一种优化的改进 *k*_means 算法

张淑清[1,2]，黄震坤[1,2]，冯 铭[1,2]
（1 武汉大学 国家多媒体软件工程技术研究中心，湖北 武汉 430072；
2 武汉大学 计算机学院，湖北 武汉 430072）

摘 要：传统的 *k*_means 算法随机地选择初始中心，导致最终聚类结果陷入局部最优且准确率低. min_max 算法针对初始中心随机选择的缺点提出了改进. 但原始的 *k*_means 和 min_max 算法都忽略了利用原空间欧式距离度量相似性的不合理性. 对此提出改进算法，利用映射函数将输入向量转换到特征空间，在 min_max 算法基础上确定初始中心后，根据特征空间中的欧式距离来进行分类. 实验证明了改进算法的有效性，在 iris 和 wine 数据集上获得了 92.86%和 72.34%的分类准确率.

关键词：随机地；欧式距离；不合理；映射函数；特征空间
中图分类号：TP301 **文献标识码**：A **文章编号**：1000－7180(2015)12－0036－04

An Optimized *k*_means Algorithm

ZHANG Shu-qing[1,2], HUANG Zhen-kun[1,2], FENG Ming[1,2]
(1 National Engineering Research Center for Multimedia Software, Wuhan University, Wuhan 430072, China;
2 College of Computer, Wuhan University, Wuhan 430072, China)

Abstract: The traditional *k*_means randomly selects initial cluster centers and divides the sample points by Euclidean distance in the original space, so the auuracy classification isn't enough. The in_max algorlthm obtatins more improvenent than the traditional algorithm. However, the traditional algorithm and the min_max algorithm neglect the irrationality of the classification by Euclidean distance. The improved algorithm maps the input vectors into the feature space and determines the initial centers , at last clusters by the distance between two points in the feature space. The improved algorithm is evaluated on available datasets called iris and wine.

Key words: randomly; Euclidean distance; irrationality; mapping function; feature space

作者简介：
张淑清 女，(1990-)，硕士研究生. 研究方向为图像聚类. E-mail: 15527179027@163.com.
黄震坤 男，(1983-)，博士研究生. 研究方向为视频编码、视频分析.
冯 铭 男，(1988-)，硕士研究生. 研究方向为视频分析.

收稿日期：2015－02－13；**修回日期**：2015－04－09