A BRIEF SURVEY ON CLASSIFICATION METHODS FOR UNBALANCED DATASETS

Manoj Kumar Sahu[1], Rajeev Pandey[2], Sanjay Silakari[3]

[1]*M. Tech. Research Scholar, Department of Computer Science, UIT RGPV Bhopal, M.P., India*
[2]*Assistant Professor, Department of Computer Science, UIT RGPV Bhopal, M.P., India*
[3]*HOD, Department of Computer Science, UIT RGPV Bhopal, M.P., India*

Abstract

In real world, we deal with the data sets which are unbalanced in nature. Information sets are lopsided when no less than one class is spoken to by extensive number of preparing illustration (called greater part class) while different classes make up the minority. Due to this uneven nature of information sets we have great precision on dominant part class yet on the other side exceptionally poor exactness on the minority class, while we attempt to foresee the class enrollment. Accordingly, the lopsided way of information sets can have a negative impact on arrangement execution of machine learning calculations. Specialists have been made numerous endeavors to manage such issues of order of information at information level and additionally calculation level. In this paper we speak to a brief study of existing answers for the class-unevenness issue proposed both at the information and algorithmic levels.

Keywords: *Classification, Cost Sensitive Learning, Unbalanced Data, Over Sampling, Under Sampling.*

----------***----------

1. INTRODUCTION

Arrangement is a composed method for foreseeing class participation for an arrangement of illustrations or occasions utilizing the qualities of those cases. The vast majority of this present reality order issues include expansive quantities of learning cases and convoluted connections between class participation and case properties.

In numerous certifiable space, it is normal for information sets to have uneven class disseminations. This circumstance happens when no less than one class is spoken to by little number of illustration (called minority class) while different classes makes up the rest (called majority class). In this connection, classifiers can give great exactness on the dominant part class yet low precision on the minority class(es). This happens on the grounds that conventional preparing criteria, for example, the general achievement or mistake rate can be enormously affected by the bigger number of case from the larger part class. Most unique grouping calculations seek after to understate the blunder rate: the rate of the off base expectation of class names. They overlook the contrast between sorts of misclassification blunders. Specifically, they verifiably accept that all misclassification mistakes cost similarly.

In some genuine applications, this supposition is not valid. The contrasts between various misclassification blunders can be entirely huge. For instance, in restorative analysis of a specific malignancy, if the tumor is viewed as the positive class, and non-growth (sound) as negative, then missing a disease (the patient is really positive yet is named negative; therefore it is additionally called ―false negative) is a great deal more genuine (in this manner costly) than the false-positive blunder. The patient could lose his/her life on account of the postponement in the right analysis and treatment. So also, if conveying a bomb is certain, then it is a great deal more costly to miss a terrorist who conveys a bomb to a flight than seeking a blameless individual.

Imbalanced information sets exists in some true areas, for example, spotting phone misrepresentation [12] , Mastercard extortion [13], recognition of oil slicks in satellite radar pictures [5], learning word articulations [6], content grouping [7], data recovery and sifting assignments [9], deficiency finding, for example, system investigating [15], medicinal determination of uncommon conditions [14], [16], [17], object location, for example, target [18], face [19], or pedestrian[20] discovery (in expansive pictures) etc.

Numerous exploration papers on imbalanced information sets have ordinarily concurred that as a result of this unequal class dispersion, the execution of the current classifiers has a tendency to be one-sided towards the dominant part class.

The purposes behind poor execution of the current grouping calculations on imbalanced information sets are:
1. They are precision driven i.e., they will likely minimize the general blunder to which the minority class contributes practically nothing.
2. They accept that there is equivalent dissemination of information for all the classes.
3. They additionally expect that the blunders originating from various classes have the same cost[21].
4. And most important reason for poor performance is biased nature of data, because of that the class having more values of different types will help in the learning phase of machines and because of that these machines will have more knowledge to map with upcoming data in the practical field. whereas, the class having less data will also affect the

learning of machines and hence they suffer from poor mapping of data in the practical field. Therefore overall performance decreases.

