Journal of Computer Sciences

Original Research Paper

Data Pre-Processing with Sampling to Reduce Uneven Data Distribution in Disease Datasets

[1]Sushruta Mishra, [2]Hrudaya Kumar Tripathy and [3]Soumya Sahoo

[1,3]*Department of Computer Science and Engineering, C.V. Raman College of Engineering, India*
[2]*School of Computer Engineering, KIIT University, Bhubaneswar, Odisha, India*

Article history
Received: 28-07-2017
Revised: 4-10-2017
Accepted: 8-11-2017

Corresponding Author:
Sushruta Mishra
Department of Computer Science and Engineering, C.V. Raman College of Engineering, India
Email: mishra.sushruta@gmail.com

Abstract: There is a constant upgradation in data aggregated from varying sources due to which proper data analysis is tough. Evaluating a classification algorithm is difficult in such circumstances. The datasets used are massive in recent days. Handling these massive datasets is a challenging task. More problem occurs when there is an uneven distribution in data samples among the classes. Classification of data becomes difficult in such scenario. Though most classifiers focus on majority class but still class with less data samples should also be taken into consideration. Thus uneven data distribution in classes leads to data skewing which needs attention of research scholars. This paper is based on the analysis of various sampling techniques on the data skewing issue using disease datasets. Three sampling methods are used in our research which include SMOTE, Spread Sub sampling and Resampling and Multi layer Perceptron is used as a classifier while Particle swarm optimization is used as the feature selection algorithm to select an optimized data from the raw disease data samples. Some critical performance metrics are used to determine the performance of classification. It is inferred that pre-processing with sampling techniques act as an optimizing agent subsequently enhancing the classification accuracy.

Keywords: Healthcare Data, Data Skewing, Particle Swarm Optimization (PSO), SMOTE, Spread Sub Sampling, Resampling, Multi Layer Perceptron

Introduction

In this modern era massive amount of data is gathered on a day to day basis. Often it is found that this huge amount of data is unevenly distributed among the different classes in the data. Especially in real time applications data is generated in a skewed manner. A skewed dataset is one in which data samples in some classes are much more than the other classes. In such a scenario a class with higher number of samples is referred to as a major class while a class with less number of samples is called as minor class. The data skewing problem is related to the intrinsic nature of the data and it may also be due to certain restrictions in retrieving data samples like cost overhead, security issues, heavy resource requirements, etc. (Phung *et al*., 2009). This data imbalance issue can result in inaccurate classification, since it gives more importance to major classes thereby neglecting the minor classes. The cost incurred during misclassification of minor classes is quite more compared to that of the major classes with maximum data samples. Examples of skewed datasets can be a cancer versus non-cancer dataset or a fraud versus unfraud dataset (Satuluri and Kuppa, 2012). But in reality there are certain applications also for which the minor class samples are more crucial than the major class samples. Attribute optimization forms an integral part in dealing with the data skewing problem. In a high dimensional dataset (e.g. many Healthcare datasets) presence of noisy attributes can degrade the efficiency of classification, thereby increasing the error rate of misclassification. This is more in case of datasets with uneven distribution of samples (Chomboon *et al*., 2013). Subsequently it is seen that data skewing is considered as an important factor in developing machine learning techniques for data classification (Chawla *et al*., 2003; Japkowicz, 2000; Weiss, 2004). Apart from this, more knowledge needs to be mined for those minor classes for which the numbers of samples are quite less (Chawla *et al*., 2004). This issue is a vital one, and is applicable in various real-time scenarios like remote-sensing (Lu and Wang, 2008), pollution detection (Huang *et al*., 2006), risk management (Cieslak *et al*., 2006), fraud detection (Mazurowski *et al*., 2008). This can be a major problem in critical applications like applications for disease diagnosis from Healthcare domain datasets where a minute negligence can even be fatal for a patient. General classifiers employed in machine learning assume substantial equality in data distribution among all the classes. This may act as a bottleneck and degrade the efficiency of the machine learning techniques for classification. Also in the case skewed data the overheads due to error rates are uneven, creating ambiguity in classification. This may also result in more overlapping of classes due to which noisy instances increase, which complicates the endeavour of data classification.

Sampling

The issue of imbalanced distribution of data samples can be addressed by using a data filtering technique called sampling. The sole purpose of sampling is to select a data sample from the original dataset that can represent the complete dataset. By employing this technique the entire dataset can be mapped to a smaller data section. Two important features that govern the sampling for the selection of samples from a given dataset are the sample size and the sample quality. In general, many factors govern the use a sample of a dataset rather than the complete dataset:

- It such a use suited for large datasets, in which case such a use may involve dealing with multiple constraints
- Is the process of data preprocessing cost-effective
- Is information loss minimum
- It is sampling strategy flexible and Adaptable to different datasets

Ideally Over-sampling and Under-sampling are two ways to deal with sampling. The later is performed on the major classes while the former is done on the minor classes.

Under-Sampling

This method is applied on the major classes. Here the samples of major classes are randomly chosen and removed, as seen in Fig. 1, so as to create a balance between major and major classes in the dataset.

