A New SVM Reduction Strategy of Large-Scale Training Sample Sets

Fang Zhu, School of Computer and Communication Engineering, Northeastern University, Qinhuangdao, China
Junfang Wei, School of Resource and Material, Northeastern University, Qinhuangdao, China & Tianjin Foreign Studies University, TianJin, China
Tao Gao, North China Electric Power University, Beijing, China & Electronic Information Products Supervision and Inspection Institute of Hebei Province, Shijiazhuang, China

ABSTRACT

There has become a bottleneck to use support vector machine (SVM) due to the problems such as slow learning speed, large buffer memory requirement, low generalization performance and so on. These problems are caused by large-scale training sample set and outlier data immixed in the other class. Aiming at these problems, this paper proposed a new reduction strategy for large-scale training sample set according to analyzing on the structure of the training sample set based on the point set theory. By using fuzzy clustering method in this new strategy, the potential support vectors are obtained and the non-boundary outlier data immixed in the other class is removed. In view of reducing greatly the scale of the training sample set, it improves the generalization performance of SVM and effectively avoids over-learning. Finally, the experimental results shown the given reduction strategy can not only reduce the train samples of SVM and speed up the train process, but also ensure accuracy of classification.

Keywords: Fuzzy Clustering Method (FCM), Large-Scale Training Samples, Point Set, Reduction, Support Vector Machine (SVM)

1. INTRODUCTION

Support vector machine is a kind of machine learning method which is put forward by Vapnik and others based on statistical learning theory. In view of its avoiding effectively local minimum value problem, good generalization performance and good classification accuracy, SVM has been applied more and more widely in pattern recognition, regression analysis and feature extraction for recent years, which has become an international new research hotspot in the field of artificial intelligence and machine learning. However, the big learning samples bring about slowly learning speed and large storage demand, which directly obstruct the SVM technique application. Moreover, for
training the sample data mingled with outlier data in the relatively class of sample, it often can not improve the classification capability. On the contrary, it will greatly increase the burden of the training calculation, and may also cause over learning so as to increase the VC dimension of classification discriminant functions, which largen the confidence interval, finally affect the generalization of SVM. Therefore, it appears a lot of improved algorithm of support vector machine (Agarwal, 2002; Daniael & Cao, 2004; Luo, 2007; Xiao, Li, & Zhang, 2006; Li, Wang, & Yuan, 2003; Zeng, 2007; Tan & Ding, 2008; Cao, Liu, & Zhang, 2006).

The reduction strategy in the article Zeng (2007) and Cao, Liu, and Zhang (2006) is presented based on the idea of class center. After obtaining the clustering center of the positive and negative sample, it reduces the training sample by determining the provision radius relationship between the sample and the clustering center. But this method is suitable only for the sample set with convex set; In the article Xiao, Li, and Zhang (2006), it makes restriction on the training sample by C-mean clustering method, that if the all samples of a group are from the same class, a clustering center instead, otherwise, reserving all the samples of the group. Therefore, that method is able to reduce effectively the non-convex training sample set, but when the cluster number less than 1/20 of the sample, the reduction effect is not obvious. For the larger sample set, with the increase of the cluster number, it will increase the cost of calculation time against the sample reduction, the method does not have practical significance; The NN-SVM algorithm proposed in the article Li, Wang, and Yuan (2003) is according to the similarities between the nearest class with each sample to determine accepting or rejecting. This method can not only reduce the size of the samples but also reduce the SVM generalization performance influence caused by outlier data, but it will spend a lot of time when looking for the nearest point of each sample points. For the larger sample set terms, the algorithm efficiency is extremely low, which also lost practical significance; Another reduction PSOC strategy is presented in article Luo (2007), that according to the geometry characteristic of the training sample which linear divided in the high dimension space to realize the reduction of samples, by calculating the angle between per sample and positive-negative clustering center of attachment in the high dimensional space. That is a kind of algorithm with practical significance, but which requires a lot of nuclear calculation, so the efficiency is not high.

In view of the above analysis of the basic idea of the existing improved algorithm and their advantages and disadvantages, this paper put forward a new large scale training samples reduction strategy based on the theory of point set for support vector machine (SVM). Aiming at the large scale training samples mingled with the class outlier data, it can effectively reduce sample scale and the influence of the classification discriminant functions caused by the outlier data mixed in the relative class, so as to increase the training speed without affecting the SVM classification performance.

2. SUPPORT VECTOR MACHINE (SVM)

SVM is based on the principle of structural risk minimization (Cristianini & Shawe-Taylor, 2006). The ideal of SVM is to search for an optimal hyperplane to separate the data with maximal margin. When the training set is non-linear, the training vector x is mapped into a higher dimensional feature space by a nonlinear function \( \varphi(x) \), and in the feature space who’s dimension maybe infinite construct the optimal classification hyperplane and the classifier’s decision function.

2.1. Optimization Classification Hyperplanes

Hyperplanes requests the classification line can not only separate the two classes but also make the two types of classification interval largely.
A Platform for Pervasive Building Monitoring Services Using Wireless Sensor Networks

Capacity and Service (CapServ) Adaptive Trust Computation by Territory Formation in Ubiquitous Environment