On the Classification of Imbalanced Datasets

[1]C.V. KrishnaVeni, [2]T. Sobha Rani
1,2Dept. of Computers and Information Sciences University of Hyderabad, India

Abstract
The Classification of Imbalanced Data Sets have received considerable attention in recent research. In this paper, we present an overview of the problem of imbalanced data sets, explain the most commonly used techniques such as sampling and cost sensitive learning, present some evaluation metrics used on imbalanced data sets, quote some interesting points drawn from various popular and latest research papers related to imbalanced classification problem. This paper does not mention all the available research solutions, but try to give a clear picture of imbalanced data set classification problem and present a brief review of existing solutions on this problem. Here, we consider binary classification problem on imbalanced data sets.

Keywords
Imbalanced data sets, Sampling, Cost Sensitive learning

I. Introduction

Data sets with imbalanced class distributions are quite common in many real applications. For example, an automated inspection system that monitors products that come off a manufacturing assembly line may find that the number of defective products is significantly fewer than that of non-defective products. Similarly, in credit card fraud detection, fraudulent transactions are outnumbered by legitimate transactions. In both of these examples, there is a disproportionate number of instances that belong to different classes. The degree of imbalance varies from one application to another for example, a manufacturing plant operating under the six sigma principle may discover four defects in a million products shipped to their customers, while the amount of credit card fraud may be of the order of 1 in 100. Despite their infrequent occurrences, a correct classification of the rare class in these applications often has greater value than a correct classification of the majority class [1].

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 [2,3].

II. Sampling Strategies

To handle the problem of imbalanced data, sampling approaches are applied on the data to change the class distribution of data and make it balanced. The Sampling approaches are mainly divided into two categories: Undersampling and Oversampling.

A. Undersampling

This method removes examples from the majority class to make the data set balanced. This method is suitable for large scale applications, where the number of majority class examples is very large and reducing the training samples lessens the training time and storage [8]. The drawback of undersampling method is that it discards potentially useful information that could be important for classifers [3,8].

Undersampling methods are divided into random and informative. Random Undersampling is simple, it randomly eliminates examples from the majority class till the data set gets balanced. Informative Undersampling method selects only the required majority class examples based on a pre-specified selection criterion to make the data set balanced. Informative Undersampling can be passive or active. Passive selection methods are proposed as preprocessing technique for selecting informative samples for a classifer. In Active selection methods, informative samples are queried during the construction process of the classifier [6].

In Ref. [28], Kubat and Matwin presented One Sided Selection(OSS) which is an undersampling method. OSS only removes examples from the majority class while leaving the examples from the minority class untouched. They divided majority(negative) class examples into four groups like class-label noise, Borderline examples, redundant and safe examples. The OSS algorithm works as follows: first the number of redundant negatives is reduced by creating the subset C, consistent with the training set. By definition, C, a subset of S is consistent with S, if when used by the 1-NN rule, it correctly classifies examples in S. Then the system removes those negative examples that participate at Tomek links. Borderline examples and examples suffering from the class-label noise participate at Tomek links.So, they are eliminated.

Ref. [29] describes an application of a simple kNN approach to an imbalanced data classification problem. They empirically studied the effects of undersampling on the k nearest neighbour kNN approach and five different methods of choosing negative training examples, Random Selection, selection of NearMiss examples which is done in three ways NearMiss-1, NearMiss-2, NearMiss-3 and selection of most distant examples. The NearMiss-1 selects negative examples that are close to some of the positive examples, they select negative examples whose average distances to three closest positive examples are the smallest. The NearMiss-2 selects negative examples that are close to all positive examples. In this method, examples are selected based on their average distances to three farthest positive examples. In NearMiss-3, given number of closest negative examples for each positive example are chosen. In Selection of most distant negative examples, the negative examples whose average distances to the three closest positive examples are the farthest are chosen. They found through experiments that both kNN and C5.0 are sensitive to the percentage of negative examples selected and among the five negative example selection methods random and NearMiss-2 methods performed the best.