Over-Sampling

This method is used to improve the sizes of minor classes of datasets so as to match the sizes with the major classes in the dataset, as seen in Fig. 2.

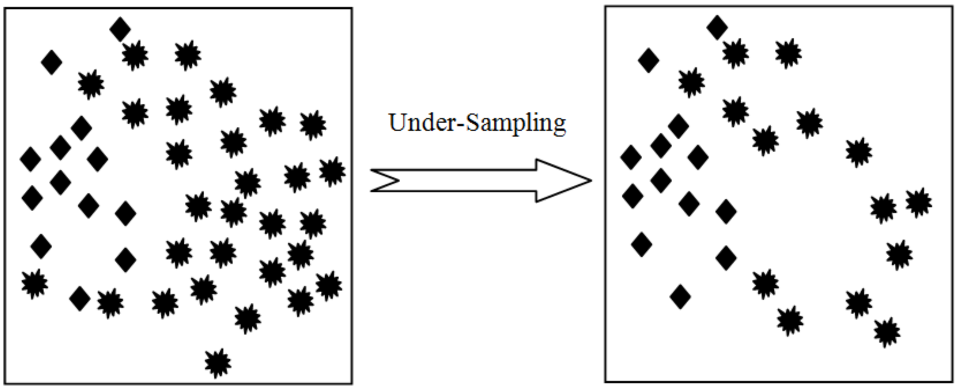

Fig. 1: Under-sampling process

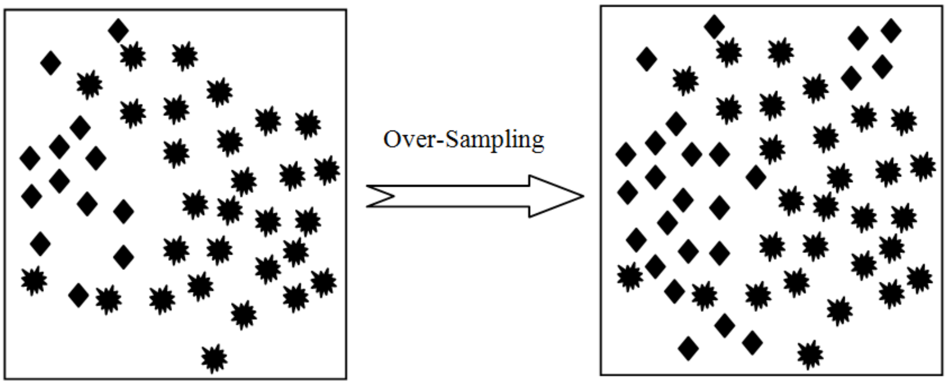

Fig. 2: Over-sampling process

Related Work

Alibeigi *et al.* (2012) the authors discussed a method to address data skewing by ranking of attributes to determine features, based on probability density. Nguwi and Cho (2010) present an attribute selection method for imbalanced datasets based on Support Vector Machines (SVMs). Chawla *et al.* (2003; Tafta *et al.*, 2009) explained the impact of SMOTE technique on sparse and unevenly distributed data where Decision Tree and Naïve Bayes were classifiers. Gaoa *et al.* (2011) a hybrid combination of Particle Swarm Optimization (PSO) and SMOTE was presented with Radial Basis Function used as a classifier. It resulted in a highly enhanced classification performance for 2-class uneven samples. Mazurowski *et al.* (2008) analyzed the effect of under sampling and oversampling methods on back propagation neural networks and PSO. The output denoted the sensitiveness of PSO algorithm towards data imbalance and small sized training data consisting of numerous attributes. Cohena *et al.* (2006) oversampling and undersampling using a resampling technique was presented which defined various metrics to tune a SVM. It resulted in the uneven data accumulation with the majority on the minor class. An effective major class balance based method was developed by authors in (Yua *et al.*, 2013) which used Ant Colony Optimization. The main drawback of this method is that its latency rate is very high. Yen and Lee (2009) the authors presented an undersampling technique based on cluster analysis. Here the entire training set is partitioned into various clusters and selected data items from major classes are chosen from every cluster representing the ratio of majority data samples to minority data samples. After the experiment was undertaken the result obtained using clustering on undersampling enhanced the classification accuracy rate. Latif and Hessampour (2014) applied genetic programming approach to select relevant attributes to recognize automatic modulation. Various experimental setups were undertaken on signals with modulation of 2PSK, 4PSK, 2FSK, 4FSK, 16QAM and 64QAM. The results depicted that features selected by genetic programming enhanced the automatic modulation recognition performance significantly. Sadegheih (2007) presents a meta-heuristic computation for formulation of a system planning network with mixed-integer programming. The results were the addition of best lines also informing the best generation at each iteration; the proposed work was illustrated to convert a 5 bus-bar system to 6 bus-bars. Myunga *et al.* (2016), a new Map-Reduce technique is presented to handle the issue of data skewing. This method is referred to as Multi-Dimensional Range Partitioning (MDRP). Vannucci and Colla (2011) (Myunga *et al.*, 2016) presents a new framework for classification which is based on binary classification technique called LASCUS which is applicable for unbalanced samples of data and used for sensitive and critical issues like malfunction identification. Wang *et al.* (2016) discusses another framework for probabilistic detection based on weighted semi-supervised k-means clustering technique and Posterior Probability SVM (PPSVM) for uneven data applicable on robot vision.

