DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4783

ISSN No. 0976-5697

Volume 8, No. 8, September-October 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

COMPARATIVE ANALYSIS OF CLUSTER CONCENTRIC CIRCLE BASED UNDER SAMPLING OVER LOW VERSUS HIGH DIMENSIONAL IMBALANCED DATASETS

S.Srividhya
Research Scholar
Research and Development Centre
Bharathiar University
Coimbatore – 46,Tamilnadu

R.Mallika
Assistant Professor
Department of Computer Science
C.B.M College
Coimbatore – 42,Tamilnadu

Abstract: An imbalanced dataset influences the supervised learning model. Most of the existing real world datasets are imbalanced and often high dimensional. The existing classification methods tend to perform extremely well on the majority class and give least importance to the minority class. Most of the solutions provided for the imbalanced datasets do not fit in for the high dimensional imbalanced datasets. This paper compares the performance of an existing balancing method (cluster concentric circle based under sampling-C3BUS) over low dimensional imbalanced dataset versus high dimensional imbalanced datasets. This work shows that C3BUS works quiet well for low dimensional imbalanced dataset when compared to high dimensional imbalanced dataset and proves that class imbalance and high dimensionality are one of the two main issues in supervised learning process.

Keywords: Classification, C3BUS, Imbalanced dataset, High dimensionality, under sampling, supervised learning

I. INTRODUCTION

Data preprocessing is an important task in data mining. If the samples in each class are not equally distributed then it is said to be imbalanced class distribution. When a dataset is extremely imbalanced the classification model tend to outperform the majority class. They aim to optimize the overall accuracy without considering the minority class [1]. The existing classification methods give equal importance to both the minority class and majority class. It is difficult to build a good classifier with the existing methods when the classes are extremely imbalanced [2]. In addition to the imbalanced nature of the datasets, high dimensionality also seems to be a major issue in supervised learning. Feature selection seems to be powerful when facing high dimensional imbalanced datasets [3].

Researchers in data mining find that class imbalance is an important issue in data mining. Imbalanced natures of datasets are predominantly found in fault diagnosis applications [4], anomaly detection applications [5], medical diagnosis [6], detection of oil spills [7] and many others. Imbalances can be classified into different types like between class, multiclass, intrinsic and extrinsic imbalances. Between class imbalances are binary class imbalances in which the samples in one class out represents samples in another class. Multiple class imbalances occur between multiples classes. Imbalance occurred due to the nature of data space are Intrinsic imbalance. Imbalances due to time and storage are extrinsic imbalance [8]. A lot of research have proposed many solutions to solve the issue of imbalance including data level handling, algorithmic level handling, ensemble method [9],[10] and cost sensitive learning [11]. Data level handling methods are the sampling method which include random under sampling, random over sampling, Informed under sampling, sampling with data cleaning techniques and cluster based under sampling. This paper uses an existing cluster based under sampling technique called cluster concentric circle based under sampling [12] to compare and analyze its effect on Low dimensional imbalanced dataset and high dimensional imbalanced dataset. High dimensionality in dataset is also a major issue for a classifier to perform well. This paper is organized in the following manner. Section 2 presents the related work. Section 3 provides the description of the cluster concentric circle based under sampling (C3BUS) technique and portrays the results of the comparative analysis. Section 4 pin drops the conclusion.

II. RELATED WORK

A lot of research work has been carried out with respect to class imbalance. Japkowicz[13] introduced the issue of class imbalance and the types of imbalance with a case study comparing the existing methods to solve the issue. Chawla *et al.* [14] introduced SMOTE (an over sampling method) which creates minority samples instead of duplicating original samples. [15, 16] examines the performance of sampling techniques (SMOTE, borderline SMOTE, Wilson's editing). In [17], a distance based under sampling method was proposed to balance the class distribution and concludes that under sampling provides better recall rates than over sampling.

M.Mostafizur Rahman *et al.*, proposed a cluster based under sampling method and proved that his method is better than other existing methods[18]. Another cluster based under sampling method was proposed by Show-Jane Yen, Yue-Shi Lee that increases the prediction of minority class. This author proved his work excels by using three datasets from repository including one synthetic dataset [19]. Parinaz, Herna *et al*., explored a single classifier which used centroid based cluster under-sampling method to choose the samples. The author reports cluster centroids are not informative. In his second experiment, he explored under

sampling ensemble algorithm based on clustering called ClusFirstClass which outperforms the other state of art solutions [20]. Multi cluster based majority under sampling is proposed by Rushi Longadge *et al*., which proved cluster based random under sampling can avoid the important information loss of majority class [21].

