IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 7, September 2014.
www.ijiset.com
ISSN 2348 – 7968

A Study On Classification Of Imbalanced Data Set

[1]Mrs.A.Hema MCA,M.Phil, PGDBI,[2] Mrs.B. Kavitha M .Phil scholar (cs),

[1]Head Department of BCA, Kongunadu Arts& Science College, Coimbatore–29
[2]Kongunadu Arts & Science College, Coimbatore –29

ABSTRACT

Data Mining is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data Mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods. Consequently, Data Mining consists of more than collecting and managing data, it also includes analysis and prediction.Data mining applications can use a variety of parameters to examine the data. They include association, sequence or path analysis, classification, clustering, and forecasting. Classification is a data mining (machine learning) technique used to predict group membership for data instances.The decision tree learning method is one of the methods that are used for classification or diagnosis. As for many other machine learning methods, the learning in decision trees is done by using a data set of already classified instances to build a decision tree which will later be used as a classifier. The set of instances used to “train” the decision tree is called the training set. A data set is imbalanced if the classes are not approximately equally represented. There have been attempts to deal with imbalanced data sets in domains such as fraudulent telephone calls telecommunications management text classification.

Keywords:

Imbalanced data, classification, under sampling, cost sensitive, over sampling, decision tree.

INTRODUCTION

A data set is called imbalanced if it contains many more samples from one class than from the rest of the classes. Data sets are unbalanced when at least one class is represented by only a small number of training examples called the minority class while other classes make up the majority. In this scenario, classifiers can have good accuracy on the majority class but very poor accuracy on the minority classes due to the influence that the larger majority class has on traditional training criteria. Most original classification algorithms pursue to minimize the error rate: the percentage of the incorrect prediction of class labels. They ignore the difference between types of misclassification errors. In particular, theyimplicitly assume that all misclassification errors cost equally.

Imbalanced class distributions are frequently encountered in real-world classification application arising from fraud detection, risk management, text classification, medical diagnosis, and many other domains. Such imbalanced class data sets differ from balanced class data sets not only in the skewness of class distributions, but also in the increased importance of the minority class. Despite their frequent occurrence and huge impact in day-to-day applications, the imbalance issue is not properly addressed by many standard machine-learning algorithms, because they assume either balanced class distributions or equal misclassification costs.

The class imbalance problem is prevalent in many applications, including: fraud and intrusion detection, riskmanagement, text classification, and medicaldiagnosis or monitoring, etc. A number of solutions to the class-imbalance problem were proposed both at the data and algorithmic levels.

LITERATURE SURVEY

Chao Chen,AndyLiaw,LeoBreiman says there are two ways to deal with the imbalanced data classification problem using random forest. One is based on cost sensitive learning, and the other is based on a sampling technique.Performance metrics such as precision and recall, false positive rate and false negative rate, F-measure and weighted accuracy are computed. Both methods are shown to improve the prediction accuracy ofthe minority class, and have favorable performance compared to the existing algorithms.

Alberto Fernandez,Victoria Lopez, MikelGalar c, Maria Jose del Jesus , Francisco Herrera**,**is first to make use of binarization schemes, i.e., one versus one and one versus all, in order to apply the standard

ISSN 2348 – 7968

approaches to solving binary class imbalanced problems. Second, they apply several ad hoc procedures which havebeen designed for the scenario of imbalanced data-sets with multiple classes.Their goal is to show the optimal combination between binarization techniques: either with preprocessing approaches (oversampling and under sampling), or with the use of cost-sensitive learning formultiple-class imbalanced data-sets, in the case of both the OVO and OVA approaches. They also seek toexperimentally determinethe degree of synergy achieved between the combination of ‘‘divide- and-conquer’’ techniques (OVO and OVA) and preprocessing/cost-sensitive learning by contrasting their results with those of the approaches specifically designed to address imbalancedclassification problems in the scenario of multiple classes.

Wei Liu and Sanjay Chawla David A. Cieslaky and Nitesh V. Chawla,they propose a new decision tree algorithm, Class Confidence Proportion Decision Tree (CCPDT), which is robust andinsensitive to class distribution and generates rules which are statistically significant. In order to make decision trees robust, they begin byexpressing Information Gain, the metric used in C4.5, in terms of confidence of a rule. This allows them to immediately explain why Information Gain, like confidence, results in rules which are biased towards the majority class. To overcome this bias, they introduce a new measure, Class Confidence Proportion (CCP), which forms the basis ofCCPDT. To generate rules which are statistically significant they design a novel and efficient top-down and bottom-up approach which uses Fisher’s exact test to prune branches of the tree which are not statistically significant. Togetherthese two changes yield a classifier that performs statistically better than not only traditional decision trees but also treeslearned from data that has been balanced by well-known sampling techniques. Their claims are confirmed through extensive experiments and comparisons against C4.5, CART,HDDT andSPARCCC.

The authors Vladimir Nikulin1, Geoffrey J. McLachlan1, and Shu Kay Ng**,**deals with a large number of relatively small and balanced subsets whererepresentatives from the larger pattern are to be selected randomly. As an outcome, the system produces the matrix of linear regression coefficientswhose rows represent random subsets and columns represent features. Based on the above matrix they make an assessment of how stable theinfluence of the particular features is. It is proposed to keep in the model only features with stable influence. The final model represents an averageof the base-learners, which are not necessarily a linear regression. Test results against datasets of the PAKDD-2007 data-mining competitionare presented.They describe the method of random sets and mean-variance filtering and discuss general principals of theAdaBoost and LogitBoost Algorithms and explain the experimental procedure and the most important business insights. Finally, they conclude the paper.