Ref. [30] proposed a majority filter-based minority prediction(MFMP) approach for imbalanced data sets, the goal of this approach is to achieve good prediction over minority class by avoiding unnecessary information loss from the majority class. The MFMP adopts an unsupervised learning technique for selecting samples for supervised learning.

The approach works in two steps: in the first step, minority samples are clustered and majority class samples that are out of minority classification regions are identified. This improves minority prediction rate, in the second step, majority samples are randomly selected in individual clusters and this enhances majority prediction rate. Experimentally, they studied the behaviour of MFMP approach and found that it outperforms the traditional random under-sampling approach. In addition to [28-30], several other undersampling approaches are available in the literature.

B. Oversampling

Oversampling is a sampling approach which balances the data set by replicating the examples of minority class. It is also called upsampling. The advantage of this method is that there is no loss of data as in undersampling technique. The disadvantage of this technique is it may lead to overfitting and can introduce an additional computational cost if the data set is already fairly large but imbalanced [2,3].

Like Undersampling, oversampling is also divided into two types. Random Oversampling and Informative Oversampling. Random Oversampling is the method which balances the class distribution by replicating the randomly chosen minority class examples. Informative Oversampling method synthetically generates minority class examples based on a pre-specified criterion [6]. There are number of Oversampling methods available in the literature like SMOTE [33], Borderline SMOTE [34], OSSLDDD-SMOTE [27] etc.

III. Cost Sensitive Learning

Cost Sensitive Learning (CSL) is another commonly used approach to handle the classification problem of imbalanced data sets. It is considered to be an algorithmic level solution. To make the concept of cost sensitive learning clear, let us take an example given in [11]. Consider an example of cancer disease from medical diagnosis. If the patient has cancer, then the tests show positive , it is treated as positive class. If the patient has not cancer then the test result gives negative, it is treated as negative class. If the results are correct, showing positive for cancer patient and negative for non-cancer patient, there is no problem that means it is not misclassified. If the penalty is to be assigned to the misclassification True Positive (TP), True Negative (TN) i.e., for correct classification there is no penalty. But giving wrong results, negative for cancer patient, False Negative (FN), positive for non-cancer patient,is misclassification. It is to be noted that these both misclassifications cannot be treated same. False Negative may results in patient's death due to the delaly in finding and taking treatment for cancer which is more serious than False Positive where the patient may go for another test to confirm and/or may take more care about his health. The cost of FN will be more than the cost of FP and the costs of TP and TN is zero [8].

Let us consider another example of fraud detection to make the concept much more clear and which helps to relate the cost factor with imbalance problem. In the case of fraud detection, suppose a person carrying a bomb is an example of positive class and a person without bomb is negative class example. Here the classification problem is to identify to which class a given person belongs to. There is no cost associated with correct classifications as identifying a person with bomb as positives and person without bomb as negative. But identifying a person with bomb as negative which is False Negative is much dangerous and cause more harm than identifying a person without bomb as positive which is False Positive because the innocent person can be left after searching him and finding that he does not have a bomb. This FP classification does not cause much, in this case it only troubles that person for sometime. It is clear that misclassification costs are not equal. They are unequal and the impact or magnitude of the cost depends upon the application and situation.

In the fraud detection classification problem, given a dataset, the number of persons carrying a bomb is usually far less than the number of persons not carrying a bomb. Here, the two classes data distribution is unequal which says that it is imbalanced data set. By taking into consideration during classifier building, the problem of classification of imbalanced data sets can be handled. The type of learning algorithm which takes misclassification cost into consideration is called Cost Sensitive Learning. It produces the classifier with minimum total cost. The advantage of this method is here no data is replicated or eliminated [12].

Let C(i,j), denote the cost of predicting an example of class i as class j. For a binary classification, misclassification costs can be presented by using cost matrix. Corresponding to a confusion matrix is cost matrix. The cost matrix provides the costs associated with the four outcomes of the confusion matrix [13].

