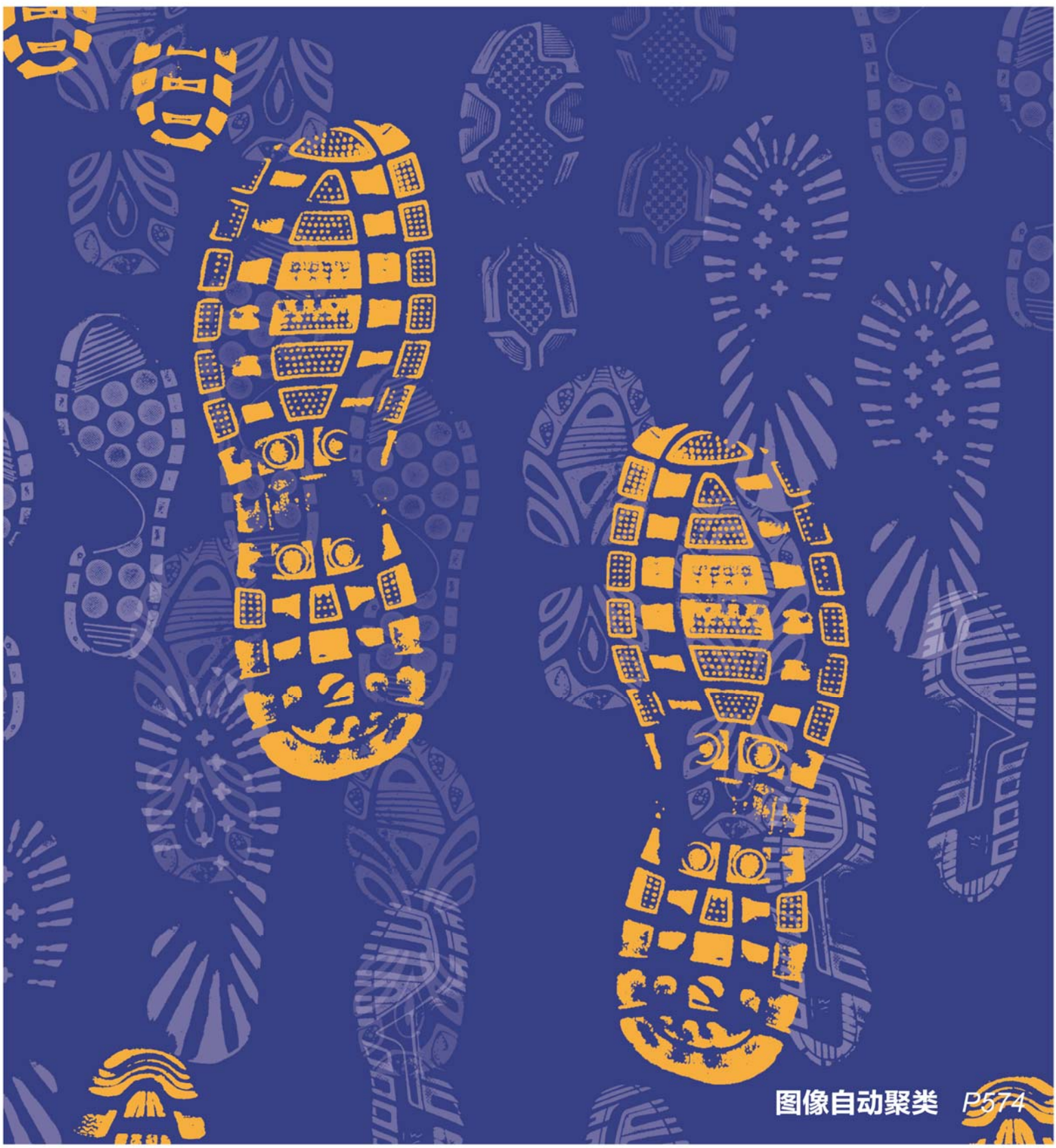
JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2016
05
VOL.21

ISSN1006-8961
CN11-3758/TB



图像自动聚类 P574

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第21卷第5期（总第241期）
2016年5月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

主管单位 中国科学院
主办单位 中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 李小文
编辑出版 《中国图象图形学报》编辑出版委员会
邮政信箱 北京9718信箱
邮　　编 100101
电子信箱 jig@radi.ac.cn
电　　话 010-64807995
网　　址 www.cjig.cn

广告经营许可证 京朝工商广字第0361号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Institute of Remote Sensing and Digital Earth, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief LI Xiaowen
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
P.O.Box 9718, Beijing, P.R.China
Zip code 100101
E-mail jig@radi.ac.cn
Telephone 010-64807995
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

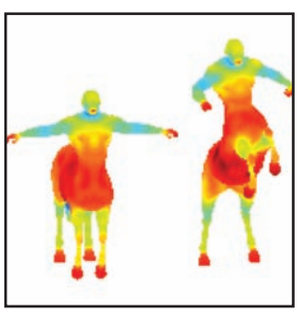
CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

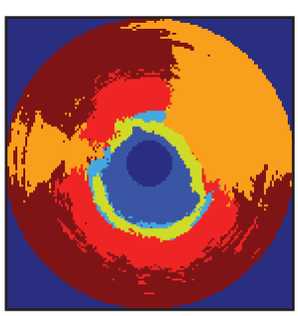
2016年5月 第21卷 第5期（总第241期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

