iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 Issue: II Month of publication: February 2018

DOI:

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue II, February 2018- Available at www.ijraset.com

A Review on Various Techniques to Resolve Multiclass Imbalance Problem

Ms. K.S. Baviskar[1], Prof. J. R. Mankar[2]
[1]M. E. Student, [2]Assistant Professor, Department of computer Engineering, K. K. Wagh Institute of Engineering Education & Research, Nashik Savitribai Phule Pune University, Maharashtra, India.

Abstract— In classification algorithm data is distributed in two or more classes. Large difference in number of instances causes skewness in data. This skewness in data also referred as data imbalance which causes difficulties in processing dataset and affects the accuracy of data classification. The data containing different number of instances with respect to class labels is called as imbalanced dataset. The major issue is the classification of minority class samples. There are various techniques proposed to deal with imbalance class problem without affecting the classification accuracy of majority class. This work aims to study various imbalance data handling techniques and its classification accuracy.

Keywords*— Data distribution, data imbalance, Mahalanobis distance, over-sampling techniques, SMOTE.*

I. INTRODUCTION

Data containing different number of instances with respect to class labels is called as imbalanced dataset. Machine learning algorithm faces many problems due to the skewness in data distribution. These skewness in data distribution causes major issue of classification of minority class samples. It affects the classification accuracy. There is need of such technique that will provide high classification accuracy for minority class without affecting the classification accuracy of majority class. There are various domain where imbalanced data is get generated such as protein fold, weld flaw. Two class problem have only two classes in dataset. One is majority class and other one is minority. The class have overwhelmed called the majority class while the other called minority class. Cost sensitive learning is algorithmic solution for class imbalance problem. Zhi-Hua Zhou and Xu-Ying Liu stated that multiclass imbalance problem is more difficult than two class imbalance problem. The difficulty also increases if higher degree of class imbalance occurs. They also stated that almost all techniques resolves the two class imbalance problem but most are ineffective and some gives negative results in case of multiclass imbalance problem.

Solution for multiclass imbalance problem is mainly classified in two categories:

A. Data Level

This solution deals with data skewness distribution. It applies oversampling or undersampling technique. The number of instances of minority class increases to certain level to balance the data distribution. This process is called as oversampling. By analyzing existing data, synthetic data is generated and added to the existing dataset. This solution may lead to overfitting or over generalization problem. In under sampling technique, instances of majority class are removed to balance the data distribution. This solution may lead to information loss and may mislead to classification technique.

B. Algorithmic Level

Ensemble learning and one class learning are two techniques using which multiclass imbalance problem is handled. In machine learning, multiple learners are trained to solve the same problem as supervised learning mechanism. Ensemble learning have two techniques such as boosting and bagging. One class learning modifies the training mechanism to achieve better accuracy in minority class classification problem. Rather than considering majority and minority classes it directly applies one class learning mechanism. Oversampling technique is most widely used technique among all the proposed techniques. Artificial synthetic data for minority class is generated to balance the data distribution. Paper is organized as follows: section I introduces solutions for multi-class imbalance problem. Section II gives the literature review. Section III concludes the paper.

II. LITERATURE WORK

This section includes various existing techniques proposed to solve class imbalance problem using data level technique. Related studies using Under-sampling and Oversampling for multi-class imbalanced problems are reviewed.

A. Under-Sampling

Under-sampling technique is proposed by Weiss and Provost [5]. This uses random under-sampling technique to remove random instances of majority class. This leads to information loss as it may remove the useful informative instances. Most widely used techniques are CNN [1], ENN [2], OSS [3]. Condensed nearest neighbor decision rule (CNN rule) [1] finds the consistent subset of majority class instances. This is iterative process it initially selects the instances those are misclassified by classifier and after words removes those instances. Wilson's edited nearest neighbor rule (ENN) [2] is another under-sampling technique. This technique uses 3 nearest neighbor rule and single nearest neighbor rule. 3NN is used to edit pre-classified samples and 1NN is used to make decisions. One sided selection (OSS) is proposed by Kubat and Matwin [3]. This technique only deals with two class imbalance problem. This technique removes all negative instances and positive instances are preserved. Negative samples are identified by finding redundancy in majority class.

B. Oversampling

The effect of class imbalance in classification technique [18] and the effect of oversampling in classification is compared. Oversampling technique leads to better accuracy irrespective of the selection of classifier. This technique proposes the binarization techniques, instance weighting, cost-sensitive learning and an SMT based technique to generate samples.

