Imbalanced Time Series Data Classification Using Oversampling Technique

Miss. Reshma K. Dhurjad **Prof. Mr. S. S. Banait**

***Abstract* — Data imbalance is a major source of performance degradation in data mining and machine learning. Existing learning algorithms assume a balance class distribution, with approximately equal number of learning instances for each class but in many real-world scenarios, the data available for learning are extremely imbalanced. The consequence of this imbalanced data is that learning algorithms tend to bias toward the less important negative class with larger population. Due to sequential nature of time series data, variable which are close by in a time series are extremely correlated in many cases. So, to preserve this correlation structure properly and to solve imbalanced learning issue, an enhanced structure preserving oversampling technique along with Majority Weighted Minority Oversampling Technique is used to re-establish the class balance. Then the time series classifier is learned from the balanced data-set. This research can be used to develop an efficient classification learning algorithm which provides a better accuracy as compared to existing methods for imbalanced time series data.**

***Key Words*— Classification, Imbalanced data, learning, oversampling, time series.**

I. INTRODUCTION

Data imbalance is a major source of performance degradation in data mining and machine learning. Imbalanced datasets means a dataset whose classification categories are unequally represented. The level of imbalance can be as large as 1:99. It is known that class imbalance is emerging as a crucial issue in designing classifiers. Furthermore, the class with the few number of instances is usually the class of interest for the purpose of learning task. This problem is of great interest because it turns up in many real-world classification problems, such as remote-sensing, pollution detection, risk management, fraud detection, and especially medical diagnosis. For example, in case of earthquakes the known samples are rare but certainly of more important for earthquake prediction than the ample normal samples. In such situations positive class samples are sparsely distributed and negative class samples are densely distributed. Existing algorithms assume a balanced class

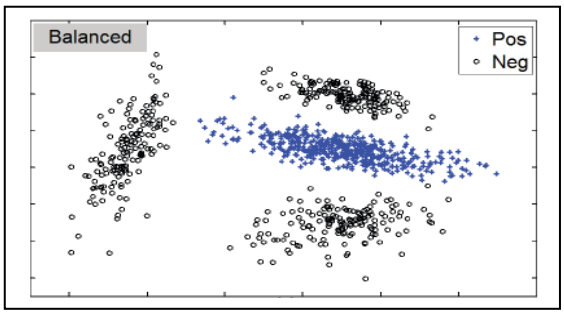

Fig. 1. Balanced Data Distribution

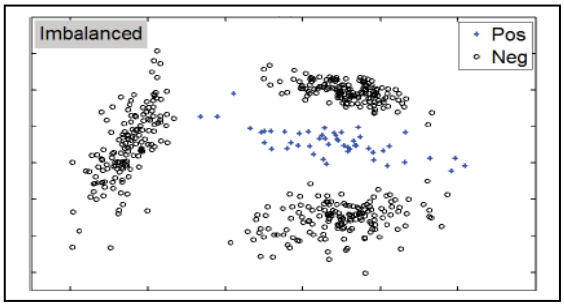

Fig. 2. Imbalanced Data Distribution

distribution, with sufficient and approximately equal number of learning instances for each class, as illustrated in Fig. 1. In case of imbalanced Data, the minority class, containing only a few sparsely distributed instances, is outnumbered by the majority class, as shown in Fig. 2. In such scenarios, the learning algorithms tend to bias toward the less important negative class with larger instances.

Imbalanced datasets learning issue can be handle with two ways first is data level and the second one algorithmic level. The techniques at data level alter the distribution of the samples in imbalanced data sets, and then it is given to the learner. The techniques at the algorithm level alter the present data mining algorithms or put up new algorithms to solve the imbalance problem. They enforce emphasis on the minority class by manipulating and incorporating learning parameters. The main advantage of data level methods is that they are not dependent on the underlying classifier. For learning from imbalanced two class data sets, oversampling is to balance the class distribution through increasing the minority class with synthetically generated instances. To solve imbalanced learning issue oversampling of minority class is done. To solve imbalanced learning issue, various oversampling

methods were proposed like SMOTE [1], Borderline-SMOTE [4], ADASYN [7], SPO [10], INOS [12], DataBoost [2], so that a class balance is re-establish.

