THINK INDIA JOURNAL
ISSN: 0971-1260
Vol-22-NO-30 Special Issue-December-2019

An Empirical Study of Decision Tree on Multi-Class Imbalanced Data

Dr. S Jahangeer Sidiq
School of Computer Science and Engineering
Lovely Professional University. Phagwara, Punjab, India
Email: jahangir.25453@lpu.co.in

Dr. Majid Zaman
Directorate of IT&SS
University of Kashmir. Srinagar, J&K, India
Email: zamanmajid@gmail.com

Abstract.
Imbalanced datasets are generated from numerous real-world applications. In datasets with imbalance the minority (positive) class is usually a small fraction of total examples, but misclassification of these examples is quite expensive. A lot of work has been done on class imbalance problem with two classes but datasets possessing multiple classes have considerably received less focus. This is because of the fact that multi-class imbalance problem is hard problem than binary class problem. Here in this work we study multi-class imbalance problem using unpruned decision trees and study the impact of resampling techniques such as random oversampling and random under sampling on the benchmark datasets. The datasets used here were imported from UCI repository and Sci-Kit, Imb-Learn machine learning tools were used for performing experiments.

Keywords: learning, over-sampling, under-sampling, decision tree.

1. Introduction

An imbalance in a dataset exists if classes are not represented equally.Several attempts were made to deal with imbalance problem in the domains of telecommunications [1],fraudulent telephone calling [2],detection of oil spills from satellite imagery[3] and text classification [4],[5],[6],[7].The most compelling query is, given the class distribution:What class distribution is right for a machine learning technique? A detailed analysis was presented by Wiess and Provost on effect of class distribution on learning of classifiers. For decision tree the structure is one of the most important factors i.e. Pruned and unpruned decision trees have different effects on learning process from data sets that are imbalanced.Pruning process can be unfavorable for learning from data sets that are imbalanced as it removes the leaves that belong to the minority class thereby decreasing the overall coverage. So,this gets us to another query:What should be the exact structure of a tree?Is tree pruning necessary process or do we have to make use of completely grown trees?An accuracy measure is typically used for evaluation of decision tree that considers every error equally.However when imbalance exists in data accuracy it is not an appropriate measure. e.g. considering an example of classification of mammogram images as possibly cancerous [8],[9],[11].This dataset may contain 2% abnormal pixels and 98% normal pixels. If the 2% abnormal pixels are classified as normal the predictive accuracy would be

ISSN: 0971-1260
Vol-22-NO-30 Special Issue-December-2019

98%.The high rate of correct predictions for cancerous patients is required and a small to reasonable error rate for the non-cancerous patients is required. Here predicting the cancerous patient as non-cancerous is more costly than otherwise.

We used unpruned decision tree on 7 multi-class imbalanced datasets imported from UCI repository using default setting in Sci-Kit for decision tree.We used G-mean,F-Score,Precision and Recall as performance metrics.Also we analyzed the impact of over and under sampling on unpruned trees and the resampling was done using Imb-learn package.

The structure of the paper is as following.Section 2 describes the sampling techniques used.Section 3 gives the data set description used for experimentation.In section 4 we discuss the results of the experiments.Finally section 5 concludes our paper.

2. Sampling strategies

A most common technique used for dealing with data which is imbalancedis either to over-sample thepositive(minority) class or down-sample(under-sample) the negative (majority) class. This is the most common strategy used because of the fact that we can use any classifiers at later stage for model creation.

2.1. Over-sampling

Using over-sampling does not increase the classification accuracy of the minority class in every case.The basic effect is interpreted in terms of decision boundaries in attribute space.When the over-sampling of minority class is performed in tremendous volumes, the effect is to classify analogous and more precise regions in attribute space as decision regions for positive class. If positive (minority) class is replicated, the positive class decision area becomes much specific and leads to splits in a tree.This ultimately leads to overfitting.The duplication of positive class doesn't lead to spread of decision boundary in to the region of majority class.

2.2. Under-sampling

Under-sampling randomly deletes the examples from the negative(majority)class untilpositive(minority) class size becomes somewhat similar to the majority class [9].The varying degree of under-sampling is experienced by the classifiers and when the under-sampling takes place at a higher rate automatically the positive class has a higher occurrence in the training dataset.

3. Data sets

In this paper we have used 7 different datasets with multiple classes and each of the dataset has a different class distribution. These datasets were imported from UCI repository [10].Table 1 shown below gives the complete description of the datasets used in our experiments.