With lopsided information sets, information mining learning calculations produce worsened models that don't consider the minority class as most information mining calculations accept adjusted information set.

As the minority class frequently speaks to the primary class of enthusiasm for some certifiable issues, precisely grouping case from this class can be at any rate as essential as, and in a few situations more critical than, precisely ordering case from the larger part class.

Tending to this learning inclination to discover arrangements with great exactness on both the minority and larger part classes has turned into a vital region of exploration [22]. Methods to address this issue include two primary angles. The primary includes changing, or examining from, the first uneven information set by making a falsely adjusted dispersion of class case for preparing, the alleged "outer" methodologies as the preparation information are balanced, not the learning calculation. Basic outer methods incorporate oversampling the minority class to help representation [23], under-testing or altering of the dominant part class to abatement representation [24], or stowing and boosting where numerous adjusted subsets of class illustrations are utilized as a part of preparing [25]. While these methodologies can be successful, they have certain impediments. Testing procedures can add a computational overhead to the preparation procedure, lead to over-fitting as conceivably valuable taking in illustrations can be prohibited from the learning procedure, and require from the earlier undertaking particular information about the information to plan a reasonable inspecting calculation.

The second method utilizes cost change inside the learning calculation to figure the uneven appropriation of class case in the first (unmodified) lopsided information set, amid the preparation procedure (the alleged "inward" methodologies) [26],[27]. In GP, cost modification can be implemented by adjusting the wellness capacity. Here, arrangements with great order precision on both classes are remunerated with better wellness, while those that are one-sided toward one class just are punished with poor wellness. Regular systems incorporate utilizing settled misclassification costs for minority and greater part class illustrations [14], [15], or enhanced execution criteria, for example, the range under the recipient working trademark (ROC) bend (AUC) [26], in the wellness capacity. While these procedures have considerably enhanced minority class exhibitions in advanced classifiers, they can acquire both a tradeoff in dominant part class precision and, in this way, a misfortune in general arrangement capacity, and long preparing times because of the computational overhead in assessing these enhanced wellness measures. Moreover, these methodologies can be issue particular, i.e., wellness capacities are carefully assembled for a specific issue area just [28].

2. APPROACHES AT DATA LEVEL

The most widely recognized and surely understood way to deal with unbalanced data is sampling. Hare sampling is used to determine mathematically the most effective way to fetch a sample which represent or reflect the whole class under study or observation.

Sampling can be applied by two ways for balancing the unbalanced classes of data, either by oversampling the minority class or by under sampling the class which is having more information or data or examples until the classes are around similarly spoken to or if nothing else they have an accomplished a fancied or expected level of harmony between them.

Both techniques are pertinent in any learning framework, since they go about as a preprocessing stage, these methodologies offer the learning framework to get the preparation examples as though they had a place with an all around adjusted information set. Along these lines, any predisposition of the framework towards the majority class because of the diverse extent of illustrations per class would be relied upon to be smothered.

Hulse et al. [29] recommend that the utility of the re-sampling strategies relies upon various components, including the proportion amongst positive and negative cases, different qualities of information, and the way of the classifier. Be that as it may, re-sampling strategies have indicated essential downsides. Under-esampling may toss out possibly helpful information, while over-sampling falsely expands the measure of the information set and subsequently, compounds the computational weight of the learning calculation.

2.1 Over Sampling

The simplest approach to build the measure of the minority class compares to irregular over-sampling, that is, a non-heuristic technique that adjusts the class circulation through the arbitrary replication of positive illustrations. All things considered, since this technique recreates existing case in the minority class, over-fitting will probably happen.

Andrew Estabrooks et al. proposed a numerous resampling technique which chose the most suitable re-sampling rate adaptively [30].

Taeho Jo et al. advanced a bunch based over-examining strategy which managed between-class lopsidedness and inside class irregularity at the same time [31].

Hongyu Guo et al. discovered hard instances of the dominant part and minority classes amid the procedure of boosting, then produced new engineered case from hard illustrations and add them to the data sets [32].