Proposed Work

The basic objective of our study is to develop well balanced Healthcare domain datasets so that there is a proper adjustment in the major and minor class data distribution. Our proposed model consists of a five-stage process of performance evaluation using data filtering techniques with sampling. The datasets considered include Breast Cancer, Diabetes and Hepatitis. These datasets were obtained from the UCI repository. In our proposed work, firstly the PSO search, which is a swarm based search technique, is applied to the raw datasets to eliminate the irrelevant and noisy attributes. The result is the optimized dataset in a reduced form. Then the reduced dataset is subjected to three different sampling techniques which include SMOTE, Resampling and Spread Subsampling. These methods are used to vary the sampling distributions in the existing dataset where the low sampled class is over-sampled while the high sampled class is under-sampled. Output of this step is a relatively balanced sample of data. After this is performed the dataset is classified. We used tMulti layer Perceptron (MLP) as classifier to perform classification on a newly arrived unknown data sample Finally after classification with MLP classifier, the efficiency of the system model is evaluated with the help of some critical performance metrics which include Positive Predictive Value (PPV), Sensitivity, Prediction Accuracy, F-Score and ROC. Our proposed work is illustrated in Fig. 3.

Different sampling techniques used in the study are:

Spread Subsampling

It is a sub sampling filter in which a maximum spread between the minor and major classes is specified so that a random subsample is filtered out which can easily be allocated in memory. For example it may be specified to denote a 2:1 ratio difference in frequencies of class. On implementation of Batch mode, there is no resampling of subsequent batches.

SMOTE

It is used for oversampling of the minority classes with random under sampling of the majority classes. In the minority classes their k-nearest minority neighbors are computed. Then a certain number of these neighbors are chosen from which synthetic data samples are generated that join the minority samples.

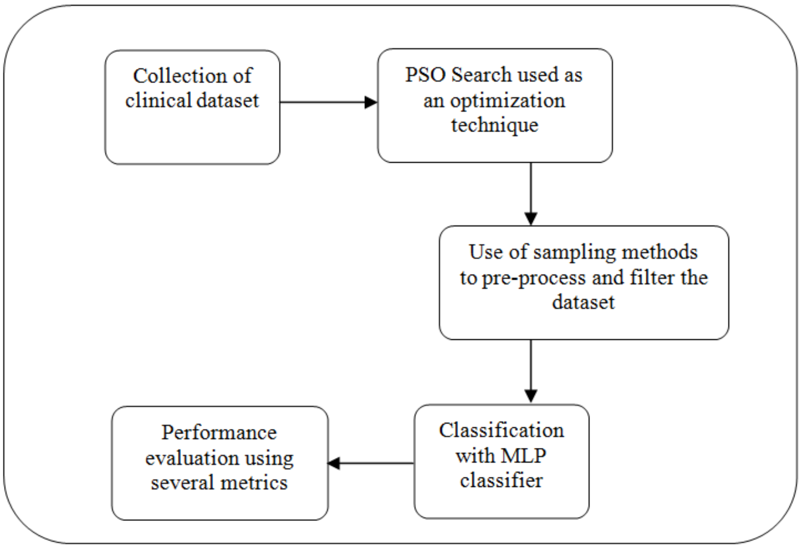
Collection of clinical dataset
PSO Search used as an optimization technique
Use of sampling methods to pre-process and filter the dataset
Classification with MLP classifier
Performance evaluation using several metrics

Fig. 3: Our proposed hybrid model based on sampling

Resampling

The objective of resampling is to minimize the uneven data distribution. It is done either by accumulating more replica of instances from the under-sampled classes or by eliminating the extra instances from the over-sampled classes. Here the bias taken is 1.0. Thus a random sub sample of a data record is generated with or without replacement.

Result and Discussion

In our study three sampling techniques for data preprocessing are used which include SMOTE, Spread Sub sampling and Re-Sampling. Breast Cancer, Diabetes and Hepatitis were the healthcare datasets under study. MLP is used for classification while Best First Search is the feature optimization method used in our research. There exist several performance evaluation parameters to evaluate the classification efficiency. But due to its uneven data distribution property all criteria are not useful in addressing the issue of data skewing in large datasets. This issue can be addressed by using some specific metrics like Sensitivity, PPV, F-Score and Prediction Accuracy rate. Our analysis has dealt with these measures to evaluate the efficiency of data sampling. To handle such problem a confusion matrix as shown in Fig. 4 is used and some basic parameters are derived from these matrixes that can represent the effectiveness of machine learning techniques are shown in Table 1.

