International Journal of Research in Signal Processing, Computing & Communication System Design
5 (1), 2019, 54-58
http://www.publishingindia.com/ijrsp

Ensemble Approaches for Class Imbalance Problem: A Review

Anjana Gosain[1] and Arushi Gupta[2*]

[1]Department of Information Technology, USICT, GGSIP University, Dwarka, Delhi, India.
Email: anjana_gosain@hotmail.com

[2]M.Tech., Department of Information Technology, USICT, GGSIP University, Dwarka, Delhi, India.
Email: arushigupta10818@gmail.com

*Corresponding Author

Abstract: In data mining, performing classification for skewed data distribution is a challenging problem. Traditional Classification Techniques (TCT) work efficiently in classifying data having symmetric distribution, as their internal design favors the balanced datasets. The Class Imbalance Problem (CIP) take place when number of instances of one class outnumbers instances of other classes. Some factors that contribute towards this imbalancing are noisy data, borderline samples, degree of class overlapping, small disjuncts, etc. In machine learning, ensembles are basically built to improve the performance and correctness of single classifier by training multiple classifiers to form the results that output the correct single class label. In this paper, our aim is to review ensemble learning methods having two-class problem. We propose different levels for ensemble learning methods that are at data level, at algorithm level and according to the base classifier.

Keywords: Bagging, Boosting, Classification, Class imbalance problem, Oversampling, Skewed data distribution, Undersampling.

I. Introduction

In data mining, classification is used to assign objects correctly to one of the several predefined categories. Traditional Classification Techniques (TCT) work efficiently in classifying the data having symmetric distribution, as their internal design favors the balanced data sets. However, when used with unbalanced datasets, TCT do not classify datasets correctly as result deviates towards majority class, which is having more data points as compared to minority classes with very few data points. TCT ignore smaller classes considering it as noise. This is called Class Imbalance Problem [1, 2]. In most of the real world situations for example fraudulent telephone calls, credit card frauds, shuttle system failure, text classification, oil spill detection, etc. [10], minority class plays an important role and draw interest of the researchers. TCT do not work well in these cases.

In the literature, researchers have proposed different techniques for handling class imbalance problem which may be classified at three levels: data level approaches (pre-processing techniques), algorithm level approaches and the ensemble based approaches [1]. In data level approaches [1], data sets are balanced using different sampling techniques (undersampling, oversampling and their hybrid forms) before applying TCT in a way results may not get biased towards majority class. While in the algorithm level approaches [3], the internal algorithm structure is modified to improve sensitivity of the algorithm towards the majority class. The ensemble based learning methods include the combination of data level with algorithm level approaches [8].

Ensembles are basically built to improve the performance and correctness of single classifier by training multiple classifiers to form the results that output the correct single class label [6]. Thus, ensemble approach is more useful in predicting the class label than other approaches towards CIP [8].

Different researchers have proposed different techniques for handling CIP using ensemble based approaches. In this paper, we have reviewed ensemble based learning approaches at three levels that are described in Section III of the paper. The organization of the paper is as follows. In Section II, we present factors for skewed data distribution, Section III includes various ensemble learning methods and finally Section IV discusses conclusion and future scope of the paper.

II. Factors for Skewed Data Distribution

In CIP, the class with less number of data points (known as minority class or positive class) is of interest for the study such as the credit card fraudulent cases, shuttle failure, web spam detection, etc. Such real life example data set are skewed

datasets with asymmetrical distributions and is closely related with the following sub problems:

- Degree of Class Overlapping: Class overlapping refers to the level of separability between the classes. Due to overlapping, it becomes difficult for the classifiers to correctly identify the data points of the minority class.
- Size of the Training Dataset: In most of the real-world problems, the number of minority class instances is quite less as compared to majority class instances, usually in the ratio of 1:1000 or 1:10000.
- Small Disjuncts: Small disjuncts covers few training examples. This problem is shown by small clusters which arise due to underrepresented sub concepts [8].
- Noisy Data: Noisy data affects the way any classification algorithm behaves. The presence of noisy data has more impact on the minority class than on the majority classes [1].
- Borderline Examples: Wherever the minority and majority classes are overlapping, the data points are present in places surrounding class boundaries.

III. Ensemble Learning Methods

In this paper, we have classified different ensemble learning approaches at three levels; data level, algorithm level and according to the base classifier as shown in Fig. 1. These three level techniques are further discussed in the subsections.