This paper is composed further as: Section 2 provides survey of literature along with pros and cons of some the existing methods. CIL (Class imbalance learning) algorithm with the main contribution of this paper is described in section 3. Section 4 reports the data sets and results. Finally, section 6 summarizes this paper and raises issues for future work.

II. LITERATURE SURVEY

In year 2002, N.V. Chawla [1] introduced SMOTE: Synthetic Minority Oversampling Technique. This work shows that a combination of method under sampling the majority class and oversampling the minority class can accomplish better classifier performance than only under sampling the majority class. SMOTE provides a new method to oversampling. SMOTE and under-sampling in combination achieves better performance than plain under-sampling.

Based on SMOTE technique H. Han, W.Y. Wang [4] were proposed two novel minority oversampling techniques, borderline-SMOTE1 and borderline-SMOTE2. In this only the minority instances near the borderline are over-sampled. For the minority class, experiments show that borderline-SMOTE approach achieve better performance than SMOTE and random over-sampling methods. The samples on the borderline and the ones nearby are more likely to be not categorized properly than the ones farthest from the borderline, and therefore more important for classification. Authors thus present two new minority oversampling techniques, borderline-SMOTE1 and borderline-SMOTE2. These techniques only over-sample borderline instances of the minority class.

In year 2008, ADASYN Approach for Imbalanced Learning is proposed by Haibo He et al.[7]. They have presented a new adaptive synthetic sampling technique for learning from imbalanced datasets. The necessary idea of this method is to use a weighted distribution for minority class instances according to the level of difficulty in learning, in which more synthetic samples is produce for minority class instances that are difficult to learn as compared to those minority instances that are not difficult to learn. ADASYN method increase learning performance in two ways: (a) adaptively shift the classification decision boundary toward the hard to learn instances. (b) decreasing the bias which is introduced by the class imbalance. They focus on the two-class classification issue for imbalanced data sets, an issue of main focus in recent research activities in the research community.

In year 2004,Hongyu Guo et al.[2] have proposed DataBoost-IM method which generates the features of the synthetic instances individually. Databoost produce feature value based on Gaussian distribution within a range. In this work, they have described a new technique that combines an ensemble-based learning algorithm, and boosting with data generation to increase the estimation power of classifiers against imbalanced data-sets including two classes. In the DataBoost-IM technique, difficult instances from both the classes are identified during execution of the algorithm. Subsequently, the difficult instances are used to separately generate synthetic instances for both the classes. The synthetic samples are then added to the training set, and the class distribution and a weights of the various classes in the new training set are re-balanced.

In 2011, Structure Preserving Oversampling technique for Imbalanced Time Series Classification has proposed by Hong Cao et al. [10]. This work presented a novel structure preserving oversampling technique for categorizing imbalanced time series data. This method generates synthetic minority instances based on multivariate Gaussian distribution by regularizing the unreliable Eigen spectrum and forecasting the covariance structure of the minority class. By creating variances in the trivial Eigen feature dimensions and maintaining the covariance structure, the synthetic instances spread out effectively into the void region in the data space and it is not closely bind with existing minority class instances.

In year 2013, David Woon et al. [12] suggested Integrated Oversampling (INOS) for Imbalanced Time series Classification [9]. They concentrate on the problem of Imbalanced learning issue. To address this issue they introduce a new technique of oversampling i.e., Integrated Oversampling method. They have noted that, the interpolation based approach work well with imbalanced learning issue, but the problem with that is, they are not sufficient for the task of oversampling imbalanced time series data sets. David Woon et al. designed an Integrated Oversampling (INOS) method with two aims in mind: First is to preserve the regularized Eigen co-variance structure which can be predicted using the limited positive time series instances, and the second is to be able to provide enough emphasis on the key minority instances with the remaining oversampling capacity. For the first objective, a new enhanced structure preserving oversampling (ESPO) has proposed. For the second objective, they use an interpolation based method to produce a small percentage of synthetic instances so as to emphasize on the borderline set of existing instances, which are critical for building accurate classifiers in the subsequent steps.