		Predicted	
		Positive	Negative
Actual	Positive	0	C(FN)
	Negative	C(FP)	0

Fig. 1: Cost Matrix

Here i represents positive(minority) class and j represents negative(majority) class. C(i,i)=C(j,j)=0. It means, no cost(penalty) is associated with True Positives and True Negatives. C(i,j)=C(FN), C(j,i)=C(FP). Costs are assumed, they can be constant or example dependent[14]. The goal of cost sensitive learning method is to choose a classifier with lowest total cost. Total cost= C(FN)xFN + C(FP)xFP, where FN is the number of positive examples wrongly predicted as negative class, FP is the number of negative examples wrongly predicted as positive class. C(FN) and C(FP) corresponds to the costs associated with False Negative and False Positive respectively, C(FN)>C(FP).

There are many ways to implement cost sensitive learning, in [9],it is categorized into three, the first class of techniques apply misclassification costs to the data set as a form of data space weighting, the second class applies cost-minimizing techniques to the combination schemes of ensemble methods, and the last class of techniques incorporates cost sensitive features directly into classification paradigms to essentially fit the cost sensitive framework into these classifiers. There are various ways to incorporate cost into classifiers available in the literature [eg., 31,32 etc.,] to handle imbalanced data sets efficiently.

IV. Performance Metrics

The performance of traditional classification algorithms is evaluated by the metric accuracy which is defined as the percentage of examples that are correctly classified. This is not suitable when dealing with imbalanced data sets as the minority class has less

effect on accuracy than that of minority class. There there is need for other metrics. All these metrics use confusion matrix. The form of confusion matrix is given below.

Table 2: Confusion Matrix

	Predicted		
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

The difference between confusion matrix and cost matrix is that the former contains the number of TP, FN, FP, TN in the cells whereas the cost matrix provides information about misclassification costs namely FN, FP. In cost matrix diagonal elements are set to zero.

Accuracy=TP+TN/TP+FN+FP+TN.

Error Rate = 1- Accuracy = FP+FN/TP+FN+FP+TN

We present here, some other widely used metrics:

Precision is the percentage of positive predictions that are actually correct. Precision=TP/TP+FP.

Recall is the percentage of true positive examples that are correctly predicted, sensitivity is the other name for recall.

Recall=TP/TP+FN

F-measure is defined as the harmonic mean of recall and precision [15].

F-measure =2xRecallxPrecision/Recall+Precision.

Precision is a measure of exactness whereas Recall is a measure of completeness [9].

Specificity is the accuracy on negative examples.

Specificity = TN/TN+FP.

Other metrics Receiver Operating Characteristics (ROC) curve, area under ROC(AUC) are widely used evaluation metrics when dealing with imbalanced data sets. Receiver Operating Characteristics(ROC) curve is formed by plotting TP-rate over FP-rate and any point in ROC space corresponds to the performance of a single classifier on a given distribution. It gives a visual indication of a classifier is superior to another classifier over a wide range of operating points. The area under ROC curve(AUC) summarize the performance of classifier into a single metric. The larger the area under the ROC, the better the performance of the classifier.

V. Some Points of Interest

Oversampling and Undersampling are effective methods of dealing with the problem of imbalanced data sets classification. However, undersampling(downsizing) approach works better than the oversampling method on large domains [16]. Oversampling appears to the the best for small data sets [13]. Ref. [23] conducted experiments and shown that oversampling clearly appears as better than undersampling for local classifiers whereas some undersampling strategies outperform oversampling when employing classifiers with global learning. The classifiers produced by sampling and by using cost sensitive matrix performance is found to be similar in [17]. By focusing exclusively on data sets with more than 10,000 examples Weiss, Kate, Zabar in [13], found that cost sensitive learning algorithm consistently outperforms the sampling methods. It should be noted that their focus was on using the cost information to improve the performance on the minority class. Ref. [18] presents an empirical study which disclosed that when the misclassification costs are equal, cost sensitive classifiers favour natural class distribution although it may be imbalanced while when misclassification costs are unequal, a balanced class distribution is more favourable.

