

# MSMOTE: Improving Classification Performance when Training Data is imbalanced

Dr S R C Murthy P, Md Ayub Khan, P. Sandeep Reddy

Professor, Assistant Professor, Associate Professor

Dept. of CSE,

mail-id: ayubkhanmd32@gmail.com, sandeepreddycse@anurag.ac.in

Anurag Engineering College, Anatagiri (V&M), Suryapet (Dt), Telangana-508206

**Abstract**—Learning from data sets that contain very few instances of the minority class usually produces biased classifiers that have a higher predictive accuracy over the majority class, but poorer predictive accuracy over the minority class. SMOTE (Synthetic Minority Over-sampling Technique) is specifically designed for learning from imbalanced data sets. This paper presents a modified approach (MSMOTE) for learning from imbalanced data sets, based on the SMOTE algorithm. MSMOTE not only considers the distribution of minority class samples, but also eliminates noise samples by adaptive mediation. The combination of MSMOTE and AdaBoost are applied to several highly and moderately imbalanced data sets. The experimental results show that the prediction performance of MSMOTE is better than SMOTEBoost in the minority class and F-values are also improved.

**Keywords**—*imbalanced data, over-sampling; SMOTE, AdaBoost, samples groups, SMOTEBoost*

## INTRODUCTION

The classification question is one of the important research contents in the data mining, the machine learning and pattern recognition. The many classification algorithms have been researched extensively and achieved success in reality applications. Then those methods come to imbalance data sets, the performance for minority class may not be so good. Unfortunately, imbalance data set is common in many practical applications such as detecting fraudulent transactions, network intrusion detection, Web mining, direct marketing, and medical diagnostics

In these applications, the correct classification for the minority class samples is more valuable than that for the majority class samples. However, because the data distribution is not balanced, the existing classification algorithms have many difficulties for correctly classifying the minority class samples. For example, the performance of the model is good for majority class samples, but the performance for minority class samples is very bad. The problem of class imbalance is a main reason.

Many techniques have been proposed to alleviate the problem of class imbalance. Sampling is a most common method to process the imbalance data sets. Eliminating or reducing the imbalance of the data through the changes of training data distribution is the main idea of sampling. Under-

sampling and over-sampling are two basic modes of sampling. Under-sampling balances two kinds of samples through reduced majority class samples' quantities, and then over-sampling achieves the balance through the duplication of minority class samples. The two methods have some drawbacks. Under-sampling neglects some useful samples, so it can cause to reduce the performance of classifier. But over-sampling introduces the extra training samples. This will lengthen the time of training model. Duplicating the samples will also cause to over-fitting. To overcome the over-fitting, Chawla et al. (2002) proposes a SMOTE [1] algorithm. SMOTE creates synthetic instances of the minority class by operating in the "feature space" rather than the "data space". By synthetically generating more instances of the minority class, the learners are able to broaden their decision regions for the minority class. Based on under-sampling, some modified methods were proposed such as [6].

Boosting [2] is another way to process imbalance data sets. Because boosting gives the misclassified training samples a high weighted value in each iteration, it changes the distribution of training data effectively. Due to the distribution changes of imbalance data sets, the boosting algorithm is effective for the minority class classification. The most common boosting algorithm is AdaBoost [3]. For improvement of the performance of classifiers, Nitesh V. Chawla et al. (2003) proposed SMOTEBoost [4]. SMOTEBoost algorithm combines the Synthetic Minority Oversampling Technique (SMOTE) and the standard boosting procedure. It utilizes SMOTE for improving the prediction of the minority class and utilizes modified boosting to not sacrifice accuracy over the entire data set. In each round of boosting, SMOTE is introduced to generate data and increase the sampling weights for the minority class.

Zhou et al (2006) proposed cost-sensitive neural network [5]. They address the class imbalance problem with this method. The experimental results show the method is efficient, and then it is difficult to obtain the cost matrix.

SMOTE doesn't consider the distribution of minority classes and latent noises in data set when it generates synthetic examples by taking each minority class sample and introducing synthetic examples. To improve the performance of SMOTE, a modified method-MSMOTE is proposed in this paper. The modified algorithm classifies the samples of

minority class into three disjoint groups. Security samples, border samples and latent noise samples by calculating the distance of all the samples. Security samples are those data points that can enhance the performance of classifier. On the contrary, the noises can reduce the performance of classifier. Those samples that classifier is hard to classify is labeled as border samples. When MSMOTE generate synthetic examples, different strategy is used for selecting its near neighbors according to the samples' type.