Actual Class		Predicted Class	
		+ve	-ve
	+ve	True Positive	False Negative
	-ve	False Positive	True Negative

Fig. 4: A Confusion Matrix

Table. 1: Parameters of a Confusion matrix

True Positives (TP)	Positive Instances that were accurately identified by Classifier
True Negatives (TN)	Negative Instances that were accurately identified by Classifier
False Positives (FP)	Positive Instances that were inaccurately identified by Classifier
False Negatives (FN)	Positive Instances that were misclassified as negatives

Classification Accuracy Analysis

At first when a classification model is developed the accuracy of that model is being looked upon which the number of correct predictions from all predictions is made. This is the classification accuracy:

$$Prediction\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

There is a significant variation seen in Classification with 3-NN classifier when Sampling techniques were implemented. The accuracy of prediction sharply rises with the use of Sampling. It can be seen in Fig. 5 that Spread Sub sampling technique yield the maximum accuracy rate of 82.4% in Breast Cancer dataset and 90.24% in Diabetes dataset as seen in Fig. 6. Sampling with SMOTE method is more beneficial in Hepatitis dataset with accuracy rate of 84.65% as seen in Fig. 7.

Positive Predictive Value (PPV) Analysis

But to have a robust model this parameter is not enough to make the predictions correctly. Other than classification accuracy several other metrics are important in handling data skewing issue.

