Advanced Science and Technology Letters
Vol.31 (MulGraB 2013), pp.148-152
http://dx.doi.org/10.14257/astl.2013.31.33

Imbalanced Data Sample Pruning Algorithm Based on Dynamic threshold K Nearest Neighbor

Li Peng[1,2], Yu Xiao-yang[1], Bi Ting-ting[2], Huang Jiu-ling[2]

[1] Higher Educational Key Laboratory for Measuring and Control Technology, Instrumentations of Heilongjiang Province, Harbin University of Science and Technology, 150080 Harbin, China
[2] School of Computer Science and Technology, Harbin University of Science and Technology, 150080 Harbin, China
{pli, yuxiaoyang }@hrbust.edu.cn.

Abstract. This paper present a new sample pruning algorithm based on dynamic threshold KNN to deal with the complexity and overlapping problem of imbalanced data set. The phenomenon of data complexity and overlapping will reduce the classification performance and generalization ability of SVM classifier. Especially in imbalanced data set, this phenomenon is more obvious due to the quantity difference between positive and negative samples. We apply KNN to prune the training samples according to the similarities of each sample between its K labeled nearest neighbors and select different dynamic threshold to adapt the characteristic of imbalanced data set. The comparative experiments show that our algorithm can effectively improve the SVM classification performance in imbalanced data set.

Keywords: Imbalanced Data Set; Sample Pruning; K Nearest Neighbor

1 Introduction

With the advent of the information age and the rapid development of science and technology, it is easier than ever to get mass data and people also pay greater attention to the value of the data. Therefore, some data mining technologies were emerged and gain great value in scientific research and industrial production. However, the acquisition of data in the actual environment often has the overlapping phenomenon owing to the noise, accidental influence and equipment error etc. Therefore, we usually need to prune the samples in the application of these data sets and pruning algorithm has become an important research hot-spot in recent years. Such as, Dai proposed an ensemble pruning algorithm based on randomized greedy selective strategy and ballot [1]; Spanish researcher put forward a cost-effective pruning method for predicting web accesses [2]; Vukovic present a sequential learning pruning algorithm of hyper basis function neural network for function approximation [3]. Pruning algorithm is also widely used in the fields of biology [4], economy [5] and medicine [6] etc. However, there is little discussion about sample pruning method for

ISSN: 2287-1233 ASTL

imbalanced data set classification. Therefore, we take the lead in exploring this special subject.

2 Complexity and Overlap Analysis of Imbalanced Data Set

Linearly separable data don't exit in reality and most data is complexity with overlapping phenomenon. Especially in the imbalanced data set, this situation is more serious. Fig.1 and Fig.2 show the complexity and overlap in sample space; it is one of the main reasons led to the decline in the performance of classifier.

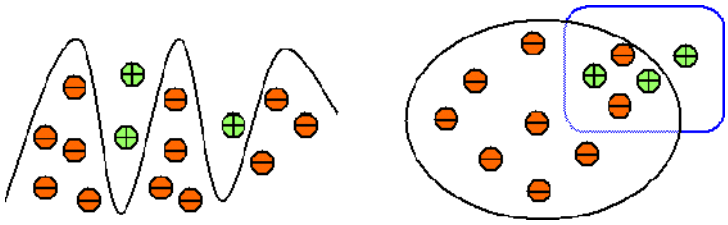

Fig. 1. The phenomenon of data complexity and overlapping

When the number of samples is large, complex distribution, and overlapped seriously, which is difficult to judge effective data and remove noise. Therefore, the problem of complexity and overlap is essential to be considered and resolved when imbalanced data classification technology is applied in the practical application. Although in theory, the SVM training only focus on these samples that near the optimal classification boundary. However, SVM classification performance will decreases obviously because serious data overlapping will rapidly increase the number of support vectors which lead to computational burden, overlearning and weakening generalization ability. Hence, it is significant for SVM classifier to find an effective samples pruning method to improve this issue.

3 Pruning Algorithm Based on Dynamic Threshold KNN

We propose a sample pruning algorithm based on dynamic threshold KNN to solve the complexity and overlap of imbalanced data set. K nearest neighbor (KNN) is a mature theory of machine learning and the basic motivation for considering the KNN rule rests on our earlier observation about matching probabilities with nature. The KNN query starts at the test point x and grows a spherical region until it encloses k training samples, and it labels the test point by a majority vote of these samples. We notice first that if k is fixed and the number n of samples is allowed to approach infinity, then all of the k nearest neighbors will converge to x .

The KNN algorithm described above is only applicable to general data set, but for imbalanced data set, the effect is not ideal if we directly apply this method. In Imbalanced data set, the positive samples are very scarce, so the positive information