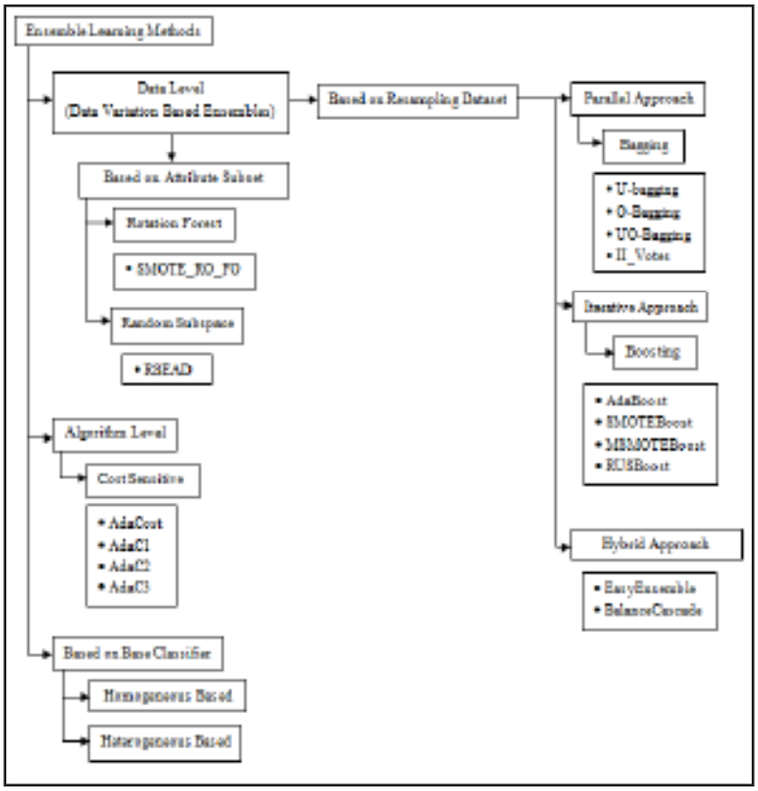

Fig. 1: Ensemble Learning Methods

A. Data Level (Data Variation Based Ensembles)

At the data level, ensembles are built by variation in the data either by resampling data sets or according to the attribute subsets.

i. Based on Resampling Dataset

In resampling dataset based approach, ensembles are built in a parallel way (bagging) or in iterative way (boosting) or hybrid based way (bagging + boosting).

a) Parallel Approach - Bagging

Bagging [15] is the parallel ensemble method in which the base classifiers are produced in parallel. The idea is to combine independent base learners to reduce errors effectively.

- UnderBagging (U-Bagging): This approach uses undersampling which is applied to majority class. However, use of resampling with replacement of positive examples can also obtain more diverse ensembles [8, 14].
- OverBagging (O-Bagging): In this process, number of minority class examples are replicated and all ma-

jority class examples are contained in new bootstrap. SMOTEBagging [16] generate the synthetic data by interpolation method than random replication.

- UnderOverBagging (UO-Bagging): This technique uses both oversampling and undersampling method [16]. A particular resampling rate percentage is selected which states number of samples taken from each class.
- Imbalanced Ivotes (IIVotes): This technique is the hybrid form of SPIDER [9], which is used as a preprocessing method to select data and Ivotes ensemble [17]. This technique modifies SPIDER to improve trade-off between sensitivity and specificity of a classifier at an acceptable level measures.

b) Iterative Approach - Boosting

Boosting belongs to the sequential ensemble paradigm wherein base classifiers are sequentially generated. The idea is to make use of dependence among base classifiers and convert weak learners to strong learners [8].

- AdaBoost [7]: It trains each classifier serially using whole training data sets. After each round of iteration, more focus is given to those instances that are classified not correctly so that it will get classified correctly in the next iteration.
- SMOTEBoost [12]: It combines SMOTE [5], oversampling technique, with AdaBoost, which results in a better hybrid approach to improve classifier performance.
- Modified SMOTEBoost (MSMOTEBoost): This technique is a variant of SMOTEBoost, wherein the MSMOTE pre-processing technique is combined with boosting [18].
- Random Undersampling Boosting (RUSBoost) [6]: It combines random undersampling with boosting approach. The main limitation is the loss of information, which is greatly overcome by combining it with boosting.
- DataBoost-IM [23]: It combines boosting with data-generation to improve the classifier performance for imbalance datasets, but different from SMOTEBoost. Also, it is not able to deal with datasets that are highly imbalanced.
- Boosting Support Vector Machine (BoostingSVM) [24]: The motivation behind this technique is that SVMs are naturally quite robust to the CIP as compared to other classifiers. This technique use SVMs with soft margins as base learner and then use boosting approach.
- Evolutionary Undersampling Boosting (EUSBoost) [25]: It improves the performance of TCT using evolutionary undersampling approach. However, it is computationally more expensive due to the execution of EUS [13] in all iterations of boosting.

c) Hybrid Approach

This approach implements both bagging and boosting approach along with pre-processing techniques. AdaBoost approach is used as base learner such that final classifier generated is an ensemble of ensembles [8].

