International Institution for Technological Research and Development

Volume 1, Issue 2, 2015

Survey on Classification Algorithms for Imbalanced Dataset

Amee Rajan[1], Chetna Chand[2] and Dr. G.R Kulkarni [3]

[1] Department of computer Engineering, Kalol Institute of Technology & Research Centre
Kalol , Gujarat , India
amee2105@gmail.com
[2] Department of computer Engineering , Kalol Institute Of Technology & Research Centre
Kalol , Gujarat , India
chetnachand87@gmail.com
[3] Department of computer Engineering, Kalol Institute Of Technology & Research Centre
Kalol , Gujarat , India
grkulkarni29264@rediffmail.com

Abstract

Real world application suffers from imbalanced dataset. There have been many attempts at dealing with classification of imbalanced data sets. Classification of imbalanced dataset is an evoloving trend in research area of data mining. In this paper we have briefly described the methods for classifying imbalanced dataset .we can categorise classification methods in three categories as data-level approach ,algorithm level approach and cost sensitive approach . we have studied various research work done on classifying imbalance dataset and briefly described it.

Keywords: *Imbalanced Data, undersampling, oversampling, cost-sensitive approach , algorithm level approach*

1.Introduction

Class imbalance problem is a hot topic being investigated recently by machine learning and data mining researchers. It can occur when the instances of one class is more than the instances of other classes. The class have more instances called the majority class while the other called minority class. However, in many applications the class has lower instances are the more interesting and important one. The imbalance problem heightens whenever the class of interest is relatively rare and has small number of instances compared to the majority class. Moreover, the cost of misclassifying the minority class is very high in comparison with the cost of misclassifying the majority class for example; consider cancer versus non-cancer or fraud versus un-fraud [2].The patient could lose his/her life because of the delay in the correct diagnosis and treatment. Similarly, if carrying a bomb is positive, then it is much more expensive to miss a terrorist who carries a bomb to a flight than searching an innocent person.

Many real world applications such as medical diagnosis, fraud detection (credit card, phone calls, insurance), network intrusion detection, pollution detection, fault monitoring, biomedical, bioinformatics and remote sensing (land mine, under water mine) suffer from these phenomena. In this scenario, classifiers can have good accuracy on the majority class but very poor accuracy on the minority class(es) due to the influence that the larger majority class has on traditional training criteria.

Many research papers on imbalanced data sets have commonly agreed that because of this unequal class distribution, the performance of the existing classifiers tends to be biased towards the majority class. The reasons for poor performance of the existing classification algorithms on imbalanced data sets are: 1.They are accuracy driven i.e.,their goal is to minimize the overall error to which the minority class contributes very little. 2. They assume that there is equal distribution of data for all the classes. 3.They also assume that the errors coming from different classes have the same cost. With unbalanced data sets, data mining learning algorithms produce degenerated models that do not take into account the minority class as most data mining algorithms assume balanced data set[3].

A number of solutions to the class-imbalance problem were previously proposed both at the data and algorithmic levels [3]. At the data level, these solutions include many different forms of re-sampling such as random oversampling with replacement, random undersampling,

directed oversampling (in which no new examples are created, but the choice of samples to replace is informed rather than random), directed undersampling (where, again, the choice of examples to eliminate is informed),oversampling with informed generation of new samples, and combinations of the above techniques. At the algorithmic level, solutions include adjusting the costs of the various classes so as to counter the class imbalance, adjusting the probabilistic estimate at the tree leaf (when working with decision trees), adjusting the decision threshold, and recognition-based (i.e., learning from one class) rather than discrimination-based (two class) learning.

2. Methods

2.1 Data Level Approach

An easy Data level methods for balancing the classes consists of resampling the original data set.Sampling is a set of methods that changes the size of training sets by adding or removing features from datasets and try to balance dataset. Under-sampling and over-sampling change the training sets by sampling a smaller majority training set and repeating instances in the minority training set. In both methods the level of imbalance is reduced to more balanced training set which can give better results. Both sampling methods have been shown to be helpful in imbalanced problems. Training time in under sampling is short, but can ignore useful data. Over sampling increases the training set size, and thus requires longer training time. Over sampling many times leads to over fitting because it repeats minority class examples.

2.2 Cost-Sensitive Approach

Cost-sensitive approaches can use both data and modifications of the learning algorithms. A higher misclassification cost is assigned for minority class objects and classification performed so as to reduce the overall learning cost. Costs are often specified in form of cost matrices. The lack of knowledge on how to set the actual values in the cost matrix is the main drawback of cost-sensitive methods, since in most cases this is not known from the data nor given by an expert[1].

2.3 Algorithm Level Approach

Trying to adapt existing algorithms to the problem of imbalanced datasets and bias them towards favoring the minority class known as Classifier level approaches. Here, some more in-depth knowledge about the nature of the used predictors and factors that cause its failure in minority class recognition is required.

3. Related Work

In 2003 Jinping Zhang and Inderjeet Mani approached kNN in a case study on information extraction in which undersampling is done on the kNN method.Under-sampling negative examples(here majority examples) is used to deal with imbalance data, in which a small subset of majority examples is selected .In this work selection of majority examples is carried out in different ways :random selection, selection of near-miss example and selection of most distant example[4].

In 2007, Yang Song et al. proposed iknn in which a point is treated informative if it is close to the query point and far away from the points with different class labels. They introduce a new metric that measures the informativeness of objects to be classified. When applied as a query-based distance metric to measure the closeness between objects, two novel KNN procedures, Locally Informative-KNN (LI-KNN) and Globally Informative-KNN (GI-KNN), are proposed. By selecting a subset of most informative objects from neighborhoods, this methods exhibit stability to the change of input parameters, number of neighbors(*K*) and informative points (*I*)[5].

