Classification of Imbalanced Data:Review of Methods and Applications

.

Classification of Imbalanced Data:Review of Methods and Applications

Pradeep Kumar[1], Roheet Bhatnagar[2], Kuntal Gaur[1], and Anurag Bhatnagar[3]
[1]Department of Computer Applications,Manipal University Jaipur, Jaipur-Ajmer Express Highway, Dehmi Kalan, Near GVK Toll Plaza, Jaipur, Rajasthan-303007
[2]Department of Computer Science and Engineering,Manipal University Jaipur, Jaipur-Ajmer Express Highway, Dehmi Kalan, Near GVK Toll Plaza, Jaipur, Rajasthan-303007
[3]Department of Information Technology,Manipal University Jaipur, Jaipur-Ajmer Express Highway, Dehmi Kalan, Near GVK Toll Plaza, Jaipur, Rajasthan-303007

E-mail: pradeep.kumar@jaipur.manipal.edu

Abstract. Imbalance in dataset enforces numerous challenges to implement data analytic in all existing real world applications using machine learning. Data imbalance occurs when sample size from a class is very small or large then another class. Performance of predicted models is greatly affected when dataset is highly imbalanced and sample size increases. Overall, Imbalanced training data have a major negative impact on performance. Leading machine learning technique combat with imbalanced dataset by focusing on avoiding the minority class and reducing the inaccuracy for the majority class. This article presents a review of different approaches to classify imbalanced dataset and their application areas.

1. Introduction

Most of the practical machine learning techniques needs supervised learning. Supervised learning method iterative makes forecast on the labelled training data. Classification is a technique which needs machine learning algorithms and labelled training data to gain how to designate class label to sample from the domain. Binary classification invokes to forecasting one of the two class; majority or minority [1]. Data are classified as balanced or imbalanced data in binary classification techniques.

The issue with imbalanced data distribution appears when the segment of majority class has a greater proportion than minority class because data instances of importance are low in quantity [2]. In various medical peculiarity, often the training samples has very small amount of instances of interest. The accuracy of such models induced from

the imbalanced dataset would be not acceptable. The tendency of classification model is more bias towards the majority class any may avoid the minority class altogether. Most classifiers like decision tree and neural network function effectively when the distribution of response variable is balanced in dataset [3].When an event happens where percentage of minority class is reduced to 5% than it will be problematic to obtain good anticipating model because of small amount of information to understand about the rare event [3].For Example, a dataset where the ration between minority and majority class is 1:97 where 1% of the samples belong to the majority class, and 97% belongs to the minority class. Such classification may obtain the accuracy up to 97% by ignoring that 1% of minority class samples but it will result in the failure to correctly classify any instances of the class of interest.

In order to tackle highly imbalance data-set majorly three approached namely data level, combining methods and algorithmic level are used [4]. Random oversampling and under sampling are two technique considered under data level approach. Primarily the main objective of this approach is to balance again the skewed distribution of instances meanwhile, combining methods include mixture of experts approach. Machine learning algorithms are altered to cater imbalanced data for re sampling the class distribution [5].Imbalance data problem arises problems in numerous applications such as prediction of natural disaster, analysing biological disorders, medical diagnosis assisted by computer and big data analytic [6]. Almost all standard algorithms look for balanced class distribution therefore these applications fails to adequately represent the distributive properties of complex data or imbalanced data and causes adverse accuracy [7]. Data sampling [8] is considered as a common solution of data imbalance which modify the anatomy of actual data-set to alter its balance ratio to the expected level by under sampling and oversampling both. The aim of our review paper is to sum up latest research work on class imbalance issues. The purview of our literature is to explore research works conducted within the past few years which focus on the class imbalance, and the subsequent solutions developed.

The remaining paper is organized as follows. Section 2 discusses adverse effects of imbalanced dataset on various machine learning techniques. Section 3 explains various pre processing techniques like sampling for classification of imbalanced data. Section 4 discusses different domains for application of data classification. Section 5 concludes the paper.