- EasyEnsemble [19]: The bags are constructed by random undersampling majority class and combining it with the whole minority class. Although, each bag learns more classifier but trains fewer bags than UnderBagging. Also, operations are not performed on original data-set instances. Thus, it allows training all classifiers in parallel.
- BalanceCascade [13]: In this approach, new majority class sample subset is filtered by removing correct instances and all misclassified instances are kept in the majority class. Iteratively, most classifiers are produced on the filtered sampled data-set and finally combined together. By working in a supervised way, Balance Cascade trains all classifiers sequentially.

ii. Based on Attribute Subset

In the attribute subsets approach, ensemble approaches exploit datasets that are having highly redundant features e.g. random subspace and rotation forest approaches.

a) Rotation Forest

Rotation Forest (RO_FO) uses independently trained decision trees to build classifier ensembles. In the RO_FO, subsets are formed from original feature set and on each subset, PCA technique is implemented independently. Then, data is linearly transformed into new features. Training of a decision tree is done with this set. Different rotations results due to the different splits of feature sets. In this way, diverse classifier is developed [21].

- SMOTE with Rotation Forest (SMOTE_RO_FO) [11]: It is a new hybrid ensemble approach by combining SMOTE with RO_FO to address CIP. It trains classifier by rotating subspaces of the original data-set and works in noiseless environment.

b) Random Subspace

In this method, classifiers are built in random subspaces of feature space of dataset. These classifiers uses simple majority voting in the final decision rule [25]. It is useful for gene expression problem in which the number of training examples is less than number of features [3].

- Random Subspace Ensemble with Artificial Datasets (RSEAD) [20]: It is an ensemble based active learning approach, which created artificial data as per the data-set distribution to increase the diversity of ensemble.

B. Algorithm Level

At the algorithm level, the ensembles are built by cost sensitive approaches [1].

i. Cost Sensitive Approaches

a) AdaCost

AdaCost algorithm adjusts the parameter of weight update by adding a cost adjustment function ϕ [8, 22]. This function increases the weight for those instances that are incorrectly classified but decreases its weight less otherwise. The cost of misclassifying i^{th} example can be represented by the given function.

$$\phi+ = -0.5Ci + 0.5 \text{ and } \phi- = 0.5Ci + 0.5 \quad (1)$$

b) AdaC1

It is one of the three modifications of AdaBoost [4] which is based on different methods to embed cost into the weight update formula. The different values of αt computed depends on where the cost is introduced. In AdaC1, the cost factors are included within exponent part of the formula.

$$D_{t+1}(i) = D_t(i) \cdot e^{-\alpha t Ci yi ht(xi)}, \text{ where } Ci \in [0, +\infty) \quad (2)$$

c) AdaC2

Similar to AdaC1, AdaC2 introduce cost factor in weight update formula. But in a different way; the costs are included outside exponent part [4].

$$Dt + 1(i) = CiDt(i) \cdot e^{-\alpha t yi ht(xi)} \quad (3)$$

d) AdaC3

AdaC3 modifies both AdaC1 and AdaC2. The weight update formula is changed by including costs both inside and outside the exponent part [4].

$$Dt+1(i) = CiDt(i) \cdot e^{-\alpha t Ci yi ht(xi)} \quad (4)$$

e) MetaCost

It applies meta learning concept to reduce cost of classifier is reduced. It is a variant of bagging and works by first creating multiple bootstrap instances from training set and then assigns class based on majority of votes received from ensemble [26].

C. Based on Base Classifier

According to the base classifier, the ensembles could have either homogeneous classifiers that include minor variants of the same classifier or heterogeneous classifiers that comes under broader category of multiple classifier system. Thus, homogeneous classifier works on same algorithm over varied datasets whereas heterogeneous classifier works on distinct learning algorithms over same data.

i. Homogeneous Based

Ensemble methods mostly refer to collection of classifiers that are minor variants of same classifier [8]. When forming ensembles, creating diverse classifiers is a key element. This approach constructs ensembles of classifiers that include classifiers of the same base learner.

ii. Heterogeneous Based

The multiple classifier systems include combinations that consider hybridization of different models [8]. While constructing the ensembles of classifier, the base learners are different and thus represented as the heterogeneous based approach. These provide their specific algorithms for the prediction of class label and thus their combination will result into better and improved performance of classification task in terms of accuracy and sensitivity of individual classifiers.

