Chapter 63
Mitigating Data Imbalance Issues in Medical Image Analysis

Debapriya Banik
Jadavpur University, India

Debotosh Bhattacharjee
Jadavpur University, India

ABSTRACT

Medical images mostly suffer from data imbalance problems, which make the disease classification task very difficult. The imbalanced distribution of the data in medical datasets happens when a proportion of a specific type of disease in a dataset appears in a small section of the entire dataset. So analyzing medical datasets with imbalanced data is a significant challenge for the machine learning and deep learning community. A standard classification learning algorithm might be biased towards the majority class and ignore the importance of the minority class (class of interest), which generally leads to the wrong diagnosis of the patients. So, the data imbalance problem in the medical image dataset is of utmost importance for the early prediction of disease, specifically cancer. This chapter attempts to explore different problems concerning data imbalance in medical diagnosis. The authors have discussed different rebalancing strategies that offer guidelines for choosing appropriate optimal procedures to train the samples by a classifier for an efficient medical diagnosis.

INTRODUCTION

The data imbalance problem is prevalent in medical image analysis. The training of machine learning (ML) algorithm from an imbalanced medical data set is an inherently challenging task(Mena & Gonzalez, 2006). A classifier in ML's objective is to learn and predict the unseen output class of an unknown instance with good generalization capability. The mining of knowledge in a machine learning paradigm is accomplished by a set of $\mathfrak{D}$ input instances such as $\eta_1, \eta_2, \eta_3, \ldots, \eta_{\mathfrak{D}}$ described by k features

DOI: 10.4018/978-1-6684-7544-7.ch063

$\lambda_1, \lambda_2, \ldots, \lambda_k \in F$ whose intended output class labels $\mathcal{O}_j \in C$ $\{c_1, c_1, \ldots, c_m\}$. A mapping function $F^k \rightarrow C$, implies the learning algorithm which is known as a classifier(Galar, Fernandez, Barrenechea, Bustince, & Herrera, 2011). This is a general idea for how a supervised learning algorithm performs its task. The imbalanced distribution of the data in medical image datasets happens when a specific disease type in a dataset appears in a small section of the entire dataset(C. Zhang, 2019). Hence, analyzing medical data posed severe challenges in the classification of a disease. A standard ML classifier will be skewed against the majority class and underestimate the importance of the minority class because the minority class has a lesser number of instances compared to the majority class. However, the minority class is generally referred to as the class of interest(Napierala & Stefanowski, 2016) in medical image analysis. So, the minority class is of utmost importance for the early prediction of disease. This problem influences all supervised classification algorithms. A well-balanced medical image dataset is very crucial for designing a reliable and standard prediction model. Typically, real-world medical data, specifically cancer data, usually suffer from data imbalance, leading to the degradation of ML algorithms' generalization. These eventually degrade the efficiency and accuracy of the computer-aided early prediction of cancer. The biaseness of the medical data in healthcare domain due to individual diversity can cause missclassification which may affect early diagnosis of cancer and disease risk prediction(Zhao, Wong, & Tsui, 2018). However, the imbalanced class problem is generally ignored in Conventional Learning(CL) algorithms. Those algorithms give the same priority to both classes: the majority class and the minority class. However, when the majority class and the minority class are highly imbalanced, it is very challenging to build a good classifier using CL algorithms(Krawczyk, 2016). It is a significant concern in most medical datasets where patients at high-risk tend to be in the minority class, and so the cost in miss-classification of the minority classes is higher than that of the majority class. In Figure 1 a graphical representation of the distribution of majority class and the minority class is shown. The noisy data is a small part of the minority class, which significantly impacts the performance of the classifier(López, Fernández, García, Palade, & Herrera, 2013).

Cancer is a formidable disease. Recently, there is a high incidence and mortality rate due to cancer. Early diagnosis of the disease at a primary stage before metastasis and growth can save more lives. However, the survival rate drastically declines in its advanced stage. Due to the rise in artificial intelligence (AI) techniques for computer-aided early cancer prediction, various researchers worldwide have focused more on various factors affecting the learning algorithms due to imbalanced data. So analyzing medical datasets with imbalanced data is a significant challenge for the machine learning classifiers(Mena & Gonzalez, 2006). It is seen that the learning classifiers are biased towards the class having more samples (majority class), i.e., healthy data, with the perception that the dataset is well-balanced. However, it can be observed that the learning classifiers are inefficient in handling imbalanced data. The classifiers tend to classify the majority class false positively due to imbalanced data, which generally leads to the patients' wrong diagnosis. But an optimal classifier should accurately classify the disease, which typically belongs to the minority class, with a higher degree of accuracy(Ali, Shamsuddin, & Ralescu, 2013). So, the primary goal of a learning classifier in diagnosing a disease is to improve the accuracy of the minority class. Incorrect labeling of class labels(cancer/noncancer) in an imbalanced medical dataset is costly as a cancerous patient will be labeled as noncancerous and vice-versa, which is a significant matter of concern(J. Zhang, Chen, & Abid, 2019).