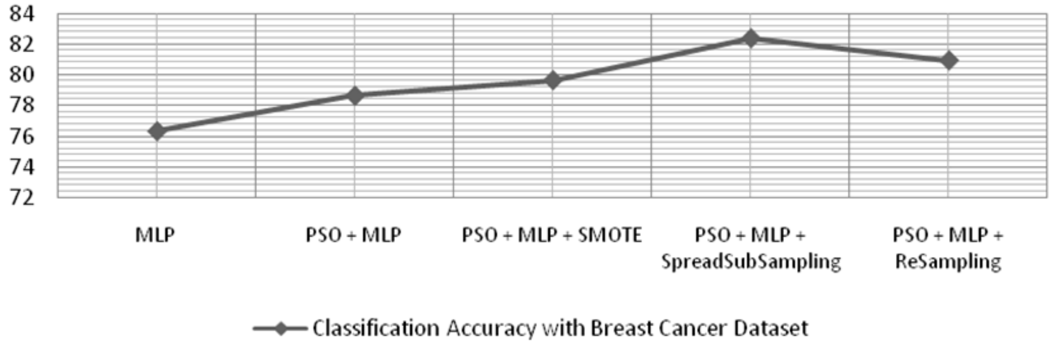

Fig. 5: Prediction Accuracy comparison in Breast cancer dataset

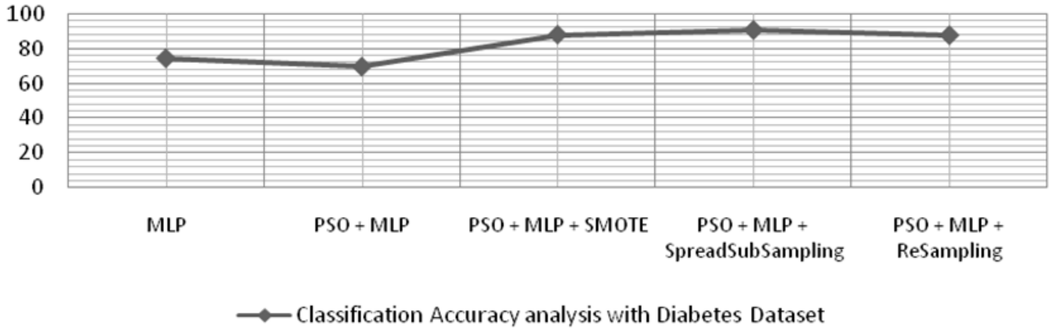

Fig. 6: Prediction Accuracy comparison in Diabetes dataset

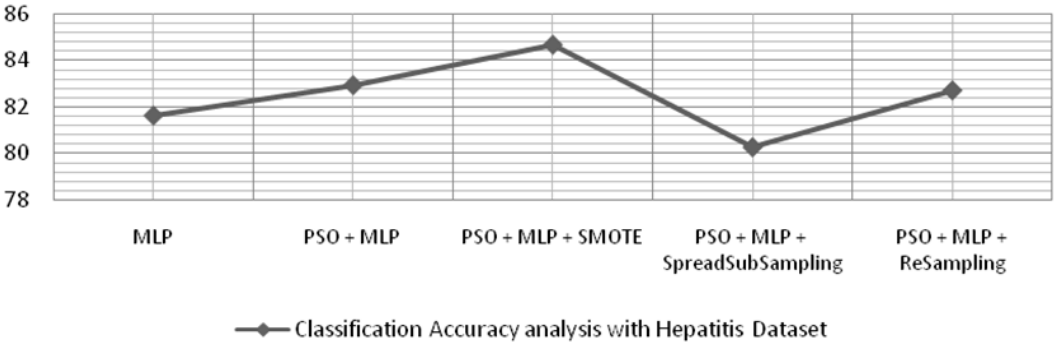

Fig. 7: Prediction Accuracy comparison in Hepatitis dataset

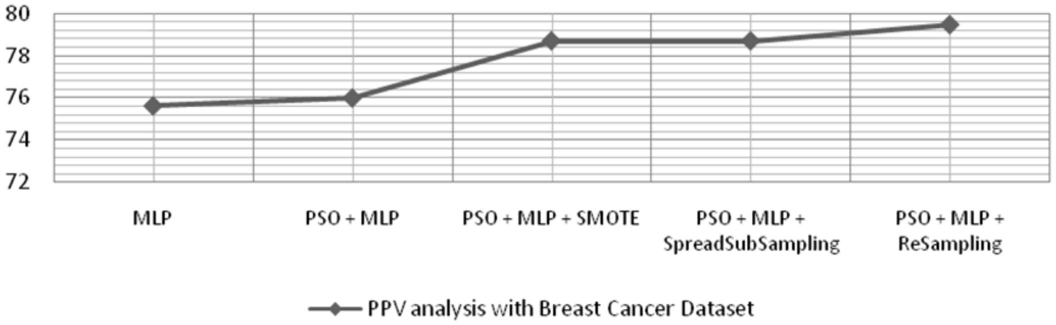

Fig. 8: PPV comparison in Breast cancer dataset

Positive Predictive value is the probability that any randomly selected retrieved document is relevant:

$$Positive\,Predictive\,Value = \frac{TP}{TP+FP} \quad (2)$$

The Positive Predictive Value differs in different datasets under consideration. The impact of Sampling on these clinical datasets is seen to be positive. As presented in Fig. 8, Re-Sampling shows the best PPV rate of 79.45% with Breast Cancer data, 92.05% is observed in Diabetes dataset using Spread sub Sampling method in Fig. 9 and 84.71% value is optimal with SMOTE method in Hepatitis data as shown in Fig. 10.

Sensitivity Analysis

Sensitivity is the probability that any randomly selected relevant document is retrieved in a search:

$$Sensitivity = \frac{TP}{TP+FN} \quad (3)$$

Sensitivity analysis is also performed on the clinical datasets. There is an uneven distribution of sensitivity graph with the use of sampling. It is observed that SMOTE method of sampling records highest of 91.24% value in Breast cancer dataset as senn in Fig. 11 while with Diabetes and Hepatitis data Re-Sampling method shows the maximum Sensitivity analysis corresponding to 91.6% and 79.18% as observed in Fig. 12 and 13 respectively.

F-Measure Analysis

But using these two values, we often cannot determine if one algorithm is superior to another. For example, if one algorithm has higher precision but lower recall than other, it is difficult to compute the superiority of an algorithm. In such case another performance metric called as F-Score which is a balanced mean between precision and recall. Higher value of F-Score is a direct measure of the effectiveness of an algorithm.

$$F-Score = \frac{2TP}{2TP+FP+FN} \quad (4)$$

It can be observed that there is a contrasting enhancement in efficiency of Classification with the application of sampling methods. As senn in Fig. 14 SMOTE records the maximum value for Breast Cancer data with 83.99% while Re-Sampling technique shows the maximum F-Measure value with Diabetes and Hepatitis datasets with 91.05% and 75.69% respectively as infered in Fig. 15 and 16 respectively.

