International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Multi-Label Classification with PSO based Synthetic Minority Over-Sampling Technique (Psosmote) for Imbalanced Samples

M.Priyadharshini, L.Pavithira,

Abstract: ***Recently, the learning from unbalanced data has emerged to be a pre-dominant problem in several applications and in that multi label classification is an evolving data mining task, learning from unbalanced multilabel data is being examined. However, the available algorithms-based SMOTE makes use of the same sampling rate for every instance of the minority class. This leads to sub-optimal performance. To deal with this problem, a new Particle Swarm Optimization based SMOTE (PSOSMOTE) algorithm is proposed. The PSOSMOTE algorithm employs diverse sampling rates for multiple minority class instances and gets the fusion of optimal sampling rates and to deal with classification of unbalanced datasets. Then, Bayesian technique is combined with Random forest for multi-label classification (BARF-MLC) is to address the inherent label dependencies among samples such as ML-FOREST classifier, Predictive Clustering Trees (PCT), Hierarchy of Multi Label Classifier (HOMER) by taking the different metrics including precision, recall, F-measure, Accuracy and Error Rate.***

Keywords: multi-label classification, multi-class imbalance, PSO, SMOTE, Bayesian approach. (DrNGPASC 2019-20 CS016)

I. INTRODUCTION

Several real-world applications, like text classification and sub cellular localization of protein sequences, deal with multi-label classification with unbalanced data. The classification of unbalanced data is a significant issue in machine learning and data mining [1]. In an unbalanced dataset, there are considerably lesser training instances of one class in comparison with another class. Accordingly, the former is called as the minority class, and the latter is known as the majority class. In [2], the imbalance problem for MLC is addressed and a novel scheme known as DEML is proposed. In [3] an algorithm BSHD (Block Sampling with choosing the Highest Degree nodes), an active learning based imbalanced networked multi-label classification algorithm is proposed.

In [4] a multi -label classification algorithm that depends on multi-rank neighbors is introduced. In [5] the random walk model is combined with multi -label learning to introduce a multi -label classification algorithm MLRW (Multi-Label Random Walk algorithm).

Manuscript published on November 30, 2019.
* Correspondence Author
M.Priyadharshini*, Assistant Professor, Department of CT, Dr.N.G.P. Arts and Science College, Coimbatore.
Dr.L.Pavithira, Associate Professor, Department of Computer Applications, CIMAT, Coimbatore.

In [6] the asymmetric stage-wise loss function is presented to move the positive class samples at some distance away from the classification boundary compared to the negative class samples by adjustment of the ramp in addition to the margin parameters. In [7] LEML algorithm is used, which is the low-rank property of the label matrix to develop a linear prediction model and then it helps in restoring the missing labels by reducing the kernel norm.In [8] the low-rank hypothesis is combined with manifold hypothesis, and then the proximal gradient descent algorithm is used for recovering the missing labels. In [9] the existing correlation among labels is completely exploited, and a considerably good data subset is selected with the help of cross-validation technique, and its prediction results is used in the next subsequent iteration, and eventually all the missing labels are recovered. But, this technique presumes that the training set gets balanced between positive and negative categories. In [10] different mechanisms are introduced and they are compared for the generation of synthetic samples for balancing the data sets during the training of multi-label algorithms. In [11] the SCUT hybrid sampling technique is brought into use and it is utilized for balancing the number of training examples in such a kind of multi-class environment. In [12] the challenge occurring due to the multiclass imbalance problems is studied and the generalization capability of few ensemble solutions, including the recently introduced algorithm Adaboost is also investigated. In [13] the process of synthetic instance production for multilabel datasets (MLDs) and MLSMOTE (Multilabel Synthetic Minority Over-sampling Technique), which is a novel algorithm targeted at the generation of synthetic instances for unbalanced MLDs, is presented. In this research work, a new PSO based on SMOTE algorithm, called as PSOSMOTE is introduced for multi-lable classification for unbalanced data to boost the performance of unbalanced data classification. The PSOSMOTE algorithm makes use of multiple sampling rates for various minority class instances and gets the combination of optimal sampling rates. Then, the newly introduced MLC is used on the dataset. The remaining portion of this work is organized as below. Section 2 explains about the proposed technique. Section 3 discusses about the data sets, the experimental setup and experimental results. In the last section, the conclusions are discussed in Section 4.

