Unbalanced Data Set- State-of-the-art and its Research Challenges

Deeksha Dhapola
M.Tech Scholar
Graphic Era Hill University
Bhimtal, India

Janmejay Pant
Asst. Professor
Graphic Era Hill University
Bhimtal, India

ABSTRACT
Real world application often found the problem of unbalanced dataset. This then create the problem in machine learning methods . In this paper we have surveyed the imbalance dataset problem at the algorithmic level . By over sampling and under sampling some researchers artificially prove that updated svm ,cost sensitive classifier ,class orientation methods can be good on imbalanced dataset. This imbalance problem is also switching towards hybrid algorithm.

Keywords
cost-sensitive learning, imbalanced data set, modified SVM, oversampling, undersampling

1. INTRODUCTION
Imbalanced dataset means when one class contains some samples then rest of the classes. These are unbalanced when any one class contains small minority classes. Classifier applies good accuracy on major classes but weak in minor classes. Classification algorithm minimised the label of wrong prediction. These classification algorithm assumes the error of cost equally.

In real world examples the assumption of classification is not true.During medical diagnosis of a cancer patient , if having cancer is under the class A and non cancer in class B then missing cancer implies patient is healthy but is classified on negative , so it is a false negative error . patient may loose its life due to the delay in diagnosis.

Unbalanced dataset have certain examples of real world like fault detection , fraud detection , medical diagnosis , cultural modelling and many more. Szil Vajda et al[1] has proposed a method for imbalanced dataset. In this method it has shown that the performance of existing classification is biased towards majority classes. Their poor performance over imbalanced dataset was due to the following points i. Classifier aim is to reduce the overall performance error where minority classes contributes less.

ii. Their methods assume the equal distribution of data for all the classes.

iii. Erorrs causing from other classes cost same [2]

Both data and algorithmic way class imbalanaced problem was proposed earlier. It includes either resampling , undersampling ,oversampling and directed sampling. At algorithmic level solutions are given for adjusting the imbalance methods to adjust its cost of various classes. Adjuting decision threshold in decision tree and recognition based learning and discrimination based learning . Some more techniques to deal with when unbalanced dataset contains resizing of process data , cost effective classifier and snowball method. Some approaches includes k-means neighbour (KNN) , modified SVM , probabilistic decision tree and learning methods.

Next section contains some methods details explaination limitation are attached for cost sensitive machine learning . Undersampling discards the useful data with oversampling . Oversampling limitation shows that it creates existing from examples. Oversampling method is quite easy to generate classification rule and it also increases not of training example and learning time.

The easiest method of Data level for balancing the different classes contains re-sampling of the data set, by oversampling or under-sampling , until the classes are exactly equally represented. Both methods can be applied in any learning system, permitting the learning system to receive the training sets as if they sets to a balanced data set. Any bias of the system for the majority class is due to the different proportion of datasets .

Hulse et al. [4] proposed that the performance of the resampling methods depends on a ratio between positive and negative classes, other features of data, and the working of the classifier. However, resampling methods have shown limitations. Under-sampling may exit potentially useful data, while over-sampling artificially uplift the size of the data set and further, avail the computational overhead of the learning algorithm.

A. Oversampling
Non heuristic methods are the easiest method to enhance the data size of the minority class wrt to random over-sampling. This method manages the class distribution through the replacement of positive classes using random method. Since this method replaces existing datasets in the minority class.

Chawla et al[5] proposed Synthetic Minority Over-sampling Technique (SMOTE) an over-sampling method in which the minor clasess is over-sampled. introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors The minority class is over-sampled by taking each minority class sample . Based on the over-sampling requirement, the k nearest neighbors chose random neighbour. From the SMOTE algorithm, several updation have been proposed . While SMOTE does not tackle data sets with fundamental features, it was easy to operate and mixed datasets of normal features. He also proposed SMOTE-NC (Synthetic Minority Over-sampling Technique Nominal Continuous) and SMOTE-N (Synthetic Minority Over-sampling Technique Nominal).