In 2011 Yuxuan Li and Xiuzhen Zhang presented kENN which proposed a training stage where exemplar positive training instances are identified and generalized into Gaussian balls as concepts for the minority class. This paper propose to identify exemplar minority class training instances and generalize them to Gaussian balls as concepts for the minority class. *k* Exemplar-based Nearest Neighbor (*k*ENN) classifier is therefore more sensitive to the minority class[6].

.

Again In 2011 Wei Liu and Sanjay Chawla proposed CCW-kNN a novel *k*-nearest neighbors (*k*NN) weighting strategy is proposed for handling the problem of class imbalance. CCW-kNN uses the probability of attribute values given class labals to weight prototypes in kNN[7].

In 2011Saumil Hukerikar et al. proposed SkewBoost technique, in which minority and majority instances are identified during execution of the boosting algorithm. A variation of SMOTE is used to create synthetic minority instances which are then added to the training set and total weight is rebalanced. After each iteration of the boosting algorithm, the weight of each instance is modified to focus more on the misclassified instances. A cost-sensitive approach has been adopted to reweight the instances following every iteration[8].

In 2012 Ikram ChaYri et al. presented Learning from Imbalanced Data Using Methods of Sample Selection, A novel method to deal with the problem of imbalanced data was proposed that create a balance between the classes. It focuses learning in the most important samples, which improves the performance. In the realized experiments we have shown how the application of SS on the majority class can give us a better performance to our learning process by avoiding the selection of non-critical samples. Contrary to the random undersampling, this method allows us to keep all the important examples of the data set[9].

In 2013, Jinjin Wanga et al. focuses on many diseases data, whose patients are composed of majority normal persons and only minority abnormal ones. Many researchers ignored these imbalance problems, so their learning models usually led to a bias in the majority normal class. To deal with this problem, a new over-sampling technique was proposed to over-sample the minority class to balance the data samples and improve Support Vector Machine(SVM) in imbalanced diseases data sets. For the minority class, a K-Nearest Neighbor(KNN) graph is built. Second, the proposed technique gets a Minimum Spanning Tree(MST) based on the graph. Third, the proposed technique generates synthetic samples by using SMOTE along the direct path in the tree[10].

In 2014 Chunming Liu et al. proposed HC-*k*NN which considers relationship between features when computing the similarity. It calculates sized membership using fuzzy theory to deal with imbalance data problem then assigns feature weight to every feature .hybrid coupled k-nearest neighbor classification algorithm (HC-*k*NN) works on mixed type data, by doing discretization on numerical features to adapt the inter coupling similarity as we do on categorical features, then combing this coupled similarity to the original similarity or distance,then *k* nearest neighbors that correspond to the *k* highest similarity values are choosen. The most frequently occurred class in the *k* neighbors is assigned [11].

4. Conclusion

In this paper we have describe imbalanced data ,its techniques and solutions .we have studied many research paper of different technique , we have concluded form that among the several methods algorithm approach is better as it deals with the imbalance data problem without modifying inherent data structure unlike data level approach.In future,we can work on algorithm level approach to improve its accuracy of classification on imbalanced data.

References

[1] S.Jayasree and A.Alice Gavya, "Addressing imbalance problem in the class – A survey" ,International Journal of Application or Innovation in Engineering & Management (IJAIEM) Volume 03, Issue 09, September 2014

[2] Shaza M. Abd Elrahman and Ajith Abraham, "A Review of Class Imbalance Problem", Journal of Network and Innovative Computing, Volume 1, 2013, pp. 332-340

[3] Vaishali Ganganwar, "An overview of classification algorithms for imbalanced datasets" ,International Journal of Emerging Technology and Advanced Engineering Volume 2, Issue 4, April 2012

[4] Jianping Zhang and Inderjeet Mani, "kNN Approach to Unbalanced Data Distributions: A Case Study involving Information Extraction" Workshop on Learning from Imbalanced Datasets II ,ICML ,2003

[5] Yang Song, Jian Huang, Ding Zhou, Hongyuan Zha, and C. Lee Giles **,"**IKNN: Informative K-Nearest Neighbor Pattern Classification", Springer-Verlag Berlin Heidelberg 2007

[6] Yuxuan Li and Xiuzhen Zhang, "Improving k nearest neighbor with exemplar generalization for imbalanced classification," Springer-Verlag Berlin Heidelberg 2011

[7] Wei Liu and Sanjay Chawla ,"Class confidence weighted knn algorithms for imbalanced data sets", Springer-Verlag Berlin Heidelberg 2011

[8] Saumil Hukerikar, Ashwin Tumma, Akshay Nikam, Vahida Attar, **"**SkewBoost: An Algorithm for Classifying Imbalanced Datasets", IEEE 2011

[9]Ikram ChaYri, Souad Alaoui and Abdelouahid Lyhyaoui, "Learning from Imbalanced Data Using Methods of Sample Selection" , IEEE 2012

[10]Jinjin Wanga, Yukai Yaob, Hanhai Zhouc, Mingwei Lengd and Xiaoyun Chen, "A New Over-sampling Technique Based on SVM for Imbalanced Diseases Data", IEEE 2013

[11]Chunming Liu , Longbing Cao and Philip S Yu ,"A Hybrid Coupled k-Nearest Neighbor Algorithm on Imbalance Data.", International Joint Conference on Neural Networks (IJCNN), IEEE 2014