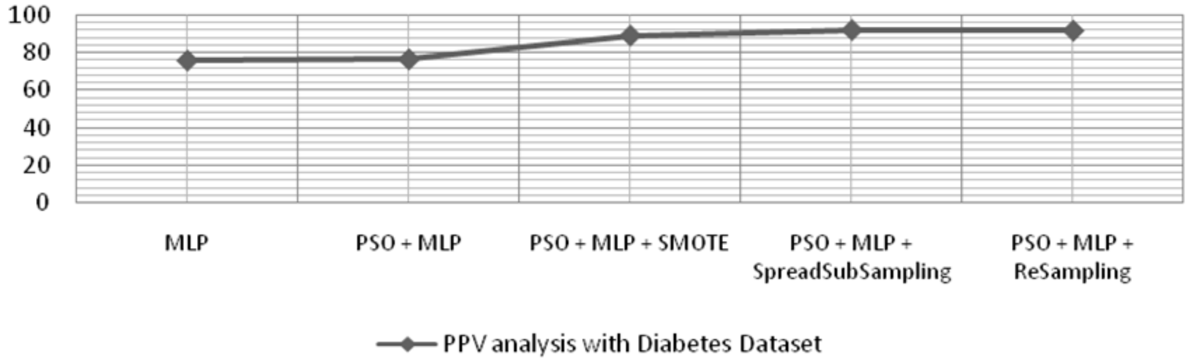

Fig. 9: PPV comparison in Diabetes dataset

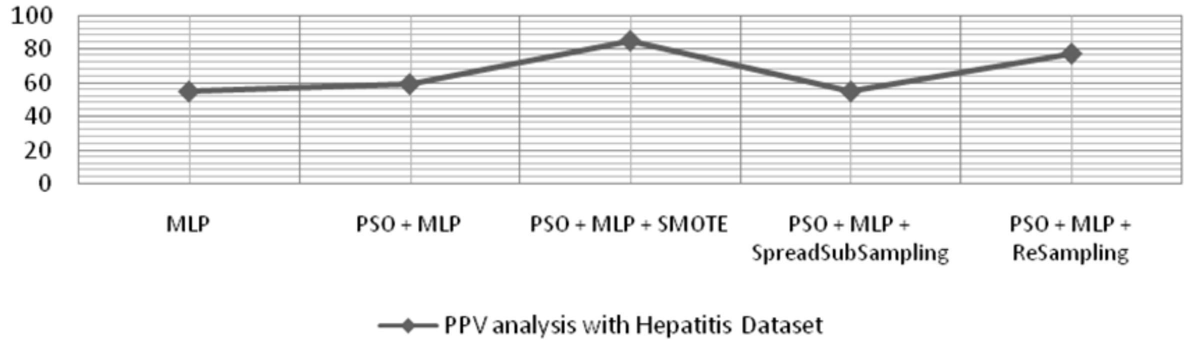

Fig. 10: PPV comparison in Hepatitis dataset

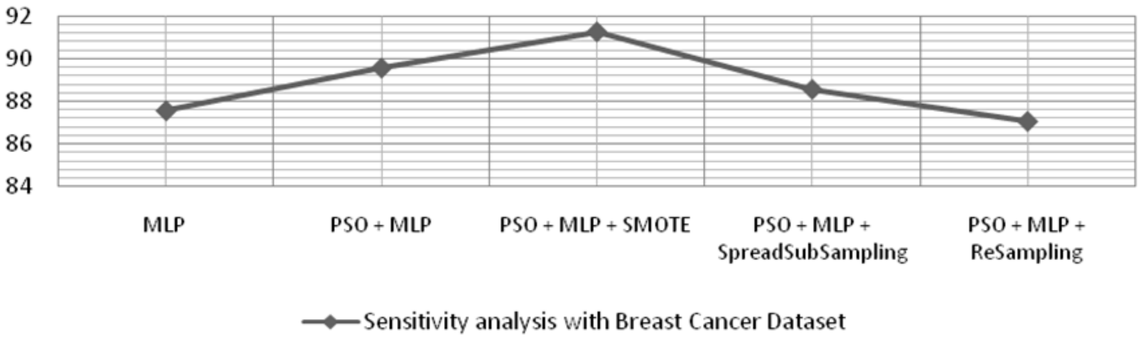

Fig. 11: Sensitivity comparison in Breast cancer dataset

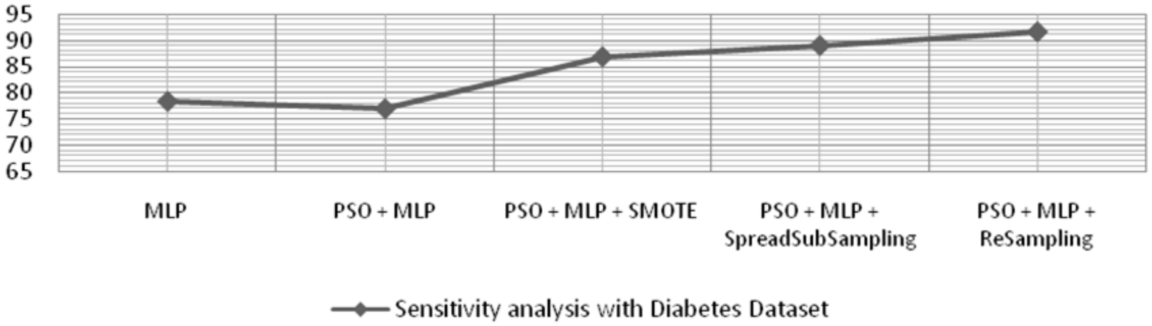

Fig. 12: Sensitivity comparison in Diabetes dataset

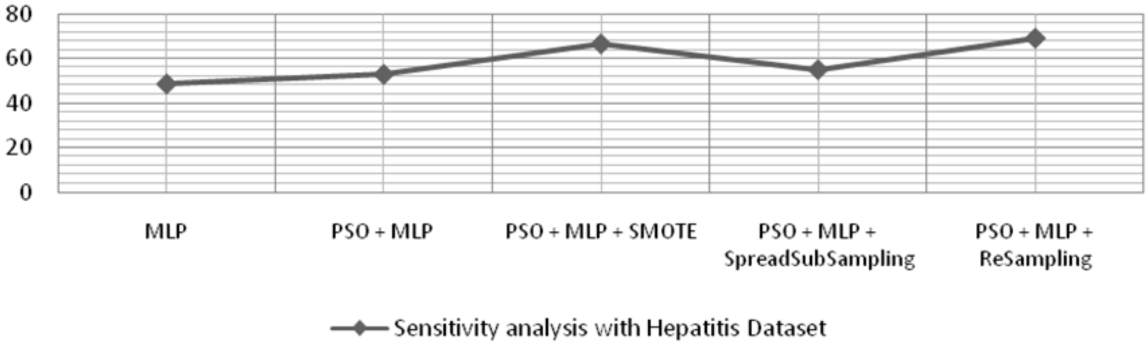

Fig. 13: Sensitivity comparison in Hepatitis dataset

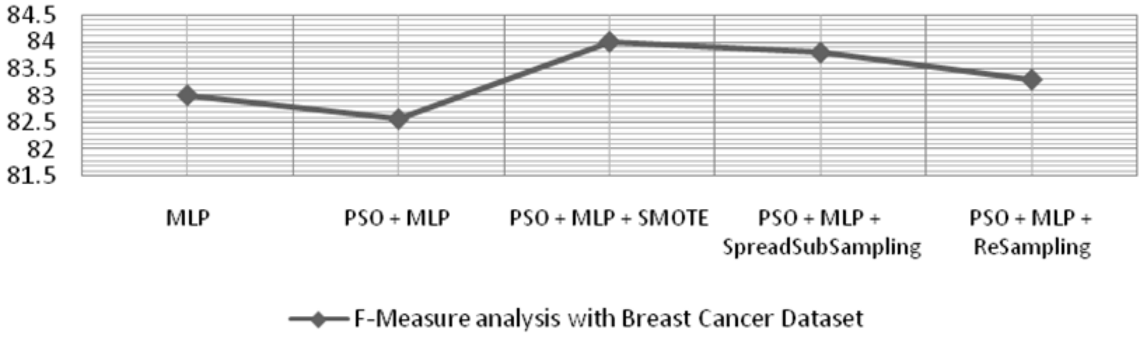

Fig. 14: F-Measure comparison in Breast cancer dataset

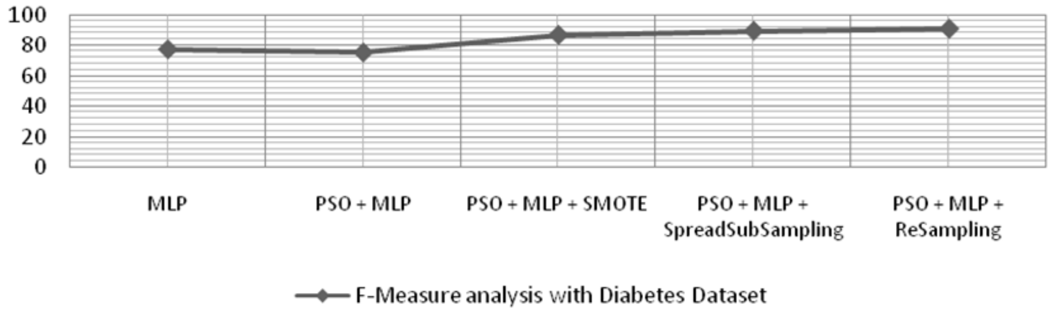

Fig. 15: F-Measure comparison in Diabetes dataset

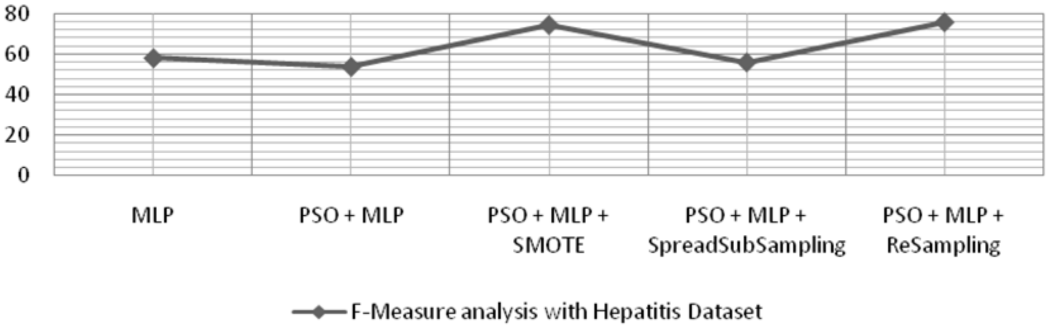

Fig. 16: F-Measure comparison in Hepatitis dataset

Conclusion

Data skewing is a hot issue that has to be addressed while handling realtime domain-specific applications using machine learning techniques. One of the critical applications is the healthcare sector. In healthcare industry efficient disease risk prediction is very important that requires a certain level of precision else the consequence may be fatal. Our study adressed the issue of data skewing applying sampling methods as data preprocessing tool on some healthcare datasets. It is observed that implementation of sampling methods is a good solution that can significantly reduce uneven data distribution among classes. By sampling new information are added or duplicate data items get eliminated which helps in balancing data. In our work we have used three sampling techniques which include SMOTE, Spread Subsampling and Resampling. While SMOTE was used as an over-sampling technique Spread Subsampling was treated as an undersampling method of data balancing. Though there is no unified rule for class balancing, still it can be inferred that classification with sampling techniques yielded optimal result than going without them for the Healthcare datasets unvestigated by us, hence for proper disease diagnosis, classification with sampling techniques is a safe option to avoid skewing of data samples, which frequenltly occur in case of Healthcare datasets.

Future Work

Our future scope include dvelopment of an ensemble classification technique that can accurately classify the data samples without suffering from uneven data distribution. Apart from this, preneting a classification framework for multi-class classification of data with different heterogeneous attributes is the objective in future research.

Acknowledgement

I would like to thank all my co-authors for their continous support and dedication that led to the completion of this research paper. We havenot taken any financial assistance from anywhere and our work has been done by us. We have followed some otehr works as mentione din the refernces and besides that we have not taken any other technical assitance from any otehr source.