Andrew brooks et al.[6] proposed a re- sampling method which checks the most exact re-sampling rate . Taeho et al.[7]proposed a cluster-based sampling method which works with class imbalance and within-class imbalance parallel . Hongyu et al[9] has found and shown complex examples of

the major and minor classes during the boosting processs, then created the new examples from hard datasets and include them to the major classes .Based on SMOTE method, Hui Han [9] proposed two new over-sampling methods, borderline-SMOTE1 and borderline-SMOTE2, in which only the minority examples near the borderline are verified and over- sampled. These methods maintains good performance rate and F- value than SMOTE .

B. Undersampling

Under-sampling is a method for identifying imbalance learning. This method uses major classes to make the classifer expert. Many majority class datasets are neglected, the resultant training dataset becomes balanced and the training operation becomes faster. Random majority under-sampling (RUS), is the most popular undersampling methods.In RUS, classes of the majority are randomly neglected from the dataset.

The main limitation of under-sampling is that useful dataset contains neglected examples. Methods like Tomek links, Condensed Nearest Neighbor Rule and One-sided selection etc. are the methods to improve the performance of random sampling .Rule Kubat and Matwin proposed random side selection which tries to intelligently manage under-sample the majority class by deleting errors in major classes which are either redundant or ‗noisy.' Over-sampling improves minority class identification, by randomly creates multiplee minority data without increasing the small number of new information.

T. Maruthi et al[10] proposed the method of hybrid sampling, in which combining of SMOTE to over-sample the minority data (fraud samples) and random under- sampling to under-sample the majority data (non-fraud samples) if we eliminate extreme problems from the minor examples highly imbalanced data sets like fraud detection classification accuracy can be modified.

Class skew and properties of the dataset is contained in sampling methods., nontrivial datasets is offered by machine learning and data mining with the extra features and properties remotely. Classifier improves through sampling levels which may be different in the data, resulting in optimum performance.David A. et al[11] has suggested that for improving performance of the classifier sampling can be treated remotely, instead of uniform levels of samples. Their proposed a framework first searches useful data sets and then find optimum sampling levels.

The limitations are attached for cost sensitive Machine Learning . Undersampling discards the useful data with oversampling. Oversampling limitations shows that it creates existing copies from the examples. Oversampling method is quite easy to generate the classification rule and it also increases the number of training examples and learning time.

Imbalanced dataset is always managed by sampling methods as an effective method. Some of the reasons are due to cost ineffectiveness wrapper based method is one of the method. Many more learning algorithms like c4.5 still do not managed cost in learning process. Secondly many skilled datasets are good and training sets must be reduced to train data.

If some training datasets has to be skipped majority classes also reduces the training size of the dataset and shows sensitive arm learning algorithm . Another reason may share misclassification cost other than cost sensitive problems . If cost information is not known steps like ROC curve can be used to measure classifier efficiency[12].

2. COST SENSITIVE LEARNING

Cost sensitive learning means data-mining which takes misclassification cost in priority . Cost sensitive learning can be implemented by many methods[13]. Its classification can be shiwn in three steps.

First step is a technique to apply misclassification cost to dataset . Second step applies cost reduction to the ensemble methods and third fits to cost sensitive framework.

In Decision tree classification cooperating cost also shows sample classification. Cost can be shown in various ways.

First method shows applying simple classification in decision threshold.

Second method distributes the attributes is pruning schemes. Ref.[14] shows a testing decision tree that reduces test cost and misclassification.

In this method algorithm uses separation attributes which reduces cost and also entropy. In ref [16] class confidence proportion decision tree (CCPDT) is proposed which is insensitive with classes and shows statistics reformation.Hellinger distance shows strong operational empirically [17] and imbalanced dataset is sufficient to use Hellinger method for evaluating the performance of Hellinger .AUC has proposed the algorithm better Nguyen et al[19] has proposed an Extended Regularized Least Square .(RLS) that accepts the errors of dissimilar samples. Jie sang[20] proposed the method BABoost that is a modification of Adaboost. Adaboost method gives equal weightage to each dataset by misclassification errors are not same. So Adaboost loads to higher majority when showing the distribution. Yanmin Sun et al[20] in his research has shown three cost sensitive algorithm which are developed by imbalanced datasets into the framework of Adaboost . Boosting methods shows their performance in terms of samples on real world medical data sets and imbalance problems.