Over-generalization problem is proposed by Prati, Batista, and Monard [8]. To increase the samples of minority class, Existing instance copies are created. This can be one of the solutions. But this causes over fitting and over generalization in minority class instances. SMOTE technique [6] uses k-nearest neighbor technique to generate minority class samples. This technique calculates the similarity between instances and its feature space. It randomly selects one of the k-nearest neighbors of instance of minority class and calculates the difference between these instances with selected sample. This difference is then multiplied by one random value generated in between the range of 0 to 1 and then added those instances to the dataset. SMOTE faces overgeneralization problem. It also leads to overlapping multiclass instance problem. This technique calculates the minority class instances without considering the boundaries of minority class. This causes change in minority class boundaries. This leads to misclassification of majority class samples. To overcome all these problem of SMOTE technique, solutions are proposed in literature [11], [12]. ADASYN [11], Adaptive Synthetic Sampling algorithm uses an adaptive learning procedure and dynamic adjustment of weights according to data distributions. Safe-level SMOTE strategy [12] is proposed by Bunkhumpornpat, Sinapiromsaran and Lursinsap. This algorithm uses different weight degrees called as safe level. Before generating instances it generates the safe region. These two algorithms produces better results, but the solution is applicable on for two class imbalance problems. MSYN algorithm [15] uses large margin principle along with SMOTE. It initially generates the samples using SMOTE and then applies filtering technique. Again this solution works for two class imbalance problem. After analyzing the problem of multiclass classification over imbalanced data, OAO [4] and OAA [7] schemes are proposed to decompose multiclass problem to two class problem. In OAO solution is generated by comparing each class with every other class in dataset and in OAA every single class is compared with other set of classes. The selected class is treated as positive class and all other classes are treated as negative class. OAA gives better performance than SMOTE technique.

Dynamic oversampling technique [16] and [19] is proposed for multiclass data imbalance problem. It uses radial basis functions neural networks- RBFNNs and memetic algorithm-MA [16]. Dys algorithm is proposed by Lin, Tang, and Yao [19]. It uses multilayer perceptrons (MLP). These techniques performs oversampling in training phase. These are heuristics algorithms. Due to this training time increases in classification problem and final model complexity also increases. SERA [13] deals with non-stationary Imbalanced data stream mining problems. This is two class technique. It generates minority class data chunks the instances are added to chunks by measuring the similarity with the chunk. It uses Mahalanobis distance to measure similarity. Ensemble learning algorithm - MuSeRA technique [14] is similar to SERA technique. Like SERA, MuSeRA also generates the data chunks. In this technique, chunks are dynamically updated and used for next iteration to generate new instance. Combined technique of oversampling and under sampling is proposed by Batista, Prati, and Monard [9]. It proposes 3 combinations: CNN+Tomek, SMOTE+Tomek and SMOTE+ENN. This technique deals with the overlapping of instances in multiple classes. But this technique leads to increase the false positive rate of the learning algorithms. But these techniques face the same problems faced by SMOTE technique. TRIM [17] technique uses feature extraction approach. This technique generates precise minority class region. It iteratively removes the majority class instances and adds the minority class instances using SMOTE. This technique generates better results than existing technique but it is applicable for two class imbalance problem. Abdi and Hashemi [20] proposed Mahalanobis distance based oversampling technique – MDO. By considering each instance in minority class it generates new synthetic dataset. The generated dataset instances has equal Mahalanobis distance from the derived class mean. The class mean is calculated using existing candidate instances. This technique avoids the overlapping between multiple class instances. This technique preserves the covariance structure of data present in minority class. MDO generates synthetic instances those are lies in dense region of existing minority class instances. MDO generates better results as compared to other techniques.

The previous work focuses on oversampling and Under-sampling technique. CNN, ENN, OSS are most widely used Under-sampling techniques which removes the instances of majority class to balance the data distribution. OAO and OAA schemes proposed to decompose multiclass problem to two class problem. SMOTE, ADASYN, MSYN and MDO are Over-sampling techniques which generates synthetic samples of minority class at certain level.