II. PROPOSED METHODOLOGY

In this research work, PSOSMOTE is used for imbalanced dataset sampling. PSOSMOTE has combined both PSO and SMOTE process.

Then, Bayesian scheme is merged with Random forest (BARF-MLC) that will be used for revealing the inherent label dependencies. The overview of the proposed scheme is illustrated in figure 1.

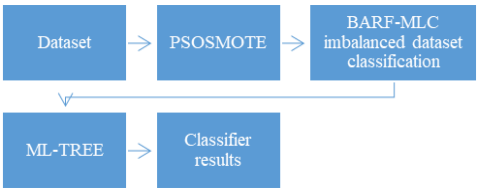

Figure 1. Architecture of proposed imbalanced classification

2.1. PSOSMOTE

The proposed PSOSMOTE algorithm makes use of a PSO algorithm to get the optimized sampling rates and then creates a new dataset by over-sampling making use of the optimized sampling rates. The algorithm comprises of five steps, which are explained as follows.

Step 1. Encoding and Initialization: In this step, a particle with size P gets generated to be used for the PSO algorithm. Let $N_i$represent the sampling rate of the minority class instance x_i. Then in the pretext of a PSO algorithm, an individual in a particle is used for the representation of a combination of the sampling rates for every instance as given, $X_j = (N_1^j, N_2^j, \ldots N_M^j), j = 1, 2, \ldots, P$,where M refers to the length of particles, which is the number of minority class instances, and P stands for the particle size. For initializing an individual, the position and velocity of every node of the chromosome is set.

Step 2. Fitness computation: In this step, the fitness function value for every individual in the particles is calculated, and then that value is compared with the PBest.In case the condition of fitness value is better compared to pBest then the current fitness is assigned to be new pBest, otherwise the earlier pBest value is maintained as best. The sample for Train based on $X_i$is considered to obtain SmotedTraini and then classification is carried out by Smoted Train that is a training set and then trained in the form of a testing set. After this, the classification index G-mean value is taken as the i-th individual of fitness function value pBest$i$ (i=1,2,…P). Then the final pBEst value is assigned in order to obtain gBest.

Step 3. Velocity computation: The velocity value is computed as per the distance of the individual's data from the target. The more the distance, the higher is the velocity value. In case the data is a pattern or sequence, then the velocity would explain how dissimilar the pattern is from the target, and therefore, how much it requires to be modified in order to be matched with the target. The pBest value of each particle just specifies the nearest the data has ever reached to the target since the start of the algorithm.

Step 4. Update: The gBest value is only changed if any particle's pBest value comes near to the target rather than gBest. With every iteration of the algorithm, gBest slowly moves more near to the target till one of the particles attains the target.

Step 5. Termination: Check If the termination condition is satisfied, target or maximum epochs is attained, and if this condition is satisfied, then the algorithm provides the result in the form of the optimal sequence of oversampling rate N for SMOTE; else, return to Step 2. Once the search of the optimal sampling rates is stopped, the dataset is produced using the SMOTE over-sampling employing the optimal sampling rates given in [11].

2.4. Bayesian approach with Random forest for multi-label classification (BARF-MLC)

BARF-MLCscheme is used for multi-label classification. The proposed work at first employs the classifier tree construction algorithm and label transfer technique for dealing with label dependencies. Then, classifier trees are integrated with the forest using the new ensemble framework to improve the prediction performance, and the computational complexity of the proposed algorithm. A new hierarchical tree algorithm, known as ML-TREE specified in Algorithm 1, presumes the intrinsic label dependency in a hierarchical way.

Algorithm 1. ML-TREE

Input: A data set D, and a relevant label vector b = none

Output: A hierarchical multi-label tree

1: (b, h, P) = SPLITTEST(D, b)

2: **if** h≠ none ^ Acceptable(P) **then**

3: **for** $D_i \in$ P **do**

4: treei=ML-TREE(D_i, b)

5: **end for**

6: **return** node(h, b, U_i{$tree_i$})

7: **else**

8: **return** leaf(h, b)

9: **end if**

With no loss of generality, linear Bayesian technique is used in the form of base classifier, and the thresholding value λ= 0.9 is set as the default value.