ISSN: 0971-1260
Vol-22-NO-30 Special Issue-December-2019

Dataset	Total examples	Total classes	Class distribution	Total attributes
Wine	178	3	59/71/48	13
Shuttle	58000	7	45586/49/171/8903/3267/10/13	9
Penbased	10992	10	115/114/114/…/106	16
Pageblock	5473	5	4913/329/28/88/115	10
Leaf	340	40	12/10/10/8/12/8/10…/11	16
Glass	214	6	70/76/17/13/9/29	10
Dermatology	366	6	112/61/72/49/52/20	34

Table 1. Dataset Description

4. Results

We have used the unpruned decision tree without sampling and with sampling techniques like random over-sampling, random under-sampling. The ultimate aim was to empirically investigate the effect of over and under sampling methods on unpruned decision tree.The over-sampling and under-sampling was carried out using Imb-learn package with default settings for both the sampling techniques. There may be asuitable class distribution that would give us far better values in terms of performance measures like Precision,G-mean,Recall and F-measure and this would be the part of our future work. In our experiments we used Stratified Shuffle Split with 10-folds so the distribution of samples in the training and testing data set are in the same ratio.We have compared the performance of no sampling, under-samplingand over-sampling using decision tree as the classifiers for model creation on different datasets using metrics such as Precision, G-mean, Recall and F-measure.The output of performances in shown in Table 1, Table 2 and Table 3.The mean of different metrics over different datasets in shown in the last row of each table. And graphically the performance is shown in Figure 1, Figure 2 and Figure 3.Figure 4 summarizes the performance comparison of decision tree using over-sampling,under-sampling and no-sampling as pre-processing/data balancing strategy.

	F1_Score	Precision_Score	Recall_Score	Gmean_Score
wine	0.887115	0.89818	0.88418	0.912668
shuttle	0.905663	0.938746	0.88925	0.941675
penbased	0.865579	0.870084	0.865782	0.923484
pageblock	0.803133	0.81646	0.797172	0.874332
leaf	0.615053	0.665395	0.620556	0.782138
glass	0.645302	0.65182	0.672488	0.790065
dermatology	0.926768	0.933117	0.927008	0.957365
***Mean**	**0.806945**	**0.824829**	**0.808062**	**0.883104**

Table 1: Decision tree with no sampling

	F1_Score	Precision_Score	Recall_Score	Gmean_Score
wine	0.926358	0.93218	0.92746	0.945317
shuttle	0.356	0.323333	0.44	0.602561
penbased	0.918745	0.922284	0.918851	0.954194
pageblock	0.92167	0.926951	0.9225	0.950932

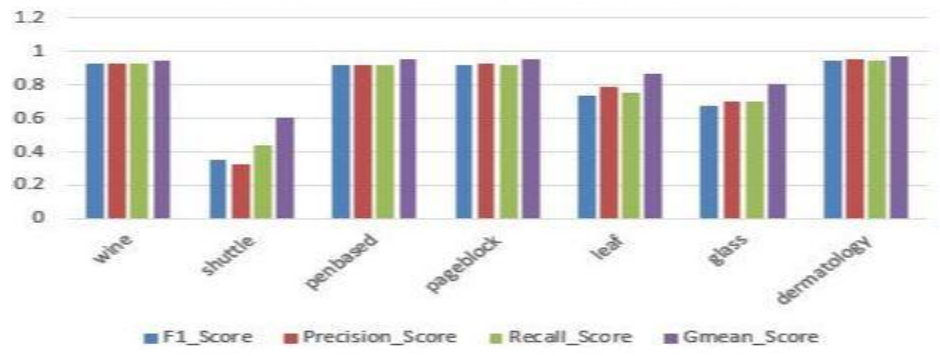
ISSN: 0971-1260
Vol-22-NO-30 Special Issue-December-2019

leaf	0.738664	0.786111	0.758333	0.866687
glass	0.675714	0.700278	0.697222	0.808135
dermatology	0.949577	0.957738	0.95	0.969701
***Mean**	**0.783818**	**0.792697**	**0.802052**	**0.871075**

Table 2: Decision tree with under sampling

	F1_Score	Precision_Score	Recall_Score	Gmean_Score
wine	0.948209	0.950238	0.948485	0.961222
shuttle	0.999101	0.999106	0.999101	0.999438
penbased	0.868801	0.873551	0.869042	0.925406
pageblock	0.993411	0.993471	0.993419	0.995884
leaf	0.753931	0.790851	0.763833	0.870144
glass	0.867433	0.869726	0.873715	0.922814
dermatology	0.983656	0.984419	0.983601	0.990137
***Mean**	**0.916363**	**0.923052**	**0.918742**	**0.952149**