III.CONCLUSION

In this survey, various solutions for imbalance data handling is studied. The most suitable and widely used technique is oversampling technique. Mahalanobis distance technique provides better solution for multiclass imbalance problem. This technique preserves the covariance structure of data by calculating class mean. As number of attributes increases the time required for evaluating the single attribute value for every instance in synthetic dataset also increases. Hence the execution time increases as number of attributes in dataset increases. Feature selection technique can be applied as a preprocessing task to improve the execution performance of system without compromising accuracy. The second important aspect is, in all existing work data imbalance solution works for numeric data. There is need to extend these techniques for categorical or Boolean data values.

ACKNOWLEDGMENT

Authors would like to thanks Prof. Dr. K. N. Nandurkar, Principal and Prof. Dr. S. S. Sane, Head of Department of Computer Engineering, K.K.W.I.E.E.R., Nashik for their kind support and suggestions. We would also like to extend our sincere thanks to all the faculty members of the department of computer engineering and colleagues for their help.

REFERENCES

[1] Peter E. Hart. The condensed nearest neighbor rule (corresp.). IEEE Transactions on Information Theory, 14(3):515–516, 1968.

[2] Dennis L Wilson. Asymptotic properties of nearest neighbor rules using edited data. Systems, Man and Cybernetics, IEEE Transactions on, (3):408–421

[3] Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. In ICML, volume 97, pages 179–186. Nashville, USA, 1997.

[4] Trevor Hastie, Robert Tibshirani, et al. Classification by pairwise coupling. The annals of statistics, 26(2):451–471, 1998.

[5] Gary M. Weiss and Foster Provost. The effect of class distribution on classifier learning: an empirical study. Rutgers Univ, 2001.

[6] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16:341–378, 2002.

[7] Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. The Journal of Machine Learning Research, 5:101–141, 2004.

[8] Ronaldo C. Prati, Gustavo EAPA Batista, and Maria Carolina Monard. Class imbalances versus class overlapping: an analysis of a learning system behavior. In MICAI 2004: Advances in Artificial Intelligence, pages 312–321. Springer, 2004.

[9] Gustavo EAPA Batista, Ronaldo C Prati, and WeMaria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. ACM Sigkdd Explorations Newsletter, 6(1):20–29, 2004.

[10] Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. Knowledge and Data Engineering, IEEE Transactions on, 18(1):63–77, 2006.

[11] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on, pages 1322–1328. IEEE, 2008.

[12] Chumphol Bunkhumpornpat, Krung Sinapiromsaran, and Chidchanok Lursinsap. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In Advances in Knowledge Discovery and Data Mining, pages 475–482. Springer, 2009.

[13] Sheng Chen and Haibo He. Sera: selectively recursive approach towards nonstationary imbalanced stream data mining. In Neural Networks, 2009. IJCNN 2009. International Joint Conference on, pages 522–529. IEEE, 2009.

[14] Sheng Chen, Haibo He, Kang Li, and Sachi Desai. Musera: Multiple selectively recursive approach towards imbalanced stream data mining. In IJCNN, pages 1–8, 2010.

[15] Xiannian Fan, Ke Tang, and Thomas Weise. Margin-based oversampling method for learning from imbalanced datasets. In Advances in Knowledge Discovery and Data Mining, pages 309–320. Springer, 2011.

[16] Francisco Fern´andez-Navarro, C´esar Herv´as-Mart´ınez, and Pedro Antonio Guti´errez. A dynamic over-sampling procedure based on sensitivity for multi-class problems. Pattern Recognition, 44(8):1821–1833, 2011.

[17] Kamthorn Puntumapon and Kitsana Waiyamai. A pruning-based approach for searching precise and generalized region for synthetic minority over-sampling. In Advances in Knowledge Discovery and Data Mining, pages 371–382. Springer, 2012.

[18] Alberto Fern´andez, Victoria L´opez, Mikel Galar, Mar´ıA Jos´e Del Jesus, and Francisco Herrera. Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. Knowledge-based systems, 42:97–110, 2013.

[19] Minlong Lin, Ke Tang, and Xin Yao. Dynamic sampling approach to training neural networks for multiclass imbalance classification. Neural Networks and Learning Systems, IEEE Transactions on, 24(4):647–660, 2013.

[20] Lida Abdi and Sattar Hashemi, "To Combat Multi-class Imbalanced Problems by Means of Over-sampling Techniques," IEEE Transactions on Knowledge and Data Engineering Vol. 28, pp. 238 - 251, Issue. 1, Jan. 2016