3. SVM AND IMBALANCED DATASETS

Apart from the better functionality of SVM it generates some limitation when applied on imbalanced dataset. Even SVM performs well when applied over majority classes under sampling . rehan et all=[12] proposed a combination of cost effective learning and undersampling . The algorithm was based on SMOTE algo by Chawla et al. TAO Xia-yan et all[20] proposed a modified support vector machine which assigns positive and negative datasets by optimizing methods for high generalized optimization . Real – coded immune clone algorithm has been improved for SVM. M. Muntean et al[20] proposed a variable algorithm for improving SVM classification of imbalanaced data sets.

Yunchun Tang [21] has proposed SVM model. Fitness Function has improve the SVM by ROC curve AUC[19]. These methods has improved both the major and minor classes. But fitness function has one limitation as it works on a specific domain only.

4. CONCLUSION

In this paper we have shown the detail classification of imbalanced datasets. Sampling techniques works well at data level for imbalanced datasets . Under classifier performance undersampling works poor as compared with oversampling . At algorithmic level solutions on SVM techniques performs well for minority based learning methods. Classifier

development can be efficient from future research in imbalanced dataset.

5. REFERENCES

[1] 1. Szil´ard Vajda, Gernot A. ―Fink Strategies for Training Robust Neural Network Based Digit Recognizers on Unbalanced Data Set 2010‖ 12th International Conference on Frontiers in Handwriting Recognition

[2] C.V. KrishnaVeni,T. Sobha Rani On the Classification of Imbalanced Datasets‖ IJCST Vol . 2, SP 1, December 2015

[3] Nitesh V. Chawla, Nathalie Japkowicz,

[4] Special Issue on Learning from Imbalanced Data Sets‖ Sigkdd Explorations. Volume 6, Issue 1

[5] .Chawla,NBowyer, K., Hall, L. Kegelmeyer, W. ―SMOTE: Synthetic minority over-sampling technique‖ of Artificial Intelligence Research 16, 321–357 (2015)

[6] Andrew Estabrooks, Taeho Jo and Nathalie Japkowicz ―Multiple Resampling Method for Learning from Comprtational Intelligence 20 (1) (2009).

[7] Taeho Jo, Nathalie Japkowicz ―Class Imbalances versus Small Disjuncts‖. Sigkdd Conference IEEE 2011.

[8] Hongyu Guo, Herna L Viktor: ―Learning from Imbalanced Data Sets with Boosting and Data Generation: The DataBoost-IM Approach‖. Sigkdd Explorations 6 (1) (2015).

[9] Hui Han, Wen-Yuan Wang, Bing-Huan 4th International conference 2011, Malaysia.

[10] David A. Cieslak, Nitesh V. Chawla ―Start Globally, Optimize Locally, Predict Globally: Improving Performance on Imbalanced Data‖ 2012 Eighth IEEE International Conference on Data Mining.

[11] Gary M. Weiss, Kate McCarthy, and Bibi Zabar Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error costs?

[12] Haibo He, Edwardo A. Garcia, ― Learning from Imbalanced Data‖, 2012.

[13] Charles X. Ling, Qiang Yang, Jianning Decisions tree with minimum costs.

[14] David A. Cieslak, Nitesh V. Chawla, Learning decision tree for unbalanced datasets. 2015.

[15] Wei Liu, Sanjay Chawla, David A. Nitesh v Chawala for imbalanced datasets. 2013

[16] David A. Cieslak, T. Ryan The journal of Data mining issue May 2016.

[17] Satyam Maheshwari, Prof. Jitendra A New Approach for Classification of imbalanced datasets Evolutionary algorithm 2011.

[18] NGUYEN HA VO, YONGGWAN WON ―Classification of Unbalanced Medical Data with Weighted .Convergence of bio science technology 2015.

[19] Jie Song, Xiaoling Lu, Xizhi Wu ―An Improved AdaBoost Algorithmfor Unbalanced Classification Data‖ 2009 Sixth International Conference on Fuzzy 2012.

[20] Yanmin Sun, Mohamed S. Kamel, Andrew K cost sensitive boosting on imbalanced dataset 2013.

[21] Rehan Akbani, Stephen Kwek Nathalie Japkowicz Applying Support Vector Machines to Imbalanced Dataset.

[22] TAO Xiao-yan, JI Hong-bing AModifiedPSVM and itsApplicationtoUnbalancedDataClassification.ThirIntern ational Conference on NaturalICNS 2017