In year 2014, Sukarna Barua et al. [13] suggested MWMOTE method. This work shows that most of the previous oversampling methods may generate the false synthetic minority instances in some situations and make learning tasks difficult. To this end, a novel technique, called Majority Weighted Minority Oversampling Technique, is presented for dealing with imbalanced learning task. MWMOTE identifies the hard-to-learn minority class samples and give them value according to Euclidean

distance from the nearest negative class instances. It then generates the synthetic instances from the weighted informative minority class instances using a clustering technique.

III. PROPOSED WORK

A) Proposed System:

The block diagram of proposed oversampling system is shown in Fig. 3.
Step by step architecture is explained in subsequent section.

1) Structure Preserving Oversampling

Hong Cao, Xiao-Li Li, David Yew-Kwong Woon, and See Kiong Ng proposed an enhanced structure preserving oversampling method which does 1) Estimation of Data Covariance and Eigen spectrum; 2) Generate a synthetic positive sample.

I) Estimation of Data Covariance and Eigen spectrum:

Major steps in this process are as follows:

1. Covariance estimation for Positive class.
2. Eigen decomposition to obtain the covariance structure V= [v1, v2] and spectrum D = Diag(d1, d2).
3. Eigen Spectrum Regularization
   Small Eigen values due to insufficient training samples often cause poor learning outcomes.
4. Eigen Spectrum Regularization steps:
   - Determine M through cross validation;
   - Find ES model parameters using reliable spectrum;
   - Apply the ES model to regularize the unreliable region.

II) Generate a Synthetic Positive Sample

1. Randomly generate vector *Z* according to multivariate Gaussian distributions with N (0_m, I_m) and N (0_{m-M}, I_{m-M}).
2. Transform with Regularized Spectrum

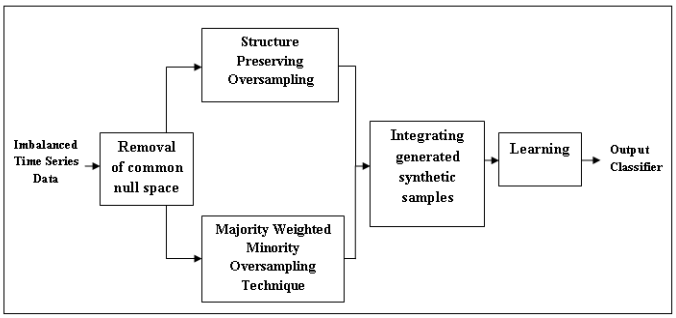

Fig. 3. Block diagram of the proposed oversampling framework

III) Cleaning:

If generated sample does not fall into the negative-class territory, keep it. Otherwise, remove new generated sample.

2) Majority Weighted Minority Oversampling Technique :

ADASYN Oversampling method try to avoid the imbalanced learning problem by adaptively assigning weights to the minority class samples. To assign the weight, ADASYN use a parameter δ, defining the number of the majority class samples among the k-nearest neighbors of the minority class sample. A large δ gains the weight, while a small δ lessens it. However, the use of δ for assigning the weights may encounter the following problems:

1. The parameter δ is inappropriate for assigning weights to the minority class samples located near the decision boundary.
It is shown in Fig. 4.

2. The parameter δ is insufficient to distinguish the minority class samples with regard to their importance in learning.

3. The parameter δ may favor noisy instances.

In the sample generation phase, previous synthetic oversampling methods like ADASYN employ the k-nearest neighbor based approach. To generate a synthetic sample from an original minority class sample, say x, the k-NN-based approach randomly selects another minority class sample, say y, from the k-nearest neighbors of x. The approach then generates a synthetic instance, g, by using the linear interpolation of x and y. This can be expressed as

$$g = x + (y-x) \times \alpha \quad (1)$$

Where α is a random number in the range 0 to 1. The above equation says that g will lie in the line segment between x and y. However, in many situations, the k-NN-based approach may generate wrong minority class instances. To show why, consider Fig. 5. Assume, one want to generate a synthetic sample from the minority class sample A and k = 5