2. Impacts of imbalanced dataset on classification

2.1. Imbalanced Data set

In recent scenario most of researchers faces challenges while classifying data which has diverse distribution. A dataset is imbalanced if one set of data has high dominance over the second set of data [9].While learning, several techniques such as support vector machine faces great challenge in presence of imbalanced data. Processing of imbalanced

dataset comprises of two parts. First to decrease the ratio of imbalance in data sets [10] and second, selection of the important features from dataset. Imbalance is usually represented as ratio between the total number of occurrence in majority class and the total number of occurrence in the minority class. Properties of imbalanced data which makes the task of classification very difficult are overlapping small dis junction [11], density lack, and noisy data and dataset deviation.

2.2. Impacts of imbalanced data set

Supervised learning is the basis of classifiers which require adequate training before the prediction process. Generally classifiers consider that the dataset is balanced throughout the training process. Perfect classification requires equal representation of all the classes accommodated in the data set [11] [12].They require balanced representations of all the classes contained in a dataset to perform effectively. Churn is very uncommon article and misclassification is high in cost for rare events or object in imbalanced scenario. Therefore inaccurate results may be provided on the imbalanced dataset by traditional approaches such as many real world applications e.g. face recognition, automatic glaucoma detection and anomaly detection are suffering from imbalanced class problem. Accuracy of many classification is severely affected by imbalanced dataset in machine learning approaches. Actual time environment is comparatively more prone to inconsistencies like imbalance and data with noise. Hugeness of data disentangle the process of prediction.forecasting performed based upon such imbalanced data usually does not effective when seen from the reliability point of view. Supervised learning techniques requires appropriate data for learning. Classifiers predicts the model based upon learning which can be interrupted by insufficient data for training. Insufficiency of data occurs due to class imbalance problem [9]. Supervised learning techniques requires appropriate data for learning. Classifiers predicts the model based upon learning which can be interrupted by insufficient data for training. Insufficiency of data occurs due to class imbalance problem. Massive data is one another problem as it nettle such problem diversified. It becomes challenging in supervised classification and contrast pattern mining using contrast patterns due to high computational cost because of exponential amount of candidate patterns [13]. But such problems causes a severe challenge on data sets which are imbalanced since there are many hindrance that emerges when contrast pattern mining [14] is done from imbalanced dataset. Pattern mining can be biased towards the majority class therefore they mainly extract certain contrast pattern from such class but small amount of pattern from another class. Despite, mainly the minority class has high importance although it is problematic to have sufficient amount of objects because it could be correlated to exponential class [15].

Classification of imbalanced dataset is a leading issue which includes identification of appropriate performance metrics [16] required. It is seen that data imbalance employ a substantial impact on the value and implication of accuracy and on other widely known performance matrices. Data imbalance can be more particularity of “insufficient” in

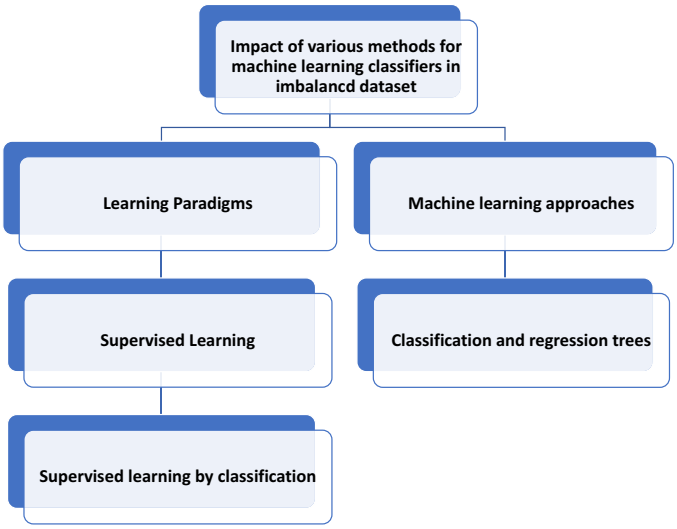

Figure 1: Impact of imbalanced dataset on Machine learning classifier [17]

feature space comparative to class imbalance. Inter class distribution and Distribution of data within each class are two major problems occurs because of Sparsity [18] in data. Such issues are also associated to problem of small disjoint. Important aspect of above methods include the determination of classification performance for obtaining a figure which need to be optimized and in order to obtain the final result by adjusting the classifier parameters. Many matrices are mainly chosen because there is no specific way to identify the optimum algorithm as many algorithm can achieve batter results in one of the class but poor result in another class. Thus the impact of imbalance data becomes a vital issue upon classification performance matrices. Impact of imbalanced dataset on Machine Learning classifier is highlighted in figure 1.