III. COMPARITIVE ANALYSIS

One of the main issues that the researchers come across in classification is imbalanced data. Many techniques have been developed to balance the imbalanced dataset. This work considers one of the existing cluster based under sampling technique (C3BUS) [12] to compare the results over low dimensional versus high dimensional imbalanced dataset. Fig. 1 depicts the existing sampling techniques in balancing the imbalanced data.

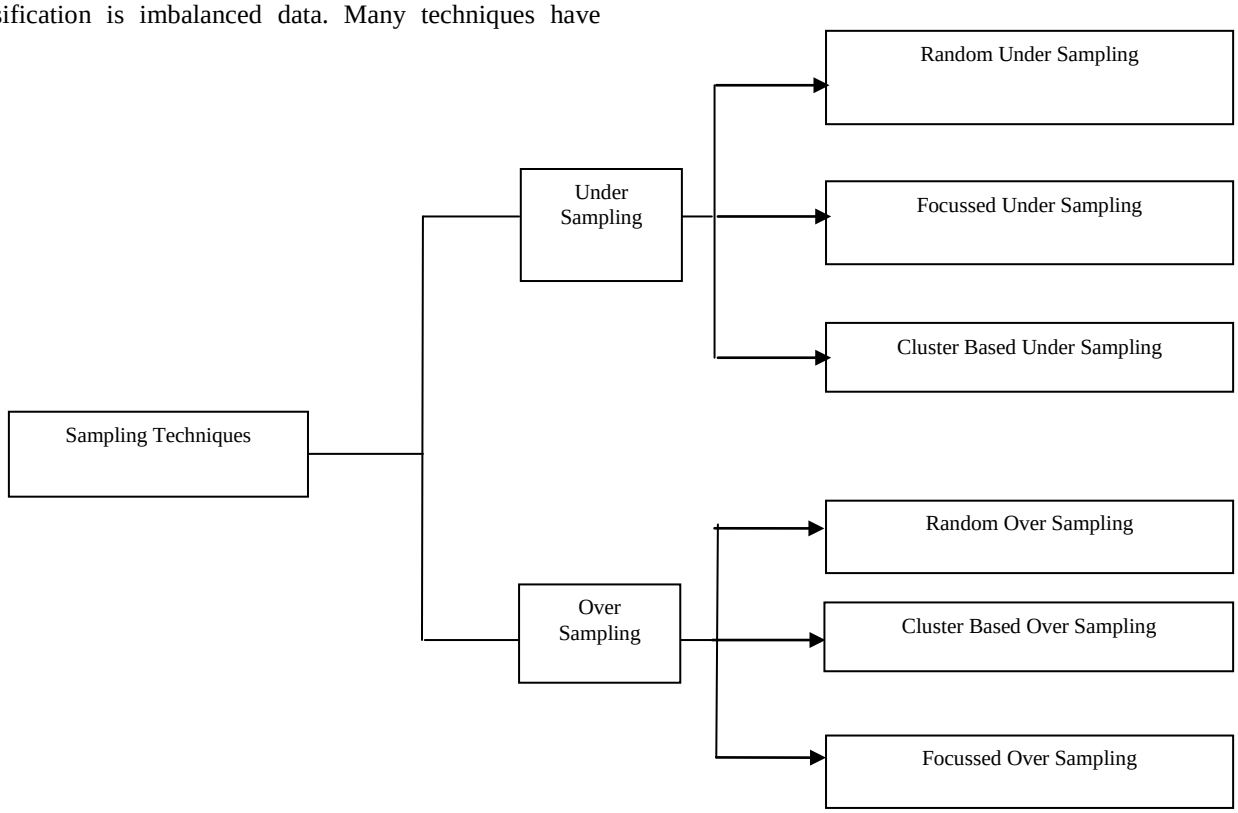

Figure 1. Sampling Techniques to balance the class distribution

A. Concentric Circle based under sampling (C3BUS)

The input dataset DS is divided into majority and minority samples. The majority samples are grouped into different k clusters using K-means algorithm. The distance between the cluster center and the samples are calculated using Euclidean distance. Each cluster is divided into concentric circles. Samples are chosen from each concentric circle. The cluster is divided into concentric circles in such a way that one sample is chosen from each circle. The chosen samples are then combined with minority class to form a balanced dataset [12].