2.2 Under Sampling

Under examining is a proficient system in unevenness learning. This technique takes those subset of the majority

class to prepare the classifier which speaks to or mirror the greater part of the attributes of the class. As an aftereffect of this pruning numerous preparation illustrations are pruned and the preparation set turns out to be more adjusted and the preparation procedure turns out to be quicker on account of the little measure of preparing cases. Random majority under sampling (RUS) is the most well-known system utilized for under examining, in this procedure , Instances of the lion's share class are haphazardly pruned from the dataset.

However, the under sampling strategy suffers from a major drawback, The under sampling strategy can discard some instances of the data set, which may contain potentially useful information.

There has been numerous routes endeavored to enhance the execution of random sampling, for example, Tomek joins, Condensed Nearest Neighbor Rule and One-sided determination and so on. one sided selection (OSS) is proposed by Rule Kubat and Matwin endeavors to keenly under-sample the majority class by removing larger part class outlines that are considered either dreary or uproarious.

For a few issues like misrepresentation recognition which is profoundly covered unequal information order issue, where non-extortion tests vigorously dwarf the extortion samples,T. Maruthi Padmaja[10] proposed hybrid sampling method, a mix of SMOTE to over-specimen the minority information (extortion sample) and arbitrary under-inspecting to under-example the dominant part information (non-misrepresentation sample) on the off chance that we kill amazing exceptions from the minority samples for exceedingly skewed imbalanced information sets like extortion identification order precision can be moved forward.

There are surely understood inconveniences connected with the utilization of sampling which energize the utilization of cost-sensitive learning. The weakness connected with under sampling is that, it can kill conceivably valuable information from the data sets. The principle disservice with over-sampling, from our point of view, is that by making indistinguishable duplicates of previous illustrations, it makes over-fitting likely. Truth be told, with oversampling it is destined to happen for a learner to create an order principle to cover a solitary, recreated, case. Another hindrance of oversampling is that it builds the span of data sets, which results in long learning time. What's more, the most breathtaking downside of oversampling is that, how to choose the criteria for regeneration of training examples.

Sapling techniques have serious drawbacks, though it is widely popular and practical approach to deal with imbalanced data.

3. APPROACHES AT ALGORITHM LEVEL

At this approach we make changes at the algorithm level, by adjusting the value of various factors which can improve the result of the learning system. And this strategy has nothing to deal with the data. This approach is also known as internal approach or cost touchy/sensitive learning. cost touchy Learning is a kind of learning in information mining that takes the misclassification costs (and conceivably diverse sorts of cost) into thought. There are various approaches to deal with cost delicate learning, in [13], it is ordered into three classes, the top notch of strategies apply misclassification expenses to the information set as a type of information space weighting, the menial applies cost-minimizing procedures to the blend plans of outfit techniques, and the last class of systems joins cost delicate elements straightforwardly into characterization ideal models to basically fit the cost delicate structure into these classifiers.

Nguyen ha vo, Yonggwan won et al. [1] improved Regularized Least Square(RLS) calculation that punishes mistakes of various samples with appropriate weights and a few guidelines which are utilized to decide those weights. The above system enhances the order precision of altered RLS classifier and stands as a substitution of prior cost touchy characterization techniques. The above methodology is like inspecting strategy and it relies on upon the cost we pick. For instance, modifying of the cost-affectability of one class is corresponding to addition or decrement to the quantity of illustrations or test in that class.

Jie song et al.[2] gave an variant of AdaBoost which is named as BABoost. This is a new approach which reduces class within group error. Adaboost algorithm awards equal weight to every misclassified example, and they ignore the fact that misclassification error of each class may be different and generally it falls out. Most of the time, misclassification error of majority class will be smaller than the minority class. Therefore, AdaBoost algorithm will head towards smaller margin and higher bias when run across skew distribution. On the other side BABoost algorithm consider the above problem. BABoost calculation in every cycle of boosting allots more weights to the misclassified illustrations, particularly those in the minority class.

Samya Elaoud et al. [10], they proposed a Pareto fitness genetic algorithm (PFGA), their strategy proposed a modified ranking procedure and a bright way of sharing; a new objective function is designed according to the density value and rank of individual. The performance of PFGA is judged on six multi-objective standards with dissimilar pareto front features.