IV. Conclusion and Future Scope

The paper discusses about limitations of existing traditional classifiers and factors contributing for skewed datasets. It is observed that ensemble approach is more useful in predicting class label than other approaches toward CIP. Thus, ensemble improves the performance and correctness of single classifier by training multiple classifiers.

References

[1] A. Gosain, and S. Sardana, "Handling class imbalance problem using oversampling techniques: A review," *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI),* pp. 79-85, 2017.

[2] J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 3rd ed., Morgan Kaufmann Publishers, 2011.

[3] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, Pearson Education Pvt. Ltd., 2013.

[4] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358-3378, 2017.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

[6] C. Seiffert, T. M. Khoshgoftaar, J. V. Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Transactions on Systems, Man,*

and Cybernetics - Part A: Systems & Humans, vol. 40, no. 1, pp. 185-197, 2010.

[7] Y. Freund, and R. E. Schapire, "Experiments with a new boosting algorithms," in *Proceedings of the 13th International Conference on Machine Learning (ICML'96),* pp. 148-156, 1996.

[8] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews*, vol. 42, no. 4, pp. 463-484, 2012.

[9] J. Stefanowski, and S. Wilk, "Selective pre-processing of imbalanced data for improving classification performance," in *International Conference on Data Warehousing and Knowledge Discovery*, Springer, Berlin, Heidelberg, pp. 283-292, 2008.

[10] A. Gosain, A. Saha, and D. Singh, "Analysis of sampling based classification techniques to overcome class imbalancing," in *3rd International Conference on Computing for Sustainable Global Development (INDIACom),* pp. 2637-2643, 2016.

[11] S. Fattahi, Z. Othman, and Z. A. Othman, "New approach with ensemble method to address class imbalance problem," *Journal of Theoretical & Applied Information Technology*, vol. 72, no. 1, pp. 23-33, 2015.

[12] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. Bowyer, "SMOTEBoost: Improving prediction of the minority class in boosting," in *7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, Cavtat Dubrovnik, Croatia, pp. 107-119, 2003.

[13] X. Y. Liu, J. Wu, and Z. H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, vol. 39, no. 2, pp. 539-550, 2009.

[14] R. Barandela, R. M. Valdovinos, and J. S. Sánchez, "New applications of ensembles of classifiers," *Pattern Analysis & Applications*, vol. 6, no. 3, pp. 245-256, 2003.

[15] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.

[16] S. Wang, and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models," *IEEE Symposium on Computational Intelligence and Data Mining*, pp. 324-331, 2009.

[17] L. Breiman, "Pasting small votes for classification in large databases and on-line," *Machine Learning*, vol. 36, no. 1-2, pp. 85-103, 1999.

[18] S. Hu, Y. Liang, L. Ma, and Y. He, "MSMOTE: Improving classification performance when training data is imbalanced," *Second International Workshop on Computer Science and Engineering*, vol. 2, pp. 13-17, 2009.

[19] K. M. Ting, "An instance-weighting method to induce cost sensitive trees," *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 3, pp. 659-665, 2002.

[20] Y. Yang, and G. Ma, "Ensemble based active learning for class imbalance problem," *Journal of Biomedical Science and Engineering*, vol. 3, no. 10, pp. 1021-1028, 2010.

[21] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1619-1630, 2006.

[22] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan, "AdaCost: Misclassification cost-sensitive boosting," in *Proceedings of the Sixteenth International Conference on Machine Learning (ICML'99)*, pp. 97-105, 1999.

[23] H. Guo, and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: The databoost-im approach," *ACM Sigkdd Explorations Newsletter,* vol. 6, no. 1, pp. 30-39, 2004.

[24] B. X. Wang, and N. Japkowicz, "Boosting support vector machines for imbalanced data-sets," *Knowledge Information Systems*, vol. 25, no. 1, pp. 1-20, 2010.

[25] M. Galar, A. Fernández, E. Barrenechea, and F. Herrera, "EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling," *Pattern Recognition*, vol. 46, no. 12, pp. 3460-3471, 2013.

[26] P. Domingos, "Metacost: A general method for making classifiers cost-sensitive," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 155-164, 1999.