The results in table I are obtained by applying Cluster Concentric Circle based under sampling (C3BUS) on five low dimensional imbalanced data and four high dimensional imbalanced dataset. Four evaluation metrics are used namely accuracy, precision, recall and F-measure. This analysis is carried out for binary classification. In high dimensional datasets, Brain tumor2, Lung Cancer and SRBCT are multiclass datasets. To adapt them for binary classification one class is considered against the rest. Fig.2, Fig.3 Fig.4 depicts the pictorial representation of performance measures over Low dimensional versus high dimensional datasets.

Table I. Classification results of C3BUS on Low and High dimensional datasets

NN	Low Dimensional Imbalanced Dataset					High Dimensional Imbalanced Dataset			
	Synthetic	Abalone	Bioassay	Glass	Ecoli	Brain Tumor2	Lung Cancer	Prostrate Tumor	SRBCT
Accuracy (%)	99	90	66	85	86	56	46	36	49
Precision (%)	99	86	54	85	86	61	42	40	59
Recall (%)	99	95	100	85	86	33	65	38	72
F measure (%)	99	90	70	85	86	52	59	41	58
KNN									
Accuracy (%)	99	85	83	100	90	64	15	37	66
Precision (%)	98	76	70	100	83	48	18	39	65
Recall (%)	100	100	100	100	100	32	20	28	20
F measure (%)	99	86	82	100	90	46	22	41	41
SVM									
Accuracy (%)	94	52	83	92	63	50	53	37	86
Precision (%)	62	51	70	100	59	28	53	49	73
Recall (%)	95	75	100	85	86	25	66	25	62
F measure(%)	75	61	82	91	70	26	57	29	63

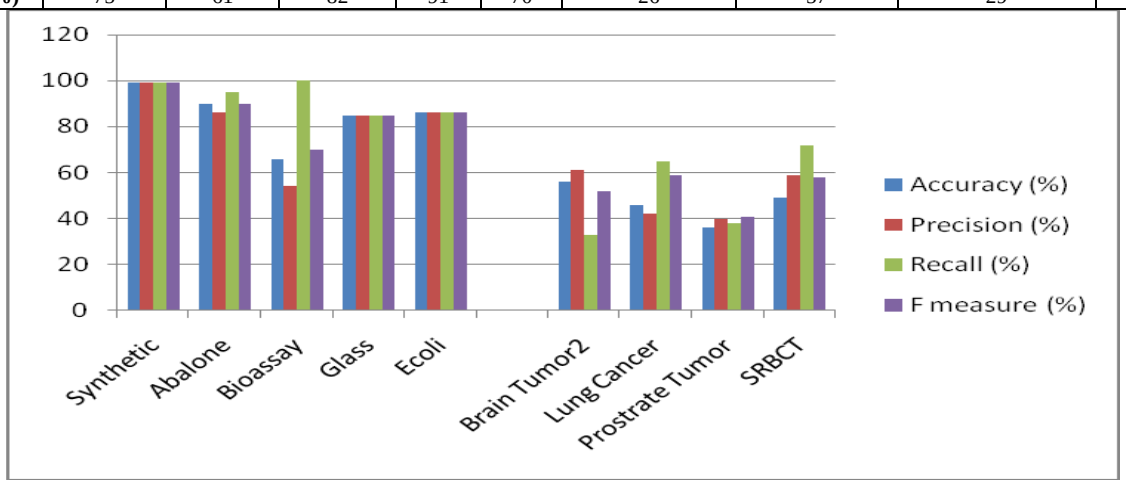

Figure 2. Performance measures for Low Vs High dimensional dataset with Neural Network Classifier

