数据库、信号与信息处理

半监督SVM的工作集样本预选取方法

冼广铭,曾碧卿,李星丽

华南师范大学 南海校区计算机工程系，广东 佛山 528225

摘要　　针对传统的半监督SVM训练方法把大量时间花费在非支持向量优化上的问题，提出了在凹半监督支持向量机方法中采用遗传FCM（Genetic Fuzzy C Mean，遗传模糊C均值）进行工作集样本预选取的方法。半监督SVM优化学习过程中，在原来训练集上（标签数据）加入了工作集（无标签数据），从而构成了新的训练集。该方法首先利用遗传FCM算法将未知数据划分成某个数量的子集，然后用凹半监督SVM对新数据进行训练得到决策边界与支持矢量，最后对无标识数据进行分类。这样通过减小工作样本集，选择那些可能成为支持向量的边界向量来加入训练集，减少参与训练的样本总数，从而减小了内存开销。并且以随机三维数据为例进行分析，实验结果表明，工作集减小至原工作集的一定范围内，按比例减少工作集后的分类准确率、支持向量数与用原工作集相比差别不大，而分类时间却大为减少，获得了较为理想的样本预选取效果。

关键词　　半监督SVM　遗传FCM　样本预选取

分类号

Working dataset pre-selection method of semi-supervised SVM

XIAN Guang-ming,ZENG Bi-qing,LI Xing-li

Computer Engineering Department of Nanhai Campus，South China Normal University，Foshan，Guangdong 528225，China

Abstract

Aiming at traditional semi-supervised support vector machine training methods spend most of time on non support vectors，we propose the pre-selection sample method of genetic algorithm fuzzy C mean for working dataset.In the optimum learning process，working dataset（unlabeled data） is added to original training dataset（label data） to construct new training dataset.Firstly the new method is to divide unlabeled data（working datasets） into many subsets with a new label by Genetic Fuzzy C Means clustering，then train the concave SVM using the new data set to get decision boundary and support vectors，at last use the SVM classifier to classify the unlabeled data.The method proposed can select boundary vectors which have most probability to be support vectors to training set and save memory through decreasing working dataset.When working dataset is decrease to some scale of original working dataset，experimental results of three dimensional random data show that both classification accuracy rate and number of support vectors between reducing working dataset and original working dataset is almost same.But classification time of reducing working dataset is largely decreasing.It is illustrated that the result of sample pre-selection is satisfied.

Key words　semi-supervised Support Vector Machine　Genetic algorithm Fuzzy C Mean　sample pre-selection

通讯作者　冼广铭 Xgm20011@163.com