2.4.1. Bayesian approach

Bayesian networks are used for specifying the labels joint distribution space that is conditioned on feature space, and is a strong random order modeling of label relationships. Multi-dimensional Bayesian network classifier is a Bayesian network (BN) of restrained topology aimed at resolving the multi-dimensional (along

multi-label) classification problem.

2.4.2. Random forest

This research work addresses the text classification problems where the comments of dataset have to be categorized into predetermined classes. Feature set of those domains includes the word occurrences, which lead to dimensionality problem. Therefore, decision tree learning approach called as random forest is found to be a suitable approach since it produces diversity by training on multiple data subsets and feature sets. In machine learning, diversity is a factor to be considered for constructing good Random Forests (RF), where bagging is combined with random feature selection for decision trees. Dataset in random forest is then trained with Bayesian technique for efficient multi-label classification.

III. RESULTS AND DISCUSSION

In this research work, the performance of proposed BARF-MLC is then compared with earlier PCT, HOMER and ML-FOREST algorithm. Using the table 1 provided below, all the Imbalanced data sets with imbalance ratio ranging between1.5 and 9 is found. For each data set, its name and its number of instances, attributes (Real/Integer/Nominal valued) and imbalance ratio value are determined. Eachdata file depicts a subsequent structure as shown in table 1.

Table 1: Details of imbalanced dataset

NAME	ATTRIBUTES (R/I/N)	EXAMPLE	IMBALANCE RATIO
Pima Indians	8 (8/0/0)	768	1.87
Yeast 1	8 (8/0/0)	1484	2.46
New – thyroid 1	5 (4/1/0)	215	5.14

This research work makes use of three different types of datasets, which include: (1) Pima Indians, (2) Yeast and (3) Thyroid are imbalanced data type. The ultimate goal is to find the onset of diabetes in Pima Indians within five years using machine learning. A model whose performance is good could help in focusing on the preventive measures to the impacted. The link can be referred for description of the entire dataset. https://sci2s.ugr.es/keel/imbalanced.php?order=featR#sub10

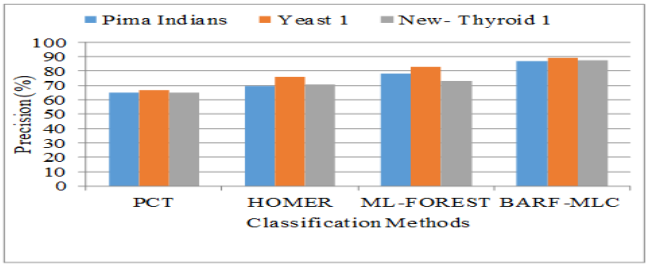

Figure 2 Comparison of various classifiers is based on precision

Evaluation on different classifiers is performed in terms of precision based on percentage is shown in figure 2. The proposed BARF-MLC approaches render the precision value to be 86.9565%, in Pima Indians that is greater compared to the earlier approaches, whereas for the other approaches for the same prima Indians: PCT gives 65.2173%, for HOMER renders 69.5652% and ML-FOREST yields 78.2608%.

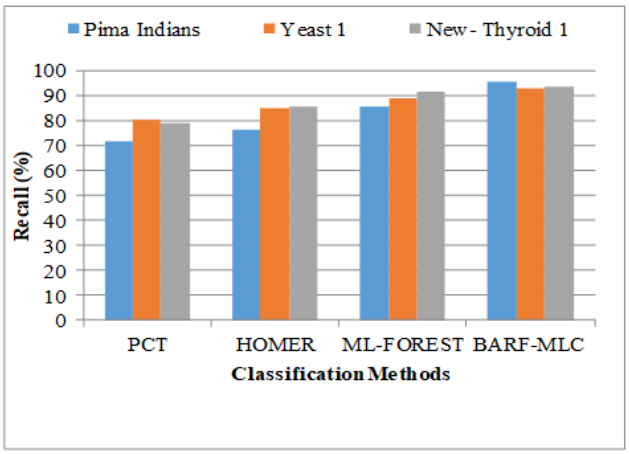

Figure 3 Comparison of various classifiers in recall

Figure 3 shows the evaluation on different classifiers in terms of recall based on percentage. The proposed BARF-MLC techniques provides are call value of 92.5925% for yeast 1 that is much greater compared to the earlier approaches, whereas for other methods for the same yeast 1: PCT gives 84.6153%, for HOMER yields 84.6153% and ML-FOREST renders 88.8888%.