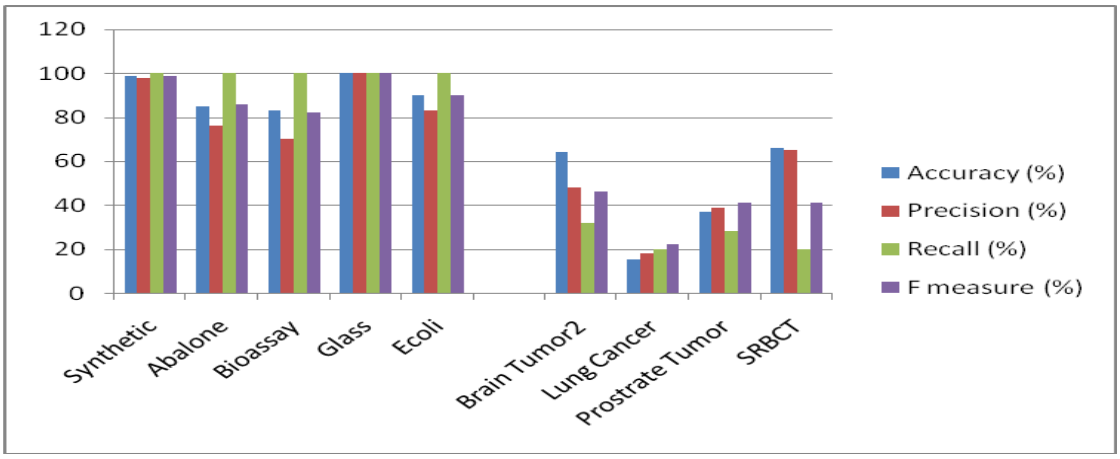

Figure 3. Performance measures for low vs. high dimensional dataset with KNN classifier

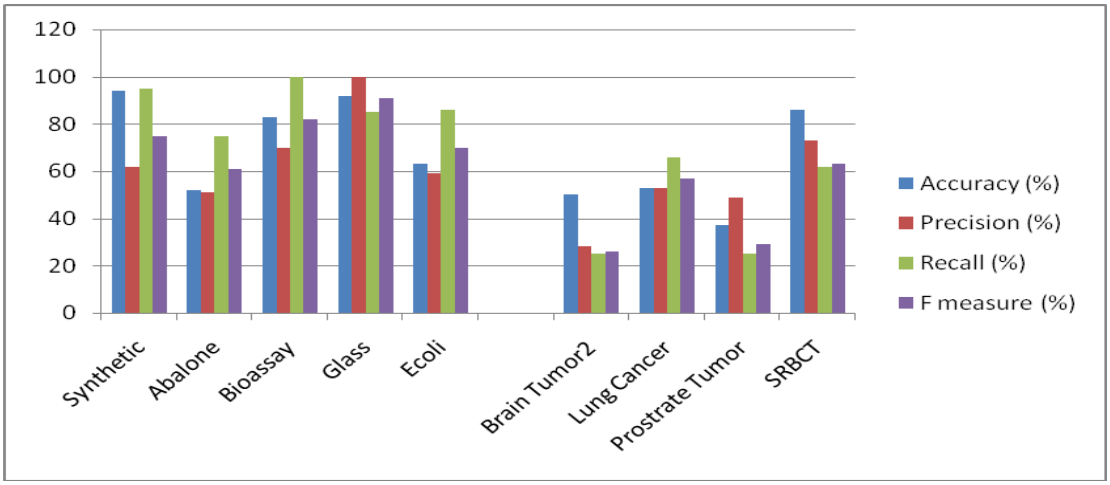

Figure 4. Performance measures for low vs. high dimensional dataset with SVM classifier

The above results show accuracy, precision, recall and F-measure are measured in terms of percentage. The analysis is carried out by applying C3BUS on low dimensional and high dimensional datasets with three classifiers namely NN, SVM, KNN. The performance of C3BUS over high dimensional datasets is very low compared to low dimensional datasets in all aspects. This brings out a clear vision that high dimensionality is one of the major issue to be handled.

IV. CONCLUSION

This work compares how the cluster based under sampling technique performs with low dimensional imbalanced dataset and high dimensional imbalanced dataset. Cluster concentric circle based under sampling (C3BUS) outperforms in terms of accuracy, precision, recall and F-measure with low dimensional imbalanced dataset than high dimensional imbalanced dataset. Results prove that high dimensionality along with imbalanced nature of dataset plays a major role in diminishing the performance of supervised learning. An attempt would be made to reduce the dimension as a future work.

V. REFERENCES

[1] Y.Liu et al., "Combining integrated sampling with SVM ensembles for learning from imbalanced datasets", Information processing & management, vol.47, no. 4, pp. 617-631 jul. 2011.

