Learning From Imbalanced Data

Lincy Mathews
M. S. Ramaiah Institute of Technology, India

Seetha Hari
Vellore Institute of Technology, India

INTRODUCTION

Pattern Identification on various domains have become one of the most researched fields. Accuracy of all traditional and standard classifiers is highly proportional to the completeness or quality of the training data. Completeness is bound by various parameters such as noise, highly representative samples of the real world population, availability of training data, dimensionality etc.

Another very pressing and domineering issue identified in real world data sets is that the data is well-dominated by typical occurring examples but with only a few rare or unusual occurrences. This distribution among classes make the real world data inherently imbalanced in many domains like medicine, finance, marketing, web, fault detection, anomaly detection etc.

This chapter aims to highlight the existence of imbalance in all real world data and the need to focus on the inherent characteristics present in imbalanced data that can degrade the performance of classifiers. It provides an overview of the existing effective methods and solutions implemented towards the significant problems of imbalanced data for improvement in the performance of standard classifiers. Efficient metrics for evaluating the performance of imbalanced learning models followed by future directions for research is been highlighted.

BACKGROUND

The field of data mining has identified learning from data that suffer from imbalance distribution as one of the top problems of today (Yang & Wu, 2006). However, a widely accepted issue is that the traditional learning algorithms assume a balanced distribution among the classes. It does not address nor recognize the presence of imbalance in the data. Data imbalance is evident when the number of instances representing the class of concern is much lesser than other classes.

To cite an example, the 1999 KDD Cup data set (UCI machine repository) is considered. The information collected by a simulated LAN environment consists of normal traffic with a relatively small number of intrusion attempts. The original data set consists of 23 classes, of which one of the classes belonged to normal traffic. When the data set was grouped down to a total of 2 classes, `normal' and `attack', the KDD data had 972,781 minority `attack' class examples and 3,925,650 majority `normal' class examples, which is approximately 80.14% majority examples. The training data thus will have only very few samples from the 'attack' class, due to which the classifier will be biased to the normal cases. This under representation of the interested class is evident in many applications such as intrusion detection, pollution detection, fault monitoring, biomedical, bioinformatics and remote sensing.

The under-represented class and well-represented class are known as the positive class (denoted by +1) and negative class (denoted by -1) respectively. As the class of interest indicates a positive case and is rarer by nature, it represents the minority data. The research community addresses the other well-defined classes as majority

DOI: 10.4018/978-1-5225-2255-3.ch159

class. The ratio between the instances of majority versus minority is termed as imbalance ratio.

The skew distribution present in the training data, leads to the bias by most classifiers. Studies have however shown that the base classifiers perform well when presented with balanced data than with imbalanced data (Weiss & Provost, 2001). This justifies the need for learning models that can address the challenges posed by imbalanced data.

CHARACTERISTICS OF IMBALANCED DATA

The imbalance ratio between the majority and minority instances need not necessarily affect the performance of classifiers if the degree of imbalance is moderate (Chen & Wasikowski, 2008). The inherent characteristics within minority data however; can cause degrade in performance by the learning models. Two basic categorization of minority instances exist; Safe and unsafe minority instances. Safe minority instances are instances where the misclassification is minimal by the base learners. These instances exist much away from the borderline of majority instances. Unsafe minority instances are so called because the misclassifications occur highly with these kinds of minority instances.

The causes of unsafe instances in imbalanced data sets are contributed by four significant occurrences. They are the presence of small disjuncts, lack of density in the sample space, noisy data and data shift. Addressing these issues alone can sometimes bring a positive effect on the accuracy of the classifier without having to address the imbalance factor (Japkowicz N, 2003).

Small disjuncts exist when there is a small cluster of similar instances amidst cluster of majority or minority instances. However, in case of imbalanced datasets, the presence of sub concepts (disjuncts) in the majority class will be rare as they are represented well. The occurrence of small disjuncts is frequent in minority class. The presence of small sub concepts can undermine the performance of the classifier. As the data space; is very limited, this can lead to generalization and over fitting by the learning models. Weiss (2003) had identified the presence of disjuncts in imbalanced datasets as noise or sub concepts. Synthetic (artificial) samples were generated for these small disjuncts to overcome the lack of representation.

Lack of density is the existence of small sample size. Lack of samples coupled with high dimensional data becomes an even more difficult issue in imbalanced data. This creates rules that are too specific and can lead to over fitting. In domains such as face recognition, gene expression analysis etc; feature reduction method such as the Principal Component Analysis were used to address the high dimensional problem with respect to sample size (Yang et al., 2008). One sided or two sided feature selection technique addressed this problem (Alibeigi, 2012).

Presence of noise or outliers creates a negative impact on the performance of the classifier. As outliers occur much away from minority instances, it is most likely treated as noise. However, it might represent the rarest of cases that should be recognized and classified. There are cases of overlapping instances; presence of similar instances in majority and minority instances. These samples seemingly are the hardest to classify. The authors (Garcıa et al., 2008) proposed two different frameworks for the k-Nearest Neighbor (k-NN) classifier. They focused on the ratio between the imbalance in the overlapping region versus the overall imbalance ratio. Khoshgoftaar et al. (2011) analyzed the relation between noise and imbalanced data. Bagging and boosting techniques were applied after implementing noise reduction techniques.

The data shift problem is realized when the training data distributions differs from the distribution of the test data (Alaiz & Japkowicz, 2008). The data shit phenomenon is mostly due to sample selection process. Stratified cross validation is used to measure performance thus avoiding sudden drift in performance. The other cause could be due to a very high degree of imbalance. Presence of data shift can highly affect the singular classification