4. HYBRID APPROACHES

Hybrid algorithms, as the name specifies, a method which has features of two or more than two algorithms or approaches. When we talk about the problem related to classification of imbalance data set then the hybrid algorithm may contain a combination of both internal and external approach, which means an approach combining the best of internal and external approach. In a general way hybrid approach can choose either under-sampling or over sampling method for selection of appropriate candidates and later on apply some other approach (by tuning fitness

function) to improve the accuracy of classification or they can apply any possible combination of internal/ external approaches. In this paper we discuss some of the available approaches.

Sung-Hwan min et al. [35] proposed an approach for bankruptcy prediction. They proposed a hybrid approach with integration of GA (genetic algorithm) and SVM (support vector machine). They improved the performance of SVM of with the help of GA in aspects namely (a) parameter optimization (b) feature subset selection. They have utilized GA to support both parameter enhancement and highlight subset choice of SVM at the same time. They have contrasted this procedure and previous methodologies and discovered it to a great degree helpful for bank insolvency expectation issue.

James D. Kelly et al. [34] proposed a decent mix of K nearest neighbors algorithm and genetic algorithm for arrangement of information sets or data sets. They used the GA to learn the real valued weights associated with soul attributes in the data sets. Then they apply the KNN algorithm to classify results of GA on the basis of their weighted distance from the members of the training sets. They have applied this approach on three different test cases and found that the results obtained by hybrid approach are better than the normal KNN.

Yi-Tung Kao et al. [35] they combined two heuristic method namely particle swarm optimization (PSO) and genetic algorithm (GA) to get a new method which works in multimodal functions as global optimizer and they named it GA-PSO. This new technique takes best of both GA (evolution) and PSO (Self improvement), and creates new data set with the help of GA and PSO. In this approach, members of data set explore themselves on the basis of their private cognition and social interaction. When GA-PSO gets compared with other methods to find global optimum, GA-PSO shows the edge over them.

5. CONCLUSION

In this paper, an outline of the classification techniques for unbalanced data set has been shown. At data level sampling method is common which has its own advantages and disadvantages because when we go for oversampling additional processing cost increases and selection of criteria to boost up the data set is itself a big task. On the other hand under sampling may loose the important information. At the algorithm level various solutions are there but most of them are problem specific and are biased in nature means they provides better result for majority class while poor result for minority class. But, this problem can be addressed upto some level by adjusting the weight or cost of classifier which is again a big task. Then we have hybrid algorithms which use best of all and provide better results where the previous two fails. There are various possibilities for developing new classifiers which will work efficiently on imbalanced data set and hybrid algorithms have a vast scope for research in unbalanced data set.

REFERENCES

[1]. Vo, N. H. and Y. Won (2007). Classification of unbalanced medical data with weighted regularized least squares. Frontiers in the Convergence of Bioscience and Information Technologies, 2007. FBIT 2007, IEEE.

[2]. Song, J., et al. (2009). An improved AdaBoost algorithm for unbalanced classification data. Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on, IEEE.

[3]. Sun, Y., et al. (2007). "Cost-sensitive boosting for classification of imbalanced data." Pattern Recognition **40**(12): 3358-3378.

[4]. Ezawa, K. J., et al. (1996). Learning goal oriented Bayesian networks for telecommunications risk management. ICML.

[5]. Kubat, M., et al. (1998). "Machine learning for the detection of oil spills in satellite radar images." Machine learning **30**(2-3): 195-215.

[6]. Van Den Bosch, A., et al. (1997). When small disjuncts abound, try lazy learning: A case study. Proceedings of the Seventh Belgian-Dutch Conference on Machine Learning, Citeseer.

[7]. Zheng, Z., et al. (2004). "Feature selection for text categorization on imbalanced data." ACM Sigkdd Explorations Newsletter **6**(1): 80-89.

[8]. Fawcett, T. and F. Provost (1997). "Combining data mining and machine learning for effective fraud detection." Proc. AI Approaches to Fraud detection and Risk Management: 14-19.