Table 3: Decision tree with oversampling



ISSN: 0971-1260
Vol-22-NO-30 Special Issue-December-2019

Figure 1: Decision tree with no sampling

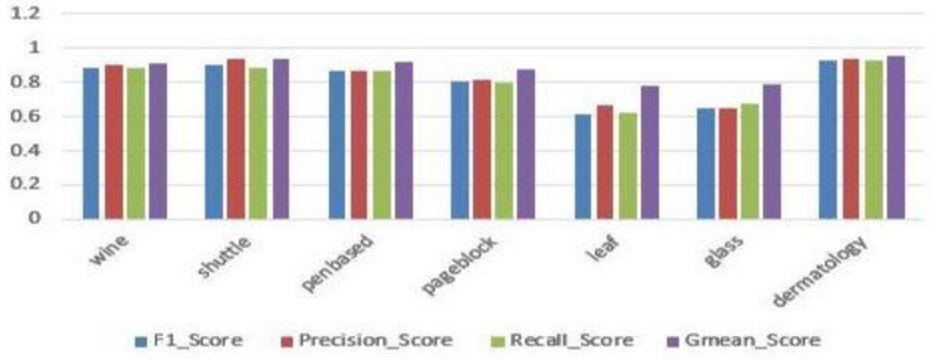


Figure 2: Decision tree with undersampling

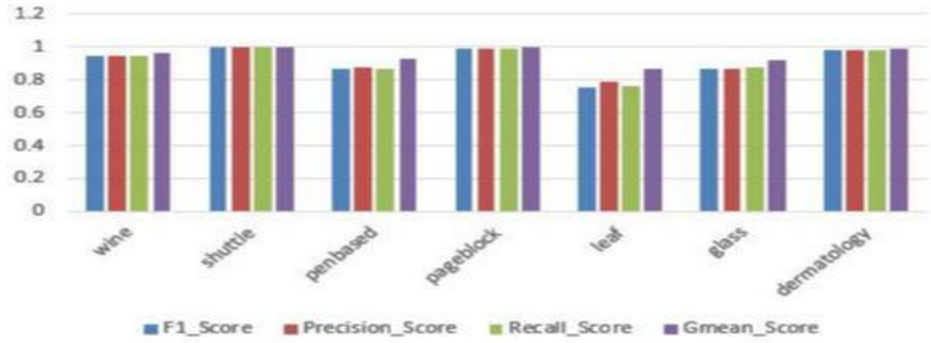


Figure 3: Decision tree with over sampling

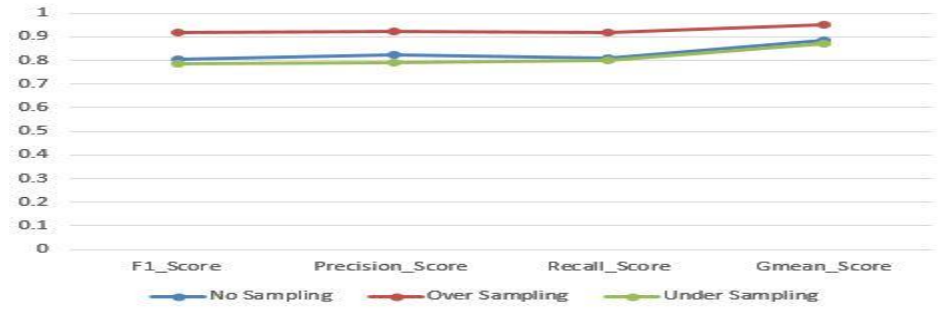


Figure 4: Performance comparison

5. Conclusion

In this paper we have presented an experimental analyses of unpruned decision tree using no preprocessing/sampling,randomunder-sampling and random over-sampling as data balancing strategy for multi-class imbalance datasets. Random over-sampling is the best data balancing strategy than under-sampling or no-sampling for decision trees on multi-class imbalanced data. While as random under-sampling degrades the classification performance of decision tree in terms of all the metrics mentioned in this paper.We also observed that for decision tree no preprocessing/data balancing is better than under-sampling as it leads to the degradation of performance in terms of all metrics.

ISSN: 0971-1260
Vol-22-NO-30 Special Issue-December-2019

6. References

[1] Ezawa, K. J., Singh, M., & Norton, S. W. (1996, July). Learning goal oriented Bayesian networks for telecommunications risk management. In *ICML* (pp. 139-147).

[2] Fawcett, T., & Provost, F. Combining Data Mining and Machine Learning for E ective User Profiling. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining* (pp. 8-13).

[3] Kubat, M., Holte, R. C., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine learning*, *30*(2-3), 195-215.

[4] Lewis, D. D., & Ringuette, M. (1994, April). A comparison of two learning algorithms for text categorization. In *Third annual symposium on document analysis and information retrieval* (Vol. 33, pp. 81-93).

[5] Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithms and representations for text categorization.

[6] Mladenic, D., & Grobelnik, M. (1999, June). Feature selection for unbalanced class distribution and naive bayes. In *ICML* (Vol. 99, pp. 258-267).

[7] Cohen, W. W. (1995). Learning to classify English text with ILP methods. *Advances in inductive logic programming*, *32*, 124-143.

[8] Woods, K. S., Doss, C. C., Bowyer, K. W., Solka, J. L., Priebe, C. E., & Kegelmeyer Jr, W. P. (1993). Comparative evaluation of pattern recognition techniques for detection of microcalcifications in mammography. *International Journal of Pattern Recognition and Artificial Intelligence*, *7*(06), 1417-1436.

[9] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

[10] Blake, C., & Merz, C. (1998). UCI Repository of Machine Learning Databaseshttp://www.ics.uci.edu/_mlearn/_MLRepository.html.Department of Information and Computer Sciences,
University of California, Irvine.

[11] Bala, Jeevan & Salaria, Avinash & Singh, Parminder. (2016). Protein-protein Interaction Prediction using Variable Length Patterns with GPU. Indian Journal of Science and Technology. 9. 10.17485/ijst/2016/v9i45/106351.