The combination of MSMOTE and AdaBoost [3] are applied to several highly and moderately imbalanced data sets, the experimental results show that the prediction performance of MSMOTEBoost is better than SMOTEBoost in on the minority class and F-values are also improved.

The rest of this paper is organized as follows. Section 2 briefly reviews SMOTE and SMOTEBoost. Section 3 presents the algorithm MSMOTE. Section 4 reports on the experiments. Finally, Section 5 concludes.

**SMOTE AND SMOTEBOOST**

SMOTE (Synthetic Minority Oversampling Technique) was proposed to counter the effect of having few instances of the minority class in a data set. SMOTE creates synthetic instances of the minority class by operating in the “feature space” rather than the “data space” [4]. By synthetically generating more instances of the minority class, the learners are able to broaden their decision regions for the minority class. The new synthetic minority samples are created as follow steps [1, 4]. Firstly, take the difference between a feature vector (minority class sample) and one of its k nearest neighbors (minority class samples).Then, multiply this difference by a random number between 0 and 1. Finally, add this difference to the feature value of the original feature vector, thus a new feature vector is created.

SMOTEBoost algorithm combines the Synthetic Minority Oversampling Technique (SMOTE) and the standard boosting procedure. By introducing SMOTE in each round of boosting, SMOTEBoost enable each learner to be able to sample moreof the minority class cases, and also learn better and broader decision regions for the minority class. For details on the SMOTEBoost algorithm we refer the reader to Nitesh V. Chawla’s work [4]. The details of AdaBoost are in Y. Freund’s work [3]

**MODIFIED SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE-MSMOTE**

SMOTE generates synthetic examples by taking each minority class sample and introducing synthetic examples, whereas it doesn’t consider the distribution of minority class samples and latent noises in data sets. To improve the performance of SMOTE, a modified method-MSMOTE is proposed in the paper. The modified algorithm classifies the samples of minority class into three groups where are security samples, border samples and latent noise samples [8] by calculating the distances of all the samples. When MSMOTE generates synthetic examples, the different strategy for selecting near neighbors is used. The details are as follows: the algorithm randomly selects a data point from the k near neighbor for the security samples, selects a nearest neighbor for the border samples and does nothing for the latent noise

samples. The pseudo-code for MSMOTE is show in the table

**Performance**

The experimental results for all three data sets are presentedin Tables III to V and Figures 3 to 5. It is important to note that only the prediction performance for the minority classes from three data sets are reported by these tables, since prediction of the majority class was not of interest in this studyand we prefer to consider the performance of the minority class. Due to space limitations, the final values for recall,precision and F-value of minority class is presented whenMSMOTE with j48 ( a classification algorithm) 、 SMOTE with j48 、 SMOTEBoost and MSMOTEBoost (a hybrid MSMOTE /boosting algorithm, similar with SMOTEBoost ) are applied on three different data sets

Analyzing Tables III to V and Figures 3 to 5, it is apparent

that MSMOTE achieved higher F-values than SMOTE although the improvement varied with different data sets. The final value for precision, recall and F-value for the various

blue points in circle are noise samples [2]