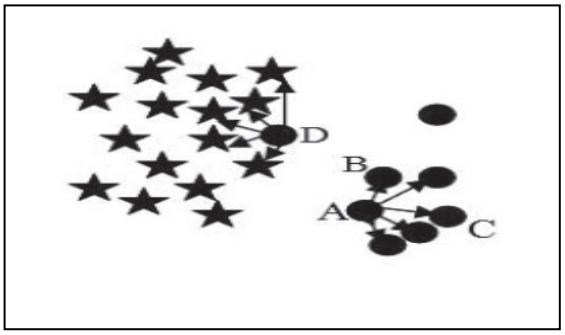

Fig. 4. K-nearest neighbors of A and D are shown by arrow (for k=5).

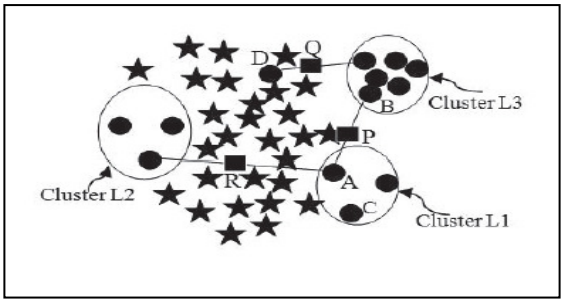

Fig. 5. Several minority clusters in data set. Synthetic samples generated are shown by square.

in this case. The k-NN-based approach will randomly select another minority class sample from the five-nearest minority neighbors of A. Let the instance B is selected. According to (1), the linear interpolation of A and B may generate a synthetic sample, say P, shown by the square in Fig. 5. It is shown in Fig. 5. that P is clearly a wrong minority class sample because it overlaps with a majority class sample. This overlap will definitely make learning tasks harder. Let one like to generate a synthetic instance from the noisy instance D. In this situate ion, the synthetic instance Q may be generated, which will suppose to be noisy and lie inside the majority class region (Fig. 5).

In summary one can say that, the k-NN based Sample generation approach is inappropriate in some cases. This approach may generate duplicate and wrong synthetic minority class samples from the members of dense and small- sized clusters, respectively. It is important to note that when the k-NN-based approach generates the synthetic samples from the noisy samples, the generated samples will also be noisy. The above mentioned problems occur due to the fact that the approach uses all the k-nearest neighbors blindly without considering the location and distance of the neighbors from the minority class sample under consideration.

To avoid this problem, a novel minority oversampling method i.e. MWMOTE is used in proposed method instead of ADASYN along with Enhanced structure preserving oversampling. The objective of this technique is that: to improve the sample selection scheme and to improve the synthetic sample generation scheme. MWMOTE includes three key phases. In the first phase, MWMOTE identifies the most important and hard-to-learn minority class samples from the original minority set, S_{min} and construct a set, S_{imin}, by the identified samples. In the second phase, each member of S_{imin} is given a selection weight, S_w, according to its importance in the data. In the third phase, MWMOTE generates the synthetic samples from S_{imin} using S_ws and produces the output set, S_{omin}, by adding the synthetic samples to S_{min}.

Algorithm: MWMOTE (S_{maj}, S_{min})
Input:
1) S_{maj}: Set of majority class samples
2) S_{min}: Set of minority class samples
Procedure Begin:
1) For each minority instance $x_i \in S_{min}$, compute the nearest neighbor set.
2) Construct the filtered minority set, S_{minf} by removing those minority class samples which have no minority instance in their neighborhood
3) For each $x_i \in S_{minf}$, compute the nearest majority set, $N_{maj}(x_i)$.
4) Find the borderline majority set, S_{bmaj}, as the union of all $N_{maj}(x_i)$s.
5) For each majority instance $y_i \in S_{bmaj}$, compute the nearest minority set, $N_{min}(y_i)$.
6) Find the informative minority set, S_{imin}, as the union of all $N_{min}(y_i)$s.
7) For each $y_i \in S_{bmaj}$, and for each $x_i \in S_{imin}$, compute the information weight, $I_w(y_i,x_i)$.
8) For each $x_i \in S_{imin}$, compute the selection weight $S_w(x_i)$.
9) Convert each $S_w(x_i)$ into selection probability, $S_p(x_i)$.
10) Find the clusters of Smin. Let, M clusters are formed which are $L_1,L_2,\ldots..,L_M$.
11) Initialize the set, $S_{omin}= S_{min}$.
12) Do for j=1…..N.

