中南大学学报 自然科学版 英文版

教育部 主管
中南大学 主办

论文摘要

中南大学学报(自然科学版)

ZHONGNAN DAXUE XUEBAO(ZIRAN KEXUE BAN)

Vol.40 No.6 Dec.2009

[PDF全文下载] [全文在线阅读]

文章编号：1672-7207(2009)06-1636-06

基于均矢量相似性的机器学习样本集划分

陈先来[1，2]，杨路明[1]

(1．中南大学 信息科学与工程学院，湖南 长沙，410083；
2．中南大学 湘雅医学院，湖南 长沙，410013)

摘 要：提出一种基于均矢量相似性的机器学习样本集分割方法(MSSS)，根据样本集中每个样本矢量与均矢量之间的余弦相似性，将样本划分成训练集和测试集。为评价MSSS方法性能，分别用随机分割法(RSS)和MSSS方法，按不同比例划分来自UCI的4个数据集，对产生的训练集-测试集进行Hotelling T^2检验；另外，采用得到的训练集对分类BP神经网络进行训练，以相应的测试集测试神经网络。研究结果表明：对用RSS划分4个数据集产生的训练集-测试集进行Hotelling T^2检验，发现均存在F值超出界值的现象，而MSSS均未出现；使用MSSS训练的神经网络所产生的训练-测试误差差异、准确率差异均比使用RSS训练的神经网络所产生的小，说明用MSSS划分产生的训练集与测试集的一致性比用RSS划分产生的好。

关键字：均矢量；样本集分割；相似性；机器学习

Partitioning machine learning sample set using similarity to mean vector

CHEN Xian-lai[1, 2], YANG Lu-ming[1]

（1. School of Information Science and Engineering, Central South University, Changsha 410083, China;
2. Xiangya School of Medicine, Central South University, Changsha 410013, China）

Abstract:MSSS (Mean-similarity-based splitting sample), an algorithm for partitioning machine learning sample set, was presented based on similarity to mean vector. A sample set was splited into training set and test set by cosine similarity of each sample vector to mean vector. Simulation study were set up to evaluate MSSS. Four data sets from UCI were individually split by different proportions with MSSS and randomly splitting sample (RSS). The training set and test set were tested by Hotelling T^2. Back propagation neural networks for classification were built up. Training set was used for training networks and test set for testing networks. The result shows that the F value of Hotelling T^2 test for RSS might overtop its border, but that for MSSS does not. In contrast with RSS, MSSS has significantly lower error difference between training error and test error and accuracy difference between training accuracy and test accuracy of the network. It can be confirmed that the consistency between training set and test set from MSSS is superior to that from RSS.

Key words:mean vector; splitting sample set; similarity; machine learning

地 址：湖南省长沙市中南大学 邮编： 410083
电 话： 0731-88879765 传真： 0731-88877727
电子邮箱： zngdxb@mail.csu.edu.cn 湘ICP备09001153号