**EXPERIMENTS**

**Datasets**

Our experiments were performed on three data sets summarized in Table II . For all data sets, except for thesatimage data set, the reported values for recall, precision and F-value were obtained by performing 10-fold cross-validation. For the satimage data set, however, the separate test data set that is supplied by UCI was used to evaluate the performance of the proposed algorithm. Because the satimage is multiclass dataset, the two-class data set is used in our experiment. We chose the smallest class as the minority class and collapsed the remaining classes into one class as was done in [7]. Thisprocedure gave us a skewed 2-class dataset, with 4020 majority class examples and 415 minority class examples. All data sets have all continuous features. The metrics such as precision, recall and F-value [8, 9] have been used to understand the performance of the learning algorithm on the minority class. We present the values of those metrics in the tables. methods at different amounts of SMOTE and MSMOTE (that is N%)with different classifier are shown in the Tables III to V. These reported values indicate that MSMOTE applied with the different classifier has the effect of improving the value for precision 、 recall and F-values of the minority class due to improved coverage of the minority class examples, In the conditions of the same N%, regardless of the concrete classifiers such as j48 and the boost, MSMOTE improves the value precision, recall and F-values. Shows according to the entire figure, their change tendency are the same. In the details,for the imbalance data set Wpbc, table V and figure [5], the evaluating indicators that are obtained from the original imbalance data set with the j48 classifier and boost are very low. Then F-values are bigger and bigger, with the N% is increased. While the F-values and N% are not positive proportion changes, in the table IV, when N%=300%, the F- values of the test data set is biggest. In the real application, thevalue of the N is considered in order to over-fitting. We have also compared SMOTEBoost, MSMOTEBoost and standard classifier, the value for recall, precision, F-value of minority class is enhanced when proposed methods are applied on standard UCI data set. According to ours experimental results, boost is also an effective method for imbalance data set from another perspective.

TABLE II. SUMMARY OF DATA SETS USED IN EXPERIMENTS

Dataset	Number of attribute	Number of classes	Number of majority class instances	Number of minority class instances
Pima	8	2	500	268
satimage	36	6	4020	415
Wpbc	32	2	147	46

TABLE III. THE VALUES FOR RECALL, PRECISION, F-VALUE MINORITY CLASS WHEN PROPOSED METHODS ARE APPLIED ON PIMA

INDIAN DIABETES DATA SET

Method		Precision	Recall	F-value	Method	Precision	Recall	F-value	
standard J48		0.632	0.597	0.614	standard AdaBoost		0.604	0.608	0.606
SMOTE and J48	N=100	0.762	0.793	0.858	SMOTE and AdaBoost	N=100	0.781	0.78	0.781
	N=200	0.831	0.887	0.777		N=200	0.844	0.898	0.87
	N=300	0.837	0.905	0.87		N=300	0.876	0.92	0.845
	N=500	0.884	0.944	0.913		N=500	0.907	0.957	0.931
MSMOTE and J48	N=100	0.801	0.804	0.803	MSMOTE and AdaBoost	N=100	0.808	0.823	0.815
	N=200	0.858	0.892	0.874		N=200	0.875	0.892	0.884
	N=300	0.905	0.91	0.907		N=300	0.896	0.915	0.905
	N=500	0.927	0.945	0.936		N=500	0.926	0.951	0.938

TABLE IV. THE VALUES FOR RECALL, PRECISION, F-VALUE MINORITY CLASS WHEN PROPOSED METHODS ARE APPLIED ON SATIMAGE DATA SET

Method		Precision	Recall	F-value	Method	Precision	Recall	F-value	
standard J48		0.586	0.55	0.567	standard AdaBoost		0.679	0.592	0.633
SMOTE and J48	N=100	0.58	0.564	0.572	SMOTE and AdaBoost	N=100	0.75	0.64	0.655
	N=200	0.558	0.569	0.563		N=200	0.687	0.645	0.665
	N=300	0.53	0.673	0.593		N=300	0.721	0.697	0.697
	N=500	0.539	0.621	0.577		N=500	0.716	0.682	0.659
MSMOTE and J48	N=100	0.619	0.569	0.593	MSMOTE and AdaBoost	N=100	0.711	0.607	0.691
	N=200	0.588	0.569	0.578		N=200	0.716	0.621	0.665
	N=300	0.621	0.645	0.633		N=300	0.739	0.659	0.708
	N=500	0.552	0.607	0.578		N=500	0.678	0.64	0.699

TABLE V. THE VALUES FOR RECALL, PRECISION, F-VALUE MINORITY CLASS WHEN PROPOSED METHODS ARE APPLIED ON SATIMAGE DATA SET

Method		Precision	Recall	F-value	Method	Precision	Recall	F-value	
standard J48		0.235	0.087	0.127	standard AdaBoost		0.344	0.239	0.282
SMOTE and J48	N=100	0.552	0.63	0.589	SMOTE and AdaBoost	N=100	0.667	0.652	0.659
	N=200	0.681	0.681	0.681		N=200	0.804	0.804	0.804
	N=300	0.704	0.815	0.756		N=300	0.796	0.87	0.831
	N=500	0.801	0.862	0.831		N=500	0.846	0.895	0.87
MSMOTE and J48	N=100	0.625	0.534	0.581	MSMOTE and AdaBoost	N=100	0.708	0.685	0.696
	N=200	0.802	0.703	0.749		N=200	0.782	0.804	0.793
	N=300	0.813	0.777	0.794		N=300	0.848	0.848	0.848
	N=500	0.938	0.884	0.91		N=500	0.919	0.909	0.914

