International Journal of Innovative Computing, Information and Control
Volume **4**, Number **7**, July **2008**

ICIC International ©2008 ISSN 1349-4198
pp. 1751–1760

SAMPLE REDUCING METHOD IN SUPPORT VECTOR MACHINE BASED ON K-CLOSEST SUB-CLUSTERS

Xiaofeng Song, Weimin Chen and Bin Jiang

College of Automation Engineering
Nanjing University of Aeronautics and Astronautics
Nanjing 210016, P. R. China
{xfsong; binjiang}@nuaa.edu.cn; cwm1018@sohu.com

Received March 2007; revised August 2007

Abstract. *In order to decrease the training time of support vector machine (SVM) on large scale datasets, sample reducing methods are always employed in SVM by the researchers. The results of existing sample reducing methods are not very satisfactory because the reduced sample cannot contain the most support vectors which mainly determine the classification results. So we investigate the method of sample reducing in support vector machine based on K-Closest Sub-Clusters (KCSC) in this paper. The proposed method employs K-Closest Sub-Clusters to reduce the training samples, remaining the sample datasets near the classification boundary, which would probably be the support vectors, and deletes the sample datasets beyond the classification boundary, which would not contribute to the classification. The proposed method is validated on two artificial datasets and two real world datasets. The results of validation show that the training time of SVM is reduced greatly on the prerequisite of maintaining training and testing classification accuracy.*

Keywords: Support vector machine, K-closest sub-clusters, Sample reducing

1. **Introduction.** From 60s to 90s, Vapnik had studied a great number of learning problems based on sample data, and then in 1992 developed Statistical Learning Theory (SLT) which can especially solve pattern recognition based on small sample data in high dimensional space. Support Vector Machine (SVM) based on SLT was proposed in 1995 as a novel technique for pattern classification [1]. Now SLT and SVM are becoming a new hot area following the neural networks. The application of SVM in many fields is in the ascendant, and has gained substantial success in bioinformatics, face detection, handwriting digital recognition, etc. But as a new technique, SVM also has some shortcomings that need to be improved, especially for too much time demanding when training large-scale data set.

In this paper, we discuss the sample reducing algorithm in SVM on large-scale data set. The iterative training process of SVM belongs to a convex quadratic programming algorithm, which can guarantee the existence of a global optimal solution. Compared with other traditional algorithms, including Neural Network, SVM has many prominent advantages, which can efficiently avoid the problems of over-learning, dimension disaster, and local optimum. For those large datasets, the classical QP algorithm employed by the SVM could cost greatly in time and space. Many heuristic improved algorithms have been proposed, such as SMO [2], Osuna's algorithm [3], Incremental SVM Learning Algorithm [4], etc. However, the existing problems have not yet been fully solved, as all of these improvements have still remained to train the whole large datasets. The classifying hyper plane constructed by SVM depends on only a fraction of training samples which called support vectors lying in the decision boundary, so other strategies have been focused on