改进 K-means 的双向采样非均衡数据分类方法

柳　毅，曾　昊
（广东工业大学 计算机学院，广东 广州 510006）

摘　要：针对分类器在不均衡数据集上对小类分类准确率较差的问题，提出了改进 K-means 的双向采样算法 KMBS（k-means bi-directional sampling），并将集成学习应用在分类算法上.首先，使用改进的 K-means 聚类算法将原始数据集划分为不同的聚类簇.其次，在聚类簇中使用改进的 SMOTE 算法对小类样本过采样，对聚类簇内的大类样本欠采样，使数据集平衡.多次执行该算法可以产生多个差异较大的数据集，因此训练出多个差异较大的分类器，提升集成学习的效果.通过分析实验结果，该算法较现有几种算法不仅能提高整体分类性能，并且有效提高小类样本的分类性能.

关键词：　不均衡学习；双向采样；分类算法；集成学习

Improved the bi-directional sampling unbalanced data classification method of K-means

LIU Yi，ZENG Hao
(College of Computer Science, Guangdong University of Technology, Guangzhou 510006, China)

Abstract：　Aiming at the poor classification accuracy of minority classes by classifier on unbalanced data sets, an improved k-means bi-directional sampling algorithm KMBS (k-means bi-directional sampling) is proposed, and integrated learning is applied to the classification algorithm. First, the improved k-means clustering algorithm is used to divide the original data set into different clustering clusters. Secondly, oversampling of the minority and under-sampling of the majority in the cluster using the modified SMOTE algorithm in the cluster, so as to make the dataset balance. Multiple executions of this algorithm can produce multiple data sets with large differences, so multiple classifiers with large differences can be trained to improve the effect of ensemble learning. By analyzing the experimental results, this algorithm can not only improve the overall classification performance, but also improve the classification performance of a few kinds of samples.

Key words：　imbalanced learning; bi-directional sampling; classification; ensemble learning

作者简介：
柳　毅　男，（1976-），博士，教授.研究方向为网络安全.
曾　昊(通讯作者)　男，（1993-），硕士.研究方向为网络安全.E-mail：2966739148@qq.com.