In Ref. [19], Weiss and Provost discusses the effect of class distribution on tree induction. Here, he says that for any fixed class distribution, increasing the size of the training set always leads to improved classifier performance. The choice of class distribution may become less important as the training set size grows. But, in practice, the number of training examples we use for learning must be limited due to the costs associated with procuring, preprocessing and storing the training samples and the computational costs associated with learning from them. In such situation, if only n training examples can be selected, in what proportion should the classes be represented is the problem. They said that the naturally occurring class distribution is shown to generally perform well when classifier performance is evaluated using undifferentiated error rate. When the are under the ROC is used to evaluate classifier performance, a balanced distribution is shown to perform well. Since, neither of these choices for class distribution always generates the best performing classifier, a budget-sensitive progressive sampling algorithm is introduced for selecting training examples based on the class associated with each example.

As mentioned in the previous section cost sensitive learning handle imbalanced classification problem. Let us discuss here, how to incorporate cost into decision tree classification algorithm which is one of the most widely used and simple classifier. Cost can be incorporated into it in various ways [1,8,9,14,20,21]. First way is cost can be applied to adjust the decision threshold, second way is cost can be used in splitting attribute selection during decision tree coinstruction and the other way is cost sensitive pruning schemes can be applied to the tree. Ref. [21] propose a method for building and testing decision trees that minimizes total sum of the misclassification and test costs. The algorithm used by them chooses an splitting attriute that minimizes the total cost, the sum of the test cost and the misclassification cost rather than choosing an attribute that minimizes the entropy. Information gain, Gini measure are considered to be skew sensitive [22]. In Ref. [23] a new decision tree algorithm called Class Confidence Proportion Decision Tree (CCPDT) is proposed which is robust and insensitive to size of classes and generates rules which are statistically significant. Ref. [24] analytically and empirically demonstrate the strong skew insensitivity of Hellinger Distance and its advantages over popular alternative metrics. They arrived at a conclusion that for imbalanced data it is sufficient to use Hellinger trees with bagging without any sampling methods. Ref. [26] uses different operators of Genetic algorithms for oversampling to enlarge the ratio of positive samples and then apply clustering to the oversampled training data set as a data clearning method for both classes, removing the redundant or noisy samples. They used AUC as evaluation metric and found that their algorithm performed better. .

VI. Conclusion

This paper provides an overview of the classification of imbalanced data sets. We just reviewed most commonly used sampling and cost sensitive strategies here. In addition to the above specified techniques, there are a number of research solutions available to handle classification problem on imbalance data sets like ensemble methods, one-class learners etc., In section V we presented some interested points from popular and latest papers. There also exists several research

solutions for multiclass classification of imbalanced data sets while this paper dealt only binary classification problem on imbalanced data sets.

References

[1] Pang-Ning, Tan, Vipin Kumar, Michael Steinbach, "Introduction to Data Mining", Pearson.

[2] Guo, Yin, Dong, Yang, Zhou, "On the Class Imbalance Problem", 2008.

[3] Kotsiantis, kanellopoulos, pintelas, "Handling Imbalanced Datasets : A review ", 2006.

[4] Sofia Visa, AncaRalescu, "Issues in Mining Imbalanced Data Sets – A Review Paper", 2005.

[5] Garcia, sanchez, mollineda, alejo, sotoca, "The class imbalance problem in pattern classification and learning ", 2006.

[6] T. Maruthi Padmaja, "Sampling Approaches for Unbalanced Data Classification Problem", thesis-2011.

[7] Shuo Wang, thesis proposal, "Class Imbalance Learning", 2008.

[8] Nguyen, Bouzerdoum, PLhung, "Learning Pattern Classification Tasks with Imbalanced data sets", 2009.