Kittipong Chomboon, Kittisak Kerdprasop, and Nittaya Kerdprasop, proposed a method to induce a classification model of minority data cases that are always predominant by a much larger majority cases. Their proposed method applies feature selection technique to choose 25% of attributes that show highly correlation between classes. Then use over-sampling technique on minority cases before making a classification. They have found from the experimental results that data of rare cases, which is normally disappeared, can be detected through their proposed method. In this research, they use R language for implementing their proposed method and other four discovery techniques.

Yetian Chen tells that there are two classification tasks based on data from scientific experiment.
The first task is a binary classification task which is to maximize accuracy of classification on an evenly-distributed test data set, given a fully labeled imbalanced training dataset. The second task is also a binary classification task, but tomaximize the F1-score of classification on a test data set,given a partially labeled training set. For task 1, he investigated several re-sampling techniques in improving thelearning from the imbalanced data. These include SMOTE(Synthetic Minority Over-sampling Technique), Oversamplingby duplicating minority examples, random undersampling. These techniques were used to create newbalanced training data sets. Then three standard classifiers Decision Tree, Naïve Bayes, Neural Network were trained on the rebalanced training sets and used to classify the testset. The results showed the re-sampling techniquessignificantly improve the accuracy on the test set except forthe Naïve Bayes classifier. For task 2, he implemented twostepstrategy algorithm to learn a classifier from the onlypositive and unlabeled data. In step 1, he implemented Spytechnique to extract reliable negative (RN) examples. In step 2, he then used the labeled positive examples and the reliablenegative examples as training set to learn standard NaïveBayes classifier. The results showed the two-step algorithmsignificantly improves the F1

ISSN 2348 – 7968

score compared to the learningthat simply regards unlabeled examples as negative ones

Nitesh V. Chawla studied three issues, usually considered separately, concerning decision trees and imbalanced data sets quality of probabilistic estimates, pruning, and effect of preprocessingthe imbalanced data set by over or undersampling methods such that a fairly balancedtraining set is provided to the decision trees. They consider each issue independently and inconjunction with each other, highlighting thescenarios where one method might be preferred over another for learning decision treesfrom imbalanced data sets.

Sofia Visa, Anca Ralescu, proved the use of Fuzzy classifiers based on frequency distributionsproves to be robust for imbalanced data sets. Besides the imbalance factor, the class distribution of the training set is anotheractor that hinders the classification abilities of a given classifier.Since there is no guarantee that test data are represented well bythe training data (data available for learning), reduced varianceof classifiers output over different training class distributions isa very important feature of a classifier.The analysis of the results is based on the accuracy and theROC curves. The experimental results reported here show thatfuzzy classifiers are less variant to the class distribution and lesssensitive to the imbalance factor than decision trees.

CONCLUSION

Data mining sometimes called data or knowledge discovery. It is the process of analyzing data from different perspectives and summarizing it into useful information. It uses mathematical analysis to derive patterns and trends that exist in data. Typically, these patterns cannot be discovered by traditional data exploration because the relationships are too complex or because there is too much data.Imbalanced data sets are a special case for classification problem where the class distribution is not uniform among the classes. Typically, they are composed by two classes. The majority negative class and the minority positive class. Many difficult machine learning real-world problems are characterized by imbalanced learning data, where at least one class is under-represented relative to other. It is concluded that the imbalanced data set problems can be solved using different classification techniques proposed by different authors

REFERENCES

1. Breiman L, (2001), “Random forest”. Machine Learning, 45, 5–32.
2. Nitesh V. Chawlaet. al. (2002). "SMOTE: Synthetic Minority Over-sampling Technique". Journal of Artificial Intelligence Research . Vol.16, pp.321-
3. Chawla N.V. (2003) “C4.5 and Imbalanced Data Sets: Investigating the Effect of Sampling Method,Probabilistic Estimate, and Decision Tree Structure”, Proc. Int'l Conf. Machine Learning, Workshop Learning from Imbalanced Data Sets II.
4. Chen, C., Liaw, A., Breiman, L. (2004) “Using Random Forest to Learn Imbalanced Data”, Tech. Rep. 666,University of California, Berkeley.
5. Nitesh V. Chawla, Nathalie Japkowicz, AleksanderKotcz: Editorial: special issue on learning from imbalanced data sets. SIGKDD Explorations 6(1): 1-6 (2004)
6. Han, H., Wang, W,Y., Mao, B, H., (2005), “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning” ; Proc. Int'l Conf. Intelligent Computing, 878-887.
7. Visa, S., and Ralescu, A. 2005. The effect of imbalanceddata class distribution on fuzzy classifiers
8. David A. Cieslak, Nitesh V. Chawla, ¯Learning DecisionTrees for Unbalanced Data., 2008.
9. Wei Liu, Sanjay Chawla, David A. Cieslak, Nitesh V. Chawla, A Robust Decision Tree Algorithm for Imbalanced Data Sets., 2010
10. Alberto Fernández, Victoria Lopez, Mikel Galar, María José del Jesus, Francisco Herrera(2013) Analysing the classication of imbalanced data-sets with multiple classes:Binarization techniques and ad-hoc approaches.

ISSN 2348 – 7968

11. Yetian chen Learning Classifiers from Imbalanced, Only Positive and Unlabeled Data Sets

12. Kittipong Chomboon , Kittisak Kerdprasop, and Nittaya Kerdprasop, Rare Class Discovery Techniques for Highly Imbalanced Data.,2013

13. Vladimir Nikulin1, Geoffrey J. McLachlan1, and Shu Kay Ng , Ensemble Approach for the Classification of
Imbalanced Data.