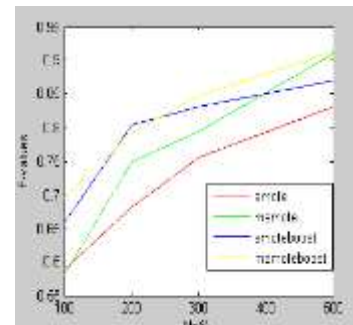
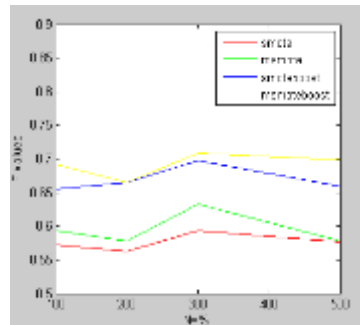
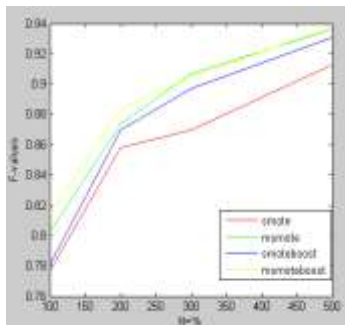


Figure 3. F-values for the minority class when the smote, msmote, smoteboost, msmoteboost algorithm is applied on the Pima dataset.

## CONCLUSION

In this work we present MSMOTE, a modified technique for learning from skewed datasets. MSMOTE is a variant of the SMOTE algorithm, for improving the performances of model for the minority class MSMOTE not only considers the distribution of minority classes, but also rejects latent noise spots based on k-nn classifier method. Experimental results from several UCI imbalanced data sets indicate that the proposed MSMOTE algorithm can result in better prediction of minority class than SMOTE. We combine the MSMOTE and AdaBoost and propose MSMOTEBoost. Our experiments have also shown that MSMOTEBoost is able to achieve higher F-values than SMOTEBoost.

Although the experiments have provided evidence that the proposed method can be successful for learning from imbalanced data sets. The MSMOTE still has some drawbacks, for example, it doesn't considerate the differences of importance features [10]. Because the features are crucial to the performances of model. Future work is needed to address this problem.

## REFERENCES

- [1] N.V.Chawla, K.K.W. Bowyer, L. O. Hall, W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-Sampling Technique", Journal of Artificial Intelligence Research, 2002, vol. 16, 321-357
- [2] SCHAPIRE, Robert E., "A Brief Introduction to Boosting. In", Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann, 1999, pp. 1401-1406.
- [3] Y. Freund and R. Schapire. "Experiments with a new boosting algorithm". In Proceedings of the Thirteenth International Conference on Machine Learning, , 1996, pages 148-156
- [4] Chawla Lal N V, Lazarevic A, Hall O. "SMOTEBoost : improving prediction of the minority class in bosting". The 7th European Conf on Principles and Practice of Knowledge Discovery in Databases. Berlin : Springer, 2003 : 107-119.
- [5] Zhou Z H, Liu X Y. Training "cost-sensitive neural networks with methods addressing the class imbalance problem". IEEE Trans Knowl Data Eng, 2006, 18(1) : 63-77
- [6] Yen S J, Lee Y S. "Cluster-based under-sampling approaches for imbalanced data distributions" Proceedings of the 8th International Conference. Berlin : Springer, 2006 : 427-436. ELENA project, ftp.dice.ucl.ac.be in directory pub/neural-nets /ELENA/databases
- [7] M.Buckland, F. Gey, "The Relationship between Recall and Precision", Journal of the American Society for Information Science, 1994. 45(1):12-19
- [8] M.Joshi, V. Kumar, R. Agarwal, "Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements", First IEEE International Conference on Data Mining, San Jose, CA, 2001
- [9] M.Buckland, F. Gey, "The Relationship between Recall and Precision", Journal of the American Society for Information Science, 1994. 45(1):12-19