Author’s Contribution

Sushruta Mishra: Data analysis, problem formulation and Implementation.

Hrudaya Kumar Tripathy: Result analysis and alignment with formatting of document.

Soumya Sahoo: Requirement analysis and literature survey works.

Ethics

The work undertaken by the authors in this paper is original and not been published elsewhere. Hence there is no violation of ethics.

References

Alibeigi, M., S. Hashemi and A. Hamzeh, 2012. DBFS: An effective density based feature selection scheme for small sample size and high dimensional imbalanced data sets. Data Knowl. Eng., 81-82: 67-103. DOI: 10.1016/j.datak.2012.08.001

Chawla, N.V., N. Japkowicz and A. Kotcz, 2003. Proceeding of ICML Workshop Learn. Imbalanced Data Sets.

Chawla, N.V., N. Japkowicz and A. Kotcz, 2004. Special issue learning imbalanced datasets. SIGKDD Explor.

Chomboon, K., K. Kerdprasop and N. Kerdprasop, 2013. Rare class discovery techniques for highly imbalance data. Proceedings of the International Multi Conference of Engineers and Computer Scientists, Mar. 13-15, IEEE Xplore Press, Hong Kong.

Cieslak, D.A., N. Chawla and A. Striegel, 2066. Combating imbalance in network intrusion datasets. IEEE Proceedings of the International Conference on Granular Computing, May 10-12, IEEE Xplore, Press, Atlanta, GA, USA, pp: 732-737. DOI: 10.1109/GRC.2006.1635905

Cohena, G., M. Hilariob, H. Saxc, S. Hugonnetc and A. Geissbuhlera, 2006. Learning from imbalanced data in surveillance of nosocomial infection. Art. Int. Med., 37: 7-18. DOI: 10.1016/j.artmed.2005.03.002

Gaoa, M., X. Honga, S. Chenb and C.J. Harrisb, 2011. A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems. Neurocomput., 74: 3456-3466. DOI: 10.1016/j.neucom.2011.06.010

Huang, Y.M., C.M. Hung and H.C. Jiau, 2006. Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem. Nonlinear Anal. R. World Applied, 7: 720-747. DOI: 10.1016/j.nonrwa.2005.04.006

Japkowicz, N., 2000. Proceeding of AAAI Workshop Learn. Imbalanced Data Set.

Latif, A. and K. Hessampour, 2014. Dimensionality reduction and improving the performance of automatic modulation classification using genetic programming. Int. J. Eng., 27: 709-714.

Lu, W.Z. and D. Wang, 2008. Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme. Sci. Total Environ., 395: 109-116. DOI: 10.1016/j.scitotenv.2008.01.035

Mazurowski, M.A., P.A. Habas, J.M. Zurada, J.Y. Lo and J.A. Baker *et al*., 2008. Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. Neural Netw., 21: 427-436. DOI: 10.1016/j.neunet.2007.12.031

Myunga, J., J. Shimb, J. Yeonc and S.G. Lee, 2016. Handling data skew in join algorithms using Map Reduce. Expert Syst. Applic., 51: 286-299. DOI: 10.1016/j.eswa.2015.12.024

Nguwi, Y. and S. Cho, 2010. An unsupervised self-organizing learning with support vector ranking for imbalanced datasets. Expert Syst. Applic., 37: 8303-8312. DOI: 10.1016/j.eswa.2010.05.054

Phung, S.L., A. Bouzerdoum and G.H. Nguyen, 2009. Learning pattern classification tasks with imbalanced data sets.

Sadegheih, A., 2007. A novel method for designing and optimization of networks. IJE Tran. A: Basics, 20: 17-26.

Satuluri, N. and M.R. Kuppa, 2012. A novel class imbalance learning using intelligent Under-sampling. Int. J. Database Theory Applic., 5: 25-35.

Tafta, L.M., R.S. Evansae, C.R. Shyuac, M.J. Eggera and N. Chawlad *et al*., 2009. Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery. J. Biomed. Informatics, 42: 356-364. DOI: 10.1016/j.jbi.2008.09.001

Vannucci, M. and V. Colla, 2011. Novel classification method for sensitive problems and uneven datasets based on neural networks and fuzzy logic. Applied Soft Comput., 11: 2383-2390. DOI: 10.1016/j.asoc.2010.09.001

Wang, Y., X. Li and X. Ding, 2016. Probabilistic framework of visual anomaly detection for unbalanced data. Neurocomputing, 201: 12-18. DOI: 10.1016/j.neucom.2016.03.038

Weiss, G.M., 2004. Mining with rarity: A unifying framework. ACM SIGKDD Explor. Newslett., 6: 7-19. DOI: 10.1145/1007730.1007734

Yen, S. and Y. Lee, 2009. Cluster-based under-sampling approaches for imbalanced data distributions. Expert Sys. Applic., 36: 5718-5727. DOI: 10.1016/j.eswa.2008.06.108

Yua, H., J. Nib and J. Zhaoc, 2013. ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data. Neurocomputing, 101: 309-318. DOI: 10.1016/j.neucom.2012.08.018