Classification of Imbalanced Classes

Chevon De Souza,
Student (M.E. VLSI & EMBEDDEDSYSTEM),
Sinhgad Institute of Technology and Science,
Narhe, Pune

ABSTRACT

A frequent situation encountered in classification problems is that of imbalanced data arising out of selection of an imbalanced data set where disproportionately high number of samples from one class will result in a classifier that is biased towards this majority class. A classification algorithm based on such data can cause serious negative effects on classification performance of machine learning algorithms.

In this review paper, after the careful consideration of the numerous algorithms available and analysing their respective advantages and disadvantages, a conclusion is presented to create a more efficient system using the advantages of the algorithms present and minimizing their disadvantages.

Keywords: Imbalanced classes, machine learning algorithms, classification

INTRODUCTION:

Technology has grown at an alarming rate and with the growth of technology comes the tedious task of decision making. Most systems require the analysis and interpretation of raw data. The problem occurring from the rise of imbalanced data/classes is a relatively new hurdle. Though a number of learning algorithms have proven to be useful for class distribution skews, it does not ensure that the system is a hundred percent efficient.

In order to understand the concept more clearly, a simple analysis can be made. For example, if a system is considered to be 98 percent accurate and 98 percent of data classified, falls under one category, then we can truly say that the system is inefficient. The efficiency of most systems depends on the outcome of the classification of the total raw data. Sometimes however, when most of the data fits in one class and the other class is got a miniscule data set, the minority class is completely ignored and the efficiency of the system is solely based on the result of the majority class. In some cases, we need the classification of the minority class, as that weighs more importance. This review paper gives a gist of some of the techniques used for the classification of imbalanced data/classes, their algorithm comparisons, advantages and disadvantages.

I. SURVEY OF WORK DONE:

A. Need for Classification:

Raw data is present in different forms. The biggest challenge is to classify this data efficiently. This is become exceedingly tough due to skewed data as it causes imbalanced classes. Imbalanced classes tend to alter the decision of any system when its final efficiency is based on the accuracy of classification. The majority class is always in the forefront of these decisions which hampers the results if the minority class is to be analyzed. The different challenges arising from these classifications are aimed at the sampling size. A clear idea of the challenges faced during classification of raw data is given below.

B. Challenges in Achieving Classification:

During classification, a number of problems arise with the variations in samples in the data set. To overcome these difficulties is a herculean task. Some of the challenges during classification is as discussed below.

1. Imbalanced Classes: Imbalanced classes can be observed when there is a considerable difference in the ratio of the majority class to the minority class. As stated in reference [7], a balanced data set yields better results. This can also be sited in reference [11] where the EEG signals having imbalanced data sets were creating difficulties in identifying the patients EEG analysis. Due to inherent complex characteristics of imbalanced data sets, learning from such data requires new understandings, principles, algorithms and tools to
transform vast amount of raw data efficiently into information and knowledge representation [6]. Even as described in [8], class imbalance is a ubiquitous problem in supervised learning and has gained wide-scale attention in literature. So the foremost importance should be towards solving the class imbalanced problems.

2. Small Sampling Size: Sample size is important to determine a good classification model. If the size of the sample is small then it helps the classification model only if the data set is large [7]. Sampling methods consider the class skew and properties of the dataset as a whole [8].

3. Class Separability: If there is no overlapping between classes then without class distribution, any simple classifier could learn an appropriate classification [4]. The class imbalance distribution by itself does not seem to be a problem but when aligned to highly overlapped classes, it can significantly decrease the number of minority classes [7].

4. Small Disjuncts: The size of a disjunct can be defined by the number of training examples that it correctly classifies. Therefore small disjuncts are those disjuncts that classify few training examples. They have a higher error rate than that of the large disjuncts and are responsible for all the classification errors. [5]

C. Data Processing Methods:

Different sampling techniques give varied results even though the original dataset remains the same. Some of the data processing methods using sampling are given below.

1. Undersampling: Undersampling removes data from the original data set [6]. This means that it helps reduce the size of the non-target classes. This process is continued till the non-targeted class size is close to the targeted class. A better way to carry out this process is by having a selected representative data from each set to go under excessive learning. This technique is more efficient as compared to oversampling [10].

2. Oversampling: Oversampling appends data to the original data set [6]. This increases the size of the targeted class. It uses duplication to increase the size of the target class [10].