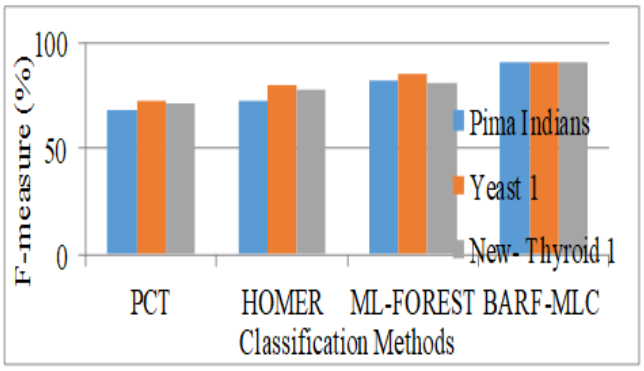

Figure 4 Comparison of various classifiers in f-measure

Figure 4 shows the evaluation on different classifiers in terms of f-measure based on percentage. The proposed BARF-MLC approaches yield an f-measure value to be 90.3225% new-thyroid that is higher compared to the earlier approaches, whereas other methods: PCT yields 70.9677%, for HOMER its value is 77.4193% and ML-FOREST renders 81.25%.

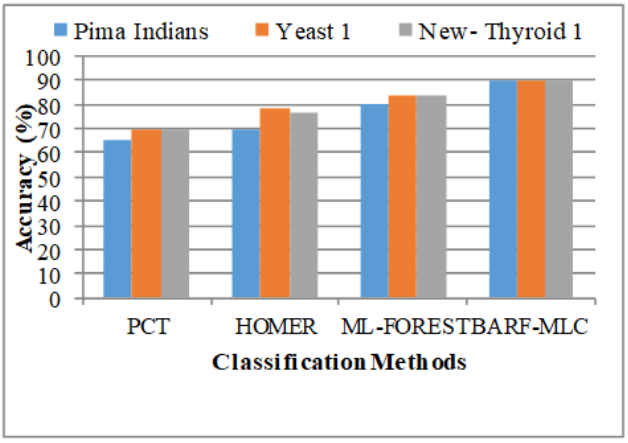

Figure 5 Comparison of various classifiers in accuracy

Figure 5 shows the evaluation on different classifiers in terms of accuracy based on percentage. The proposed BARF-MLC approaches provide an accuracy value of 90% in Pima Indians that is higher compared to the earlier approaches, whereas other approaches such: PCT renders 65%, for HOMER yields70%and ML-FOREST achieves 80%.

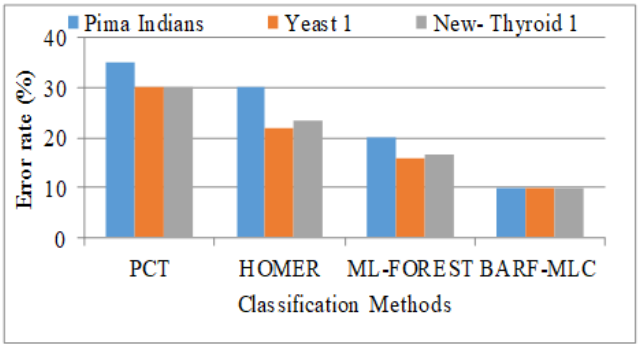

Figure 6 Comparison of various classifiers in error rate

Figure 6 illustrates the evaluation on different classifiers in terms of the error rate based on percentage. The proposed BARF-MLC approaches yields an error rate value of 10% yeast 1 that is lesser compared to the earlier approaches, whereas for other methods: PCT yields 30%, for HOMER yields 22% and ML-FOREST renders 22%.