[9] Haibo He, Edwardo A. Garcia, " Learning from Imbalanced Data", 2009.

[10] Kotsiantis S, Pintelas P, "Mixture of experts agents for handling imbalanced data sets",2003.

[11] Charles X. Ling, Victor S. Sheng, "Cost sensitive Learning and the Class Imbalance Problem", 2008.

[12] Kate McCarty, BibiZabar, Gary Weiss, "Does Cost-Sensitive Learning Beat Sampling for Classifying Rare Classes ?", 2005.

[13] Gary M. Weiss, Kate McCarthy, BibiZabar, " Cost-Sensitive Learning Vs Sampling. Which is Best for Handling Unbalanced classes with Unequal Error Costs?", 2005.

[14] Charles Elkan, "The Foundations of Cost Sensitive Learning ", 2001.

[15] Fawcett T, " An Introduction to ROC Analysis", 2006.

[16] Nathalie Japkowicz, "Learning from Imbalanced Data Sets : A Comparison of Various Strategies", 2000.

[17] Marcus A. Maloof, " Learning when Data Sets are Imbalanced and When Costs are Unequal and Unknown", 2003.

[18] Xu-Ying Liu, Zhi-Hua Zhou, "The Influence of Class Imbalance on Cost-Sensitives Learning : An Empirical Study ", 2006.

[19] Gary M. Weiss, Foster Provost, "Learning When Training Data are Costly : The Effect of Class Distribution on Tree Induction ", 2003.

[20] Chris Drummond, Robert C. Holte, " Exploiting the Cost(In) sensitivity of Decision Tree Splitting Criteria", 2000.

[21] Charles X. Ling, Qiang Yang, Jianning Wang, Shichao Zhang, "Decision Trees with Minimal Costs", 2004.

[22] David A. Cieslak, Nitesh V. Chawla, "Learning Decision Trees for Unbalanced Data", 2008.

[23] Wei Liu, Sanjay Chawla, David A. Cieslak, Nitesh V. Chawla, " A Robust Decision Tree Algorithm for Imbalanced Data Sets", 2010.

[24] David A. Cieslak, T. Ryan Hoens, Nitesh V. Chawla, W. Philip Kegelmeyer, "Hellinger distance decision trees are robust and skew-insensitive", 2011.

[25] Vicente Garcia, Jose Salvador Sanchez, Ramon A. Mollineda, " Exploring the performance of Resampling Strategies for the Class Imbalance Problem",2010.

[26] Satyam Maheshwari, Prof. Jitendra Agarwal, Dr. Sanjeev Sharma, " A New Approach for Classification of Highly Imbalanced Datasets Using Evolutionary Algorithms", 2011.

[27] ZHAI Yun, MA Nan, RUAN Da, AN Bing, " An Effective Oversampling Method for Imbalanced Data Sets Classification ", 2011.

[28] Miroslav Kubat, Stan Matwin, " Addressing the Curse of Imbalanced Training Sets : One Sided Selection ", 1997.

[29] Jianping Zhang, Inderjeet Mani, " kNN Approach to Unbalanced Data Distributions : A Case Study involving Information Extraction", 2003.

[30] T. Maruthi Padmaja, P. Radha Krishna, Raju S. Bapi, " Majority Filter- based Minority Prediction(MFMP) : An Approach for Unbalanced Data Sets", 2008.

[31] Bianca Zadrozny, John Langford, Naoki Abe, "Cost-sensitive Learning by Cost proportionate Example Weighting", 2003.

[32] Y. Sun, M.S. Kamel, A.K.C. Wong, Y. Wang, "Cost-sensitive Boosting for classification of Imbalanced Data", 2007.

[33] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer, "SMOTE : Synthetic Minority Oversampling Technique", 2002.

[34] H. Han, W.Y. Wang, B.H. Mao, "Borderline-SMOTE : A New Over-Sampling Method in Imbalanced Data Sets Learning", 2005.