[2] Yan-Ping Zhang, Li-Na Zhang, Yong-Cheng Wang, "Cluster-based majority under-sampling approaches for class imbalance learning", 2nd IEEE International Conference on Information and Financial Engineering, pp. 400-404, September 2010.

[3] Chawla NV, Japkpwicz N, Kotcz A (2004) Editorial: Special issue on learning from imbalanced datasets. SIGKDD Explor 6(1):1-6.

[4] Z.Yang, W.tang, A.Shintemirov, and Q.wu, "Association rule mining based dissolved gas analysis for fault diagnosis of power transformers," IEEE

[5] Trans.Stst.,Man,Cybern.C,Appl.Rev.,vol.39.no.6.pp.597-610.

[6] W.Khreich, E.Granger, A.Miri, and R.Sabourin, "Iterative Boolean combination of classifiers in roc space: An application to anomaly detection with hmms," Pattern Recogn., vol.43, no.8, pp.2732-2752, 2010.

[7] M.A Mazurowski, P.A Habas, J.M Zurada, J.Y Lo, J.A. Baker, and G.D Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," Neural Netw., vol.21, no 2-3, pp.427-436, 2008.

[8] M.Kubat, R.C.Holte, and S.Matwin, "Machine Learning in detection of oil spills in satellite radar images,: Mach. Learn., vol 30, pp.295-215, 1998.

[9] Haibo He, Edwardo A.Garcia, Learning from Imbalanced data IEEE transactions on Knowledge and data engineering vol. 21 NO 9, Sep 2009.

[10] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, Handling imbalanced datasets: A review GESTS International Transactions on Computer Science and Engineering, Vol.30, 2006.

[11] Bee Wah Yap, Khatijahhusna Abd Rani, Hezlin Aryani Abd Rahman, Simon Fong, Zuraida Khairudin, Nik Nairan Abdullah, An Application of Oversampling, Undersampling, Bagging and Boosting in Handling Imbalanced Datasets in Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013), Lecture Notes in Electrical Engineering 285, DOI: 10.1007/978-981-4585-18-7.

[12] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, F Herrera, A Review on Ensembles for the class Imbalance problem: Bagging, Boosting and Hybrid based approaches, IEEE transactions on systems, Man and cybernetics- Part C:Applications and Reviews.

[13] S.Srividhya, R.Mallika, "Cluster concentric circle based under sampling to handle imbalanced data" Middle East Journal of Scientific Research, Vol. 24, pp.314-319, 2016.

[14] N. Japkowicz, "Learning from imbalanced data sets: A comparison of various strategies," in Proc. AAAI Workshop Learn. From Imbalanced Data Sets, 2000, pp. 10–15.

[15] 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

[16] R. Barandela, R. M. Valdovinos, J. S. Sanchez, and F. J. Ferri, "The imbalanced training sample problem: Under or over sampling?" in Proc. Joint IAPR Int. Workshops SSPR/SPR," vol. 3138, Lecture Notes in Computer Science, 2004, pp. 806–814.

[17] H. Han, W. Y. Wang, and B. H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in Proc. ICIC, vol. 3644, Lecture Notes in Computer Science, New York, 2005, pp. 878–887.

[18] Poolsawad, N., C. Kambhampati and J.G.F. Cleland 2014. Balancing Class for Performance of Classification with a Clinical Dataset.In the Proceedings of the World Congress on Engineering 2014 Vol I, July 2 - 4, 2014, London, U.K.

[19] Mostafizur Rahman. M. and D. N. Davis. Cluster based undersampling for unbalanced Cardivascular data. In the Proceedings of the world congress on Engineering, 2013 Vol III, WCE 2013, July 3-5, 2013.

[20] Show-Jane Yen and Yue-Shi Lee, 2009. Cluster-based under-sampling approaches for imbalanced data

[21] distributions. Expert Systems with Applications. 36(3): 5718-5727.

[22] Parinaz Sobhani, Herna Viktor and StanMatwin, 2015. Learning from Imbalanced Data Using Ensemble Methods and Cluster-based Undersampling. New Frontiers in Mining Complex Patterns Lecture Notes in Computer Science 8983: 69-83.

[23] Mr. Rushi Longadge, Ms. Snehlata S. Dongre, Dr. Latesh Malik, 2013. Multi-Cluster Based Approach for skewed Data in Data Mining. IOSR Journal of Computer Engineering (IOSR-JCE), pp: 66- 73.