Over-sampling Algorithm Based on in imbalanced Classification

Abstract

The problem of imbalanced classification is that in the classification problem, different classes of samples differ in the sample size, distribution and the classification cost. In the traditional classification algorithms, however, it is assumed that the samples are evenly distributed, resulting in the poor recognition of the samples of interest. In order to improve the accuracy of the minority class samples, scholars have proposed many approaches including data-level and algorithm-level methods. Data-level algorithms are used as an optional part of preprocessing and have a wider range of applications, which has led to a lot of attention. Oversampling, the algorithm uses various methods to increase the minority class samples in the training set, increases the recognition rate of the minority class samples. However, these oversampling methods are too coarse to improve the classification effect of the minority class samples, because they can’t make full use of the information in the original samples, but increase the training time because of adding extra samples. In this paper, we propose to use the distribution information of the minority class samples, use the variational auto-encoder to fit the probability distribution function of them, and reasonably expand the minority class sample set. The experimental results prove the effectiveness of the proposed algorithm.

Keywords: imbalanced classification; generative model; variational auto-encoder;

Introduction

The classification problem is a very important part of machine learning, and it is also the first step for artificial intelligence to understand human life. At present, most classifiers assume that the samples of different classes are evenly distributed, and the classification cost is the same. However, in reality, the data people are more concerned about is often scarce, such as the detection of credit card fraud and medical disease diagnosis. In the diagnosis of medical data, most of the results are normal while only a small proportion of the results are diagnosed as diseases, which indicates the even distribution in different classes samples. Second, if healthy people are misdiagnosed as diseases, they can be removed by other inspection methods, errors do not cause very serious accidents, but if the disease is diagnosed as healthy, it may cause the patient to miss the best treatment time and cause serious consequences. This is the second feature of the imbalanced classification problems: different classes of misclassification costs are inconsistent. At the same time, if samples are classified as diseases as much as possible because they are afraid to miss the disease samples, it will cause a huge waste of medical resources and intensify conflicts between doctors and patients. Therefore, it is not feasible to determine all samples as positive, and the best way is to try to separate these two results as correct as possible. Due to the scarcity of the minority class samples and perusing the global accuracy, the classifier pays less attention to the minority class, so the recognition effect of it is not good. Imbalanced classification problems arise in many fields, such as bioinformatics [1], [2], remote sensing image recognition [3], and privacy protection in cybersecurity [4]. The imbalanced problems cover widely and have a very important practical significance.

The traditional solutions to the imbalanced problems are divided into two parts: the algorithm-level algorithms and the data-level algorithms. The algorithm-level algorithms mainly focus on the different misclassification costs, and the algorithm level is mostly based on cost-sensitive method, that is, misclassification costs are different in different classes, such as improved neural network [5]: it uses the approximation of F1 value of the minority class as the cost function; the bagging algorithm[6] continues to enhance the misclassified the minority class samples, and improve the recognition rate of the minority class samples; structured SVM [7] uses the F1 value of the minority class samples as the optimization function, and thus has a better performance in the classification of the minority class samples.