利用双通道卷积神经网络的图像超分辨率算法(第556页)



融合特征描述符约束的3维等距模型对应关系计算(第628页)



结合先验形状信息和序贯学习的心血管内超声外弹力膜检测(第646页)

综述

中国图像工程：2015

章毓晋 …… 533

图像处理和编码

基于代价敏感Adaboost目标跟踪

薛一哲，王拓 …… 544

利用双通道卷积神经网络的图像超分辨率算法

徐冉，张俊格，黄凯奇 …… 556

图像分析和识别

局部均匀模式描述和双加权融合的人脸识别

任福继，李艳秋，许良凤，胡敏，王晓华 …… 565

K步稳定的鞋印花纹图像自动聚类

王新年，舒莹莹 …… 574

空间约束混合高斯运动目标检测

董俊宁，杨词慧 …… 588

图像理解和计算机视觉

分层信息融合的物体级显著性检测

李波，金连宝，曹俊杰，冷成财，卢春园，苏志勋 …… 595

结合区域协方差分析的图像显著性检测

张旭东，吕言言，缪永伟，郝鹏翼，陈佳舟 …… 605

前景判别的局部模型匹配目标跟踪

刘大千，刘万军，费博雯 …… 616

计算机图形学

融合特征描述符约束的3维等距模型对应关系计算

杨军，闫寒，王茂正 …… 628

一类加权有理插值样条曲面及局部约束控制

刘植，肖凯，陈晓彦，江平，谢进 …… 636

医学图像处理

结合先验形状信息和序贯学习的心血管内超声外弹力膜检测

林慕丹，杨丰，梁淑君，赵海升，黄铮，崔凯 …… 646

遥感图像处理

局部显著特征下的光学遥感图像舷靠舰船检测

李轩，刘云清，卞春江，毛博年 …… 657

长时序高光谱图像清晰度影响因素分析

尚任，黑保琴，李盛阳，覃帮勇，张九星 …… 665

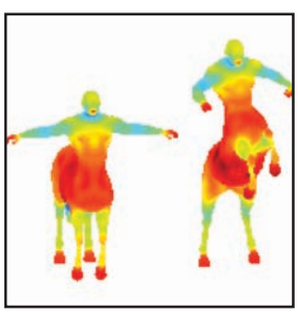
基于变差函数的中高分辨率SAR影像农村建筑区提取

林晨曦，周艺，王世新，刘文亮，田野，张燕楠 …… 674

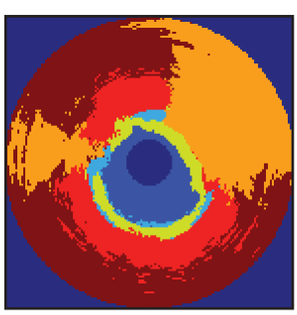
Volume 21, Number 5
Published May 16, 2016

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Image super-resolution using two-channel convolutional neural networks(P556)



Calculation of correspondences between three-dimensional isometric shapes with the use of a fused feature descriptor (P628)



External elastic membrane border detection based on for intravascular ultrasound images(P646)

Review

Image engineering in China: 2015
Zhang Yujin 533

Image Processing and Coding

Object-tracking method based on improved cost-sensitive Adaboost
Xue Yizhe, Wang Tuo 544

Image super-resolution using two-channel convolutional neural networks
Xu Ran, Zhang Junge, Huang Kaiqi 556

Image Analysis and Recognition

Face recognition method based on local mean pattern description and double weighted decision fusion for classification
Ren Fuji, Li Yanqiu, Xu Liangfeng, Hu Min, Wang Xiaohua 565

K-steps stabilization based on automatic clustering of shoeprint images
Wang Xinnian, Shu Yingying 574

Moving object detection using improved Gaussian mixture models based on spatial constraint
Dong Junning, Yang Cihui 588

Image Understanding and Computer Vision

Object level saliency detection by hierarchical information fusion
Li Bo, Jin Lianbao, Cao Junjie, Leng Chengcai, Lu Chunyuan, Su Zhixun 595

Image saliency detection using regional covariance analysis
Zhang Xudong, Lyu Yanyan, Miao Yongwei, Hao Pengyi, Chen Jiazhou 605

Foreground discrimination in local model-matching tracking
Liu Daqian, Liu Wanjun, Fei Bowen 616

Computer Graphics

Calculation of correspondences between three-dimensional isometric shapes with the use of a fused feature descriptor
Yang Jun, Yan Han, Wang Maozheng 628

Weighted rational interpolation spline surfaces and local constraint control
Liu Zhi, Xiao Kai, Chen Xiaoyan, Jiang Ping, Xie Jin 636

Medical Image Processing

External elastic membrane border detection based on sequential learning and prior shape information for intravascular ultrasound images
Lin Mudan, Yang Feng, Liang Shujun, Zhao Haisheng, Huang Zheng, Cui Kai 646

Remote Sensing Image Processing

Inshore ship detection method in optical remote sensing images using local salient characteristics
Li Xuan, Liu Yunqing, Bian