3. Imbalanced data classification techniques

Various state of art learning techniques have been suggested in past few years to address classification problem in imbalance dataset. These techniques are figured out with the introductory machine framework of machine learning modeling [19]. Two basic approaches which are addressed for learning of imbalance data are algorithm level methods and data level methods. Algorithm driven approach pursue with balancing the class distribution where data driven approach attempts to either redress the learning algorithm or classifier without altering the training set. Data driven approach includes random under sampling and oversampling. Under sampling eliminates instance of majority class randomly in order to balance the dataset [20]. Plenty

Classifier learning Algorithm	**Learning Strategy**	**Limitations**
Random Forest	Combine the output generated by multiple classifiers	Biased with non-lineal correlation between independent and dependent variables
K-Nearest Neighbour	It pin down class label by upper rank class amongst k nearest neighbours	Carries the higher possibilities of instances from frequent class
Decision Trees	Subdivide training data coercive and shorten sub trees if glitch occurs	Many division required to find imbalance
Neural Networks	Repeatedly adjustment of weight to minimize the error	Minimize errors only for frequent classes.
Naives Bayes	Consider flexibility among features	Wrong classification of instances of small class
Support Vector Machine	Detect optimal hyper plane dissociation with maximal margin	Conscious to imbalanced dataset and decision boundary have bias towards small class

Table 1: Overview of learning strategy and learning algorithm [21]

of under sampling techniques like Condensed Nearest Neighbour, Edited Nearest Neighbour, Neighbourhood Cleaning, Tomek Links and One-sided selection have been proposed. Second common sampling approach is oversampling which replicates the objects of minority class randomly until they have uniform representation. Various oversampling techniques which includes focused oversampling, synthetic sampling, random oversampling and some advance heuristic techniques like synthetic minority oversampling (SMOTE) are available. SMOTE is a strategy which over sample the object of minority class by generating synthetic samples [22]. Algorithm driven approach which is also known as classifier level approach keep the training dataset invariable and adjust the inference algorithm to facilitate the task of learning specifically related to minority class. This method includes hybrid methods like ensemble and classical methods like thresolding, one class classification and cost sensitive learning [23]. Thresolding is a set of methods which is used to remodel the decision boundary in a classifier as threshold moving .Certain algorithm such as Nave Bayes which can generate probability estimates and convert it into predictions . Cost sensitive includes modification in learning rate so that the higher cost instances contribute more in numbers in order to update the weightage. Then we can train by reducing the misclassification cost in lieu of standard loss function equivalent to oversampling. One-

class classification is a method which observe positive objects rather than differentiating between two results. Identity function is used for such purpose which is trained to implement associative mapping. Based upon absolute errors and squared sum of errors between the IO patterns the classification of new example is designed. Hybrid of methods approach combines multiple classification techniques from above mentioned categories.It can be seen as wrap up of other methods such as ensembling which is widely used as a classification techniques. Several method such as Balance Cascade and Easy Ensemble are used to train a group of classifiers on under sampled subgroups. This method include pre-training and fine-tuning on the original imbalanced dataset.Table 1 presents overview of learning techniques.

4. Applicable Domains of Imbalanced data classification

Imbalanced data distribution reduces the classification accuracy that emerges as a big concern in many real world applications. Machine learning and data mining analyse the large amount of data using several automated methods and predict future event which is based upon past events [24]. Rare events are those events which occurs less frequent such as natural disasters like earthquake, solar flares and many anthropogenic adverse effects like violent conflict, financial fraud and diseases. Different application of imbalanced learning techniques belongs to engineering, management, biology and other domains. In medical science misclassification in imbalanced dataset is a very important aspect which need to be addressed early in the medical test so that right course of treatment may be started at right time else it can lead to complex problems and prove fatal. Other application areas like pollution predictions, predicting emergency events and so on belong to this category [25]. Computer vision is an area where imbalance happens when maximum amount of images belongs to training data does not have the instance of interest which leads to failure of most of all machine learning algorithms [21]. Various issues that involves security, trust and privacy in the applications of information security also uses imbalance learning techniques while deciding the boundary for classification of anomalies. Applications of image processing uses such techniques in order to distinguish the distribution between mis-classified features and novel features. Detection of the anomalies in clouds becomes a problematic issue in imbalanced classification. Data mining, text classification and bioinformatics are emerging areas where imbalanced data has large number of applications.

