 26-29, 2000, pp.111-117.

[19.] Kaur, G. and Singh, L. 2011. Data Mining: An overview. International Journal of Computer Science and Technology 2: 336-339

[20.] Khoshgoftaar, T. M., Allen, E. B., Jones, W. D. and Hudepohl, J. P. 2000. Accuracy of software quality models over multiple releases. Annals of Software Engineering 9: 103–116.

[21.] Khoshgoftaar, T.M., Gao, K. and Seliya, N. 2010. Attribute selection and imbalanced data: problems in software defect prediction. In: Proceedings of 22nd International Conference on Tools with Artificial Intelligence held at Arras during October 27-29, 2010, pp. 137-144.

[22.] Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of 14th international joint conference on Artificial intelligence held at San Francisco during August 20-25, 1995 pp. 1137–1143.

[23.] Kubat, M. and Matwin, S. 1997.Addressing the curse of imbalanced training sets: One sided selection. In: Proceedings of 14th International Conference on Machine Learning held at Nashville during July 8-12, 1997, pp. 179–186.

[24.] Laurikkala, J. 2001. Improving identification of difficult Small Classes by Balancing Class distribution. In: Proceedings of Conference on Artificial Intelligence in Medicine held at Cascais during July 1-4, 2001, pp. 63-66.