    a) Select a sample x from S_{imin} according to probability distribution $\{S_p(x_i)\}$. Let, x is a member of the cluster L_k, 1<=k<=M.
    b) Select another sample y, at random, from the members of the cluster L_k.
    c) Generate one synthetic data , s, according to s= x+ α × (y –x), where α is a random number in the range [0,1].
    d) Add s to S_{omin}: $S_{omin}= S_{omin} \cup \{s\}$.

13) End Loop
End
Output: The output oversampled minority set, S_{omin}.

3) Learning

To test the performance of proposed system, SVM learning algorithm is used. With the balanced data set, SVM is used to learn a time series classifier. SVMs can generate accurate and robust classification results on a sound theoretical basis, even when input data is non-linearly separable. So they can help to evaluate more relevant information in a better way.

IV. RESULTS AND DISCUSSION

All Experimentation is performed using intel core i3 processor and 4 GB RAM. The operating system Windows 8 (64-bit) with Python.

a) Datasets:

UCR time series repository is used for experiment purpose, as tabulated in Table I.

Each row here shows a standalone binary imbalanced learning task. To convert them into two-class data sets, randomly one class is selected as the positive class and remaining classes are use to form the negative class. High imbalance ratios(IM-ratios) should be maintain.

TABLE I. IMBALANCED TIME SERIES DATASETS

Datasets	Training			Time Series Length
	Positive	Negative	IM-Ratio	
Adiac	10	380	38	176
FaceAll	50	1000	20	131
50Words	10~50	400	8~40	270
SLeaf	35	450	12.9	128
TwoPats	50	1800	36	128
Wafer	50	380~3000	7.6~60	152
Yoga	50	800~900	16~18	426

b) Performance Bounds:

The performance of the classifier is evaluated using various metrics like Precision, Recall. Precision and Recall are well-known performance metrics for imbalanced learning. These metrics are calculated using a confusion table as shown in Table II.

TABLE II. CONFUSION METRICS TABLE

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

In proposed system the high recall rate shows that most of the test minority instances can classified with a good accuracy. The proposed work shows that the MWMOTE (Majority Weighted Minority oversampling technique) is also quite effective for handling highly imbalanced time series classification task. The proposed system produces most of the synthetic positive instances following the covariance structure of existing positive instances.

c) Results:

To test the performance of the output classifier a test data-set can be used as an input. Classifier will predict the classes of test data-set. And by comparing predicted classes with actual classes the performance of classifier can be determined. Datasets and their corresponding SVM performance are shown in Table III for proposed oversampling method.

TABLE III. SVM PERFORMANCE COMPARISON FOR DIFFERENT DATASETS USING PROPOSED OVERSAMPLING MRTHOD

Evaluation Measure	**Dataset**	**Oversampling Methods**	
		INOS (Integrated Oversampling)	Proposed System
Precision	Adiac	0.82	0.98
	FaceAll	0.98	0.74
	50Words	0.80	0.92
	SLeaf	0.87	0.97
	TwoPats	0.97	0.97
	Wafer	0.99	1.00
	Yoga	0.98	0.93
Recall	Adiac	0.79	0.78
	FaceAll	0.89	0.77
	50Words	0.79	0.75
	SLeaf	0.93	0.94
	TwoPats	0.38	0.97
	Wafer	0.97	1.00
	Yoga	0.57	0.92

According to Table III proposed method classify test data more accurately for Wafer dataset as compare to other datasets.

V. SUMMARY AND CONCLUSION

An Imbalanced learning issue is addressed using Oversampling the minority (positive) class which preserves the main covariance structure of the original positive class instances.

