地质地球所提出利用无标签数据构建状态监测系统的主动半监督学习方法

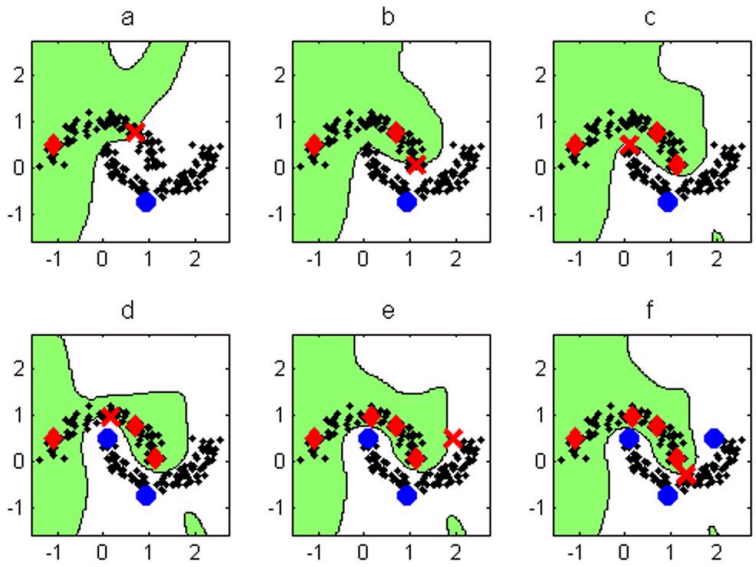
文章来源：地质与地球物理研究所 发布时间：2011-07-15 【字号： 小 中 大 】

基于模式识别的状态监测方法是通过对已知数据进行训练获取系统的状态监测模型，进而根据模型对系统状态进行判别。模式识别方法可以分为有监督学习和无监督学习两类。在有监督学习中，学习机利用的是已知标签样本，而无监督学习中只关注未知标签样本（有标签样本是有对应的类别标签的数据，无标签样本是类别标签缺失的数据）。只使用少量已知标签样本，那么训练得到的学习模型不具有很好的泛化能力，同时造成大量未知标签样本的浪费；而如果只使用大量未知标签样本，将会忽略已知标签样本的价值。

中科院地质与地球物理研究所空间环境探测实验室工程师赵秀宽与合作者提出了一种主动半监督学习方法A-LapSVM，力图挖掘未知标签数据中的有用信息，通过标记少量的标签样本来得到精确的监测模型。研究采用版本空间理论对提出的方法进行了分析，从理论上分析了方法的有效性。分别通过仿真数据、轴承数据和齿轮试验台数据对提出的A-LapSVM方法进行了验证：仿真数据中只用7个有标签数据就可以将200个数据样本完全正确分开；滚动轴承数据中只需要40个标签数据就可以达到96%的分类准确率，而其他方法需要200个以上的有标签数据才能达到相同的分类效果；齿轮试验台只需要24个标签数据就可以达到97.6%的分类准确率，而其他方法需要更多的有标签数据才能达到相同的分类效果。在实际应用中，A-LapSVM方法可以有效的降低人工标记标签的工作量。

该研究成果近期发表在国际知名人工智能学期刊*Expert Systems with Applications*（Zhao et al. *An effective procedure exploiting unlabeled data to build monitoring system. Expert Systems with Applications*, 2011, 38（8）：10199－10204）。

原文链接



附图说明：图中是采用双月模型仿真数据集来训练A-LapSVM主动半监督学习器的训练过程。该数据集包含200个样本，实验中初始化每类中有一个标签数据。图中红色的菱形代表正类，蓝色的圆形为负类，黑点为无标签数据（即假设不知道类别标签），红色叉形为本次迭代选择的将要赋予标签的样本点。由图中可以看出，经过5次迭代，分类器已经可以完全把两类数据分开。在第6步已经取得了很好的分类效果（图中的f所示）。

打印本页 关闭本页