[9]. Lewis, D. D. and J. Catlett (1994). Heterogeneous uncertainty sampling for supervised learning. Proceedings of the eleventh international conference on machine learning.

[10]. Elaoud, S., et al. (2007). "The Pareto fitness genetic algorithm: Test function study." European Journal of Operational Research **177**(3): 1703-1719.

[11]. Fawcett, T. and F. Provost (1997). "Adaptive fraud detection." Data mining and knowledge discovery **1**(3): 291-316.

[12]. Stolfo, S., et al. (1997). Credit card fraud detection using meta-learning: Issues and initial results. AAAI-97 Workshop on Fraud Detection and Risk Management.

[13]. Holmes, J. H. (1998). "Differential negative reinforcement improves classifier system learning rate in two-class problems with unequal base rates." Genetic Programming: 635-644.

[14]. Pazzani, M., et al. (1994). Reducing misclassification costs. Proceedings of the Eleventh International Conference on Machine Learning.

[15]. Gray, H., et al. (1996). "Genetic programming for classification of brain tumours from nuclear magnetic resonance biopsy spectra." Genetic Programming: 424.

[16]. Winkler, S. M., et al. (2010). Classification of tumor marker values using heuristic data mining methods. Proceedings of the 12th annual conference companion on Genetic and evolutionary computation, ACM.

[17]. Howard, D., et al. (1999). "Target detection in SAR imagery by genetic programming." Advances in Engineering Software **30**(5): 303-311.

[18]. Sung, K.-K. (1996). "Learning and example selection for object and pattern detection."

[19]. Jin, Y. (2006). Multi-objective machine learning, Springer Science & Business Media.
[20]. Ganganwar, V. (2012). "An overview of classification algorithms for imbalanced datasets." International Journal of Emerging Technology and Advanced Engineering **2**(4): 42-47.
[21]. Chawla, N. V., et al. (2004). "Editorial: special issue on learning from imbalanced data sets." ACM Sigkdd Explorations Newsletter **6**(1): 1-6.
[22]. Barandela, R., et al. (2003). "Strategies for learning in class imbalance problems." Pattern Recognition **36**(3): 849-851.
[23]. Kubat, M. and S. Matwin (1997). Addressing the curse of imbalanced training sets: one-sided selection. ICML, Nashville, USA.
[24]. Breiman, L. (1996). "Bagging predictors." Machine learning **24**(2): 123-140.
[25]. Bradley, A. P. (1997). "The use of the area under the ROC curve in the evaluation of machine learning algorithms." Pattern Recognition **30**(7): 1145-1159.
[26]. Caruana, R. and A. Niculescu-Mizil (2004). Data mining in metric space: an empirical analysis of supervised learning performance criteria. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.
[27]. Song, D., et al. (2005). "Training genetic programming on half a million patterns: an example from anomaly detection." Evolutionary Computation, IEEE Transactions on **9**(3): 225-239.
[28]. Van Hulse, J., et al. (2007). Experimental perspectives on learning from imbalanced data. Proceedings of the 24th international conference on Machine learning, ACM.
[20]. Estabrooks, A., et al. (2004). "A multiple resampling method for learning from imbalanced data sets." Computational intelligence **20**(1): 18-36.
[30]. Jo, T. and N. Japkowicz (2004). "Class imbalances versus small disjuncts." ACM Sigkdd Explorations Newsletter **6**(1): 40-49.
[31]. Guo, H. and H. L. Viktor (2004). "Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach." ACM Sigkdd Explorations Newsletter **6**(1): 30-39.
[32]. Min, S.-H., et al. (2006). "Hybrid genetic algorithms and support vector machines for bankruptcy prediction." Expert Systems with Applications **31**(3): 652-660.
[33]. Kelly Jr, J. D. and L. Davis (1991). A Hybrid Genetic Algorithm for Classification. IJCAI.
[34]. Kao, Y.-T. and E. Zahara (2008). "A hybrid genetic algorithm and particle swarm optimization for multimodal functions." Applied Soft Computing **8**(2): 849-857.