The large portion of oversampling can be done using structure preserving oversampling (SPO). In parallel, the remaining synthetic samples are also created using a Majority Weighted Minority Oversampling Technique (MWMOTE) with an aim to protect the existing instances that are hard to classify. To test the performance of this system, SVM learning algorithms is used. The oversampling

approach effectively addresses the challenging issue of data imbalance in time series classification.

ACKNOWLEDGMENT

The authors would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. We are thankful to the authorities of Savitribai Phule Pune University and concern members of IJECSCSE conference, organized by KKWIEER, Nashik for their constant guidelines and support. We are also thankful to reviewer for their valuable suggestions. We also thank the college authorities for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to friends and family members.

REFERENCES

[1] N.V. Chawla, k.w. Bowyer, l.o. Hall, and w.p. Kegelmeyer, smote: synthetic minority over-sampling technique vol. 16, j. Artificial intelligence, pp. 321-357, 2002.

[2] h. Guo and h.l. Viktor, learning from imbalanced data sets with boosting and data generation: the databoost-im approach, vol. 6, no. 1, acm sigkdd explorations newsletter, 2004.

[3] n.v. Chawla, n. Japkowicz, and a. Kolcz, editorial: special issue on learning from imbalanced data sets, vol. 6, no. 1, acm sigkdd explorations newsletter, pp. 1-6, 2004.

[4] h. Han, w.y. Wang, and b.h. Mao, borderline-smote: a new over-sampling method in imbalanced data sets learning, proc. International conf. Intelligent computing, pp. 878-887, 2005.

[5] e. Keogh, x. Xi, l. Wei, and c.a. Ratanamahatana, "ucr time series classification/clustering page," http://www.cs. Ucr.edu/~eamonn/time_series_data, 2006.

[6] y. Sun, m.s. Kamel, a.k.c. Wong, and y. Wang, "cost-sensitive boosting for classification of imbalanced data," pattern recognition, vol. 40, no. 12, pp. 3358-3378, dec. 2007.

[7] h. He, y. Bai, e.a. Garcia, and s. Li, adasyn: adaptive synthetic sampling approach for imbalanced learning, proc. International conf. Neural networks, pp. 1322-1328, 2008.

[8] h. He and e.a. Garcia, learning from imbalanced data,vol. 21, no. 9, ieee trans. Knowledge and data eng., pp. 1263-1284, sept. 2009.

[9] x.-y. Liu, j. Wu, and z.-h. Zhou, "exploratory undersampling for class-imbalance learning," ieee trans. System, man and cybernetics, vol. 39, no. 2, pp. 539-550, apr. 2009.

[10] h. Cao, x.-l. Li, y.-k. Woon, and s.-k. Ng, spo: structure preserving oversampling for imbalanced time series classification, proc. International conf. Data mining, pp. 1008-1013, 2011.

[11] M.N. Nguyen, X.-L. Li, and S.-K. Ng, Positive Unlabeled Learning for Time Series Classification, Proc. International Joint Conf. Artificial Intelligence, July 2011.

[12] Hong Cao, Xiao-Li Li, David Yew-Kwong Woon, and See-Kiong Ng, Integrated Oversampling for Imbalanced Time Series Classification, Vol. 25, No. 12, IEEE Trans. Knowledge and Data Eng., pp.2809-2822, 2013.

[13] Sukarna Barua, Md. Monirul Islam, Xin Yao, Fellow, IEEE, and Kazuyuki Murase, MWMOTE—Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning, Vol. 26, No. 2, IEEE Trans. Knowledge and Data Eng.,, pp. 405-425,2014.

AUTHOR'S PROFILE

Reshma K. Dhurjad received the B.E degree in Computer Engineering from K.K.Wagh Institute of Engineering Education & Research, Nasik, Savitribai Phule Pune University in 2012. Now pursuing M.E. from K.K.Wagh Institute of Engineering Education & Research, Nasik, India.

Prof.S.S.Banait, Assistant Professor, Department of Computer Engineering, K. K. Wagh Institute of Engineering Education & Research, Nasik, India.