3. SMOTE (Synthetic Minority Oversampling Technique): SMOTE is used in a number of applications and it has shown a great deal of success too. An artificial data set is created keeping in consideration the minority class. However the disadvantages of SMOTE include variation and generalization. In order to overcome this disadvantage, SMOTE is combined with oversampling and it synthetically generates data of the target class. The kNN (k-Nearest Neighbour) method is used by SMOTE. If the boosting technique is further added, then it makes SMOTE the best data processing method as compared to oversampling and undersampling [10].

D. Imbalanced Data Classification Techniques:

1. Decision Tree: In order to classify the data using decision trees, all the original data is first stored in the root of the tree. Then by using different criteria, the branches are created. This data goes through excessive training to ensure that the new, unlabelled data can be determined by matching its feature values against the ancestors of that respective decision tree [10]. So a decision tree can be considered as a threshold network having a hidden layer of transitional logic units followed by one logical function per class. In order to test and evaluate the performance of decision trees, experiments were performed and a conclusion was made that decision trees are better and more accurate as compared to single classifiers that are trained on their original data set [13]. The weight precision however would depend on the application of the system at hand.

2. MLPNN (Multi-Layer Probabilistic Neural Network): The MLPNNs are the most commonly used neural-network architectures since they have the ability to learn and generalize. They have smaller training-set requirements and are easy to implement. One major property of these networks is their ability to find nonlinear surfaces separating the underlying patterns, which is generally considered as an improvement on conventional method [15]. However, it is less accurate with respect to PNN even though the classification of new data is faster.

3. PNN (Probabilistic Neural Network): PNN is a feedforward neural network which was derived from the Bayesian estimation theory for decision making with non parametric estimator [1]. PNN is faster than MLPNN and are more accurate [1]. The disadvantages of MLPNN are overcome by using PNN. PNN however requires a representative training set and large memory. Also the processing
of new cases takes more time.

4. **SVM (Support Vector Machine):** SVM is a supervised learning model. SVM can be classified into four types based on the technique used: k-Nearest Neighbour, Decision Tree, Linear and Non-Linear classification. Of these the most efficient is the linear classification using SVM. SVM is used for data mining and pattern recognition as well as classification of imbalanced data [2][3][5][8][11][14].

The linear SVM classification is proven to be the best method for classification of imbalanced data [11]. The reference [11] also give a clear comparison of MLPNN, PNN and SVM for classification of EEG signals and SVM showed 99.2% classification accuracy as compared to PNN and MLPNN which have a classification accuracy of 98.05% and 93.63% respectively. The margin of the hyper-plane plays an important role to ensure that the data can be classified as accurately as possible. Also various kernel functions have also been developed to address specific needs of different domain/data [10].

However SVM also poses some limitations. In accordance to SVM, the biggest limitation of the support vector lies in the choice of kernel, the other limitations include size and speed in both testing and training.

5. **Bagging Ensemble Variation (BEV):** Data analysis is a fundamental part of most systems and the classification plays an important role [14]. In order to obtain different data sets, random sampling with a replacement from the original data set is used. This is called the bagging ensemble method and it may include similar data for some sets in order to check testability. Learning algorithms are used for this purpose [13]. After a classifier is learned from the N training data sets, majority voting is done to determine class label for each test data which is further classified into the new data [10]. Bagging is the simplest and most intuitive to implement. However it is most efficient when the data set has a limited size [12].

II. **Discussion and conclusion:**

After comparing all the advantages and disadvantages of the different classification techniques for imbalanced classes, it is seen that though the efficiency is improved from one proposed system to another, the system cannot be a hundred percent efficient.

Keeping this analysis in mind, a system can be created using the advantages of the best classifiers and the disadvantages of the other classifiers in order to create a better classifier. This can be an attempt to design, test and validate an effective classifier ensemble algorithm that can improve the classification accuracy by resolving the class imbalance problems faced by various applications having skewed distribution of data. The proposed system is shown in figure 1.

![Figure1: Proposed System](image_url)

Here the database will hold up to a thousand signals with their respective features. This can be expanded as per the requirement of other systems. The proposed system is developed on FPGA hardware and the classification performance of the system is compared based on accuracy. This includes classifiers for imbalanced data classification viz decision tree and SVM which will prove to be most efficient and a voting system to ensure that an imbalanced class is avoided.

**REFERENCES**


