Optimising Prediction in Overlapping and Non-Overlapping Regions

Sumana B.V., Vijaya College Jayanagar, Bengaluru, India

Punithavalli M., Bharathiar University, Coimbatore, India

ABSTRACT

Researchers working on real world classification data have identified that a combination of class overlap with class imbalance and high dimensional data is a crucial problem and are important factors for degrading performance of the classifier. Hence, it has received significant attention in recent years. Misclassification often occurs in the overlapped region as there is no clear distinction between the class boundaries and the presence of high dimensional data with an imbalanced proportion poses an additional challenge. Only a few studies have ever been attempted to address all these issues simultaneously; therefore; a model is proposed which initially divides the data space into overlapped and non-overlapped regions using a K-means algorithm, then the classifier is allowed to learn from two data space regions separately and finally, the results are combined. The experiment is conducted using the Heart dataset selected from the Keel repository and results prove that the proposed model improves the efficiency of the classifier based on accuracy, kappa, precision, recall, f-measure, FNR, FPR, and time.

KEYWORDS

Class Imbalance, Class Overlap, K-Means, Overlapped and Non-Overlapped Data Space

INTRODUCTION

The real-time data accumulated in the society due to day-to-day activities like credit card transactions, patient's health record, failure in a manufacturing unit, medical diagnosis, detection of oil spills, text classification etc., are always overlapped and class imbalanced in nature (Sumana, 2016). Usually in an imbalanced dataset the classifier misclassifies minority class instances because they get biased by the majority class instances which are highly represented hence classifier shows degradation performance. It frequently occurs in overlapping region as high dimensional data is the main cause for class overlap. As such class imbalance is not a crucial problem but combination of class imbalance with class overlap including high dimensional data is a crucial problem and is the cause for the degrading performance of the classifier (Sumana, 2016).

The data is said to be imbalanced if classes in the data space are not represented in equal proportion. The class representing with higher number of instances is called majority class and the class representing with fewer number of instances is called minority class. Due to class imbalance nature of the dataset classification task becomes very difficult because the classifier gets biased towards the majority class as it does not get necessary information about the minority class to make an accurate prediction therefore show poor classification rates on minority class, because it treats the instances of the minority class as noise hence due to class imbalance nature there will be degradation in the performance of the classifiers. Therefore, a balanced dataset is necessary for building a good

DOI: 10.4018/IJNCR.2020010104

prediction model as most of the classifiers perform well when the number of instances of each class is approximately equal in proportion (Guo, 2016).

When samples from different classes have similar characteristics, they do not form separate clusters and are not linearly separated, instead few samples overlap in the data space known as overlapping samples. Class imbalance is not a crucial problem on itself, but combination of class overlap with class imbalance poses a new challenge and is the cause for the degradation performance of the classifier. Liu (2008) in his work stated that overlapping region contains data from more than one class and misclassification often occurs near the class boundaries where overlapping is present and Aida Ali (2015) suggested that high dimensionality with redundant or irrelevant features makes the classifier difficult to recognize the class boundaries and hence is one of the causes for class overlap.

Methods to Address Class Imbalance

Methods to overcome class imbalance can be classified into two categories, data level approach and algorithmic level approach. Data level approach modifies the data and balances it using sampling methods or synthetic data generation methods to overcome classifier getting biased towards majority class whereas in algorithmic level approach the classifier is modified to overcome the bias towards majority class objects.

Data Level Approach

Sampling methods are further divided into over sampling, under sampling and hybrid methods. Under sampling methods balances the proportion of the class distribution by randomly eliminating the samples of majority class retaining the minority class samples. Over sampling methods balances the proportion of the class distribution by randomly replicating the samples of the minority class from the existing samples retaining the majority class samples. Hybrid method is a combination of both over sampling and under sampling methods which balances the proportion of the class distribution by randomly eliminating the majority class samples and replicating the minority class samples.

The synthetic data generation method artificially generates data using bootstrapping or Knn to balance the class distribution example ROSE, ADASYN, SMOTE, MSMOTE, BORDERLINE SMOTE, SMOTE-TL and SMOTE-E, selective pre-processing of imbalanced data (SPIDER) etc.

ALGORITHM LEVEL APPROACH

Cost Based Approach

In Cost based approach instead of creating a balanced data it learns the imbalanced problem using cost matrices taking the misclassification costs into consideration. During the model construction a higher misclassification cost is assigned for minority class objects and classification is performed such that it has a lower cost. Let C(i,j) denote the cost of a test case predicted to be class i but actually it belongs to class j. In a two class problem c(+,-) signifies the cost of misclassifying a positive sample as the negative sample and c(-,+) denotes the cost of the contrary case. Cost sensitive learning methods takes advantage of the fact that it is more expensive to misclassify a true positive instance than a true negative instance that is C(+,-) > C(-,+). For a two-class problem a cost sensitive learning method assigns a greater cost to false negatives than to false positives hence resulting in a performance improvement with respect to the positive class (HE & Garcia, 2009).

Methods to Address Class Overlap

A huge range of solutions are provided in the literature to address class overlap that includes preprocessing of data before learning a classification model. It is a two-step process which initially identifies the overlap region in the data space and removes the overlap region using data cleaning algorithms, feature selection or cluster-based methods (He, 2009).