IV. CONCLUSION AND FUTURE WORK

In this research work, at first, the correlated labels were detected and then optimized with the help of PSO,in addition to this SMOTE and integrated novel label partitions with random forest and the performance of the novel PSOSMOTE algorithm is much better compared to the actual SMOTE and the Bayesian approach is recommended for multi-label classification task. Using the PSOSMOTE algorithm, various sampling rates are utilized for over-sampling diverse minority class samples in imbalanced datasets, and the combination of the optimal sampling rate is acquired. It is experimentally shown that the proposed BARF-MLC is efficient compared to the available techniques including PCT, HOMER and ML-FOREST. In future aimed at the classification enhancement for increasing the accuracy and to reduce error rate.

REFERENCES

1. Soda, P.: A multi-objective optimization approach for class imbalance learning. Pattern Recognit. 44, 1801–1810 (2011)
2. Fang, Ming, Yuqi Xiao, Chongjun Wang, and Junyuan Xie. "Multi-label classification: dealing with imbalance by combining labels." In 2014 IEEE 26th International Conference on Tools with Artificial Intelligence, pp. 233-237. IEEE, 2014.
3. Zhang, Ruilong, Lei Li, Yuhong Zhang, and Chenyang Bu. "Imbalanced networked multi-label classification with active learning." In 2018 IEEE International Conference on Big Knowledge (ICBK), pp. 290-297. IEEE, 2018.
4. H. Wang, Z. Zhang, L. Li et al., A multi-label classification algorithm based on multi-rank neighbor[J]. Chinese Journal of Electronics, vol. 44, no. 10, pp. 2330-2334, 2016.
5. W. Zheng, C. K. Wang, Z. Liu et al., A multi-label classification algorithm based on random-walk model[J]. Chinese Journal of Computers, vol. 33, no. 8, pp. 1418-1426, 2010.
6. Xu, Guibiao, Bao-Gang Hu, and Jose C. Principe. "An asymmetric stagewise least square loss function for imbalanced classification." In 2014 International Joint Conference on Neural Networks (IJCNN), pp. 1107-1114. IEEE, 2014.
7. Yu, Hsiang-Fu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon. "Large-scale multi-label learning with missing labels." In International conference on machine learning, pp. 593-601. 2014.
8. Luo, Fang-Fang, Wen-Zhong Guo, and Guo-Long Chen. "Addressing Imbalance in Weakly Supervised Multi-Label Learning." IEEE Access 7 (2019): 37463-37472.
9. B. Wu, S. Lyu, B.-G. Hu, Q. Ji, "Multi-label learning with missing labels for image annotation and facial action unit recognition", Pattern Recognit., vol. 48, no. 7, pp. 2279-2289, Jul. 2015.
10. Giraldo-Forero, Andrés Felipe, Jorge Alberto Jaramillo-Garzón, José Francisco Ruiz-Muñoz, and César Germán Castellanos-Domínguez. "Managing imbalanced data sets in multi-label problems: a case study with the SMOTE algorithm." In Iberoamerican Congress on Pattern Recognition, pp. 334-342. Springer, Berlin, Heidelberg, 2013.
11. A. Agrawal, H. L. Viktor and E. Paquet, "SCUT: Multi-class imbalanced data classification using SMOTE and cluster-based undersampling," 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), Lisbon, 2015, pp. 226-234.
12. S. Wang and X. Yao, "Multiclass Imbalance Problems: Analysis and Potential Solutions," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 4, pp. 1119-1130, Aug. 2012.
13. Charte, Francisco, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. "MLSMOTE: Approaching imbalanced multilabel learning through synthetic instance generation." Knowledge-Based Systems 89 (2015): 385-397.

AUTHORS PROFILE

Mrs.M.Priyadharshini is an Assistant Professor. She did Bachelor of Science (Physics) in 2007 from Bharathidasan University, Master of Computer Applications in 2010 from Anna University and Master of Philosophy in 2013 from Bharathiar University and has teaching experience of 7 years. She has published in 2 international Journals and presented papers in 7 national conferences. Her area of interest includes Data Mining and Mobile Adhoc Networks.

Dr.L.Pavithira has nearly 15 years of teaching experience in Computer Science. Obtained her B.Sc. degree (Computer Science) in 2001, Master of Computer Applications in 2004 from Bharathiar University. She obtained her master of Philosophy in Computer Science from Bharathiar University during 2006 and Ph.D. in CS from Bharathiyar University in 2015. She has published 15 papers in various international journals and presented 10 papers in various international conferences.