5. Conclusion

This article presents a review of various learning issues due to Imbalanced distribution of data and different approaches to handle problem of imbalanced data in classification. Data imbalance reduces the accuracy and performance of classifier. This study has described the classification problem of imbalanced data and has focused upon different techniques to tackle with class imbalance problem. Normally Minority class samples are

predicted poorly by various machine learning models since supervised learning algorithm always pay attention to the samples of majority class. The impact of imbalance on classification is baleful and the effect increases with the extent of a task. Sampling method are prominent and very simple to implement but actual real life applications involved with biased data distribution so hybrid approaches are also required. However classification of imbalanced data is an extensive research subject in the field of machine learning.

References

[1] Johnson J M and Khoshgoftaar T M 2019 *Journal of Big Data* **6** 27
[2] Amin A, Anwar S, Adnan A, Nawaz M, Howard N, Qadir J, Hawalah A and Hussain A 2016 *IEEE Access* **4** 7940–7957
[3] Li J, Fong S, Mohammed S and Fiaidhi J 2016 *The Journal of Supercomputing* **72** 3708–3728
[4] Yap B W, Abd Rani K, Abd Rahman H A, Fong S, Khairudin Z and Abdullah N N 2014 *Proceedings of the first international conference on advanced data and information engineering (DaEng-2013)* (Springer) pp 13–22
[5] Galar M, Fernandez A, Barrenechea E, Bustince H and Herrera F 2011 *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **42** 463–484
[6] Li J, Fong S, Wong R K and Chu V W 2018 *Information Fusion* **39** 1–24
[7] He H and Garcia E A 2009 *IEEE Transactions on knowledge and data engineering* **21** 1263–1284
[8] Khoshgoftaar T M, Fazelpour A, Dittman D J and Napolitano A 2015 *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)* (IEEE) pp 705–712
[9] Somasundaram A and Reddy U S 2016 *International Conference on Research in Engineering, Computers and Technology (ICRECT 2016)* pp 1–16
[10] Li J and Fong S 2016 *Bio-Inspired Computation and Applications in Image Processing* (Elsevier) pp 311–321
[11] More A and Rana D P 2017 *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM)* (IEEE) pp 72–78
[12] Gong J and Kim H 2017 *Computational Statistics & Data Analysis* **111** 1–13
[13] Yu K, Ding W, Simovici D A and Wu X 2012 *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* pp 60–68
[14] Dong G and Bailey J 2012 *Contrast data mining: concepts, algorithms, and applications* (CRC Press)
[15] Weiss G M and Tian Y 2008 *Data Mining and Knowledge Discovery* **17** 253–282
[16] Luque A, Carrasco A, Martín A and de las Heras A 2019 *Pattern Recognition* **91** 216–231
[17] Loyola-González O, Martínez-Trinidad J F, Carrasco-Ochoa J A and García-Borroto M 2016 *Neurocomputing* **175** 935–947
[18] Chawla N V 2009 *Data mining and knowledge discovery handbook* (Springer) pp 875–886
[19] Thabtah F, Hammoud S, Kamalov F and Gonsalves A 2020 *Information Sciences* **513** 429–441
[20] Kotsiantis S, Kanellopoulos D, Pintelas P *et al.* 2006 *GESTS International Transactions on Computer Science and Engineering* **30** 25–36
[21] Kaur H, Pannu H S and Malhi A K 2019 *ACM Computing Surveys (CSUR)* **52** 1–36
[22] Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 *Journal of artificial intelligence research* **16** 321–357
[23] Buda M, Maki A and Mazurowski M A 2018 *Neural Networks* **106** 249–259
[24] Weiss G M and Hirsh H 2000 *AAAI workshop on learning from imbalanced data sets* (AAAI Press) pp 64–68

[25] Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H and Bing G 2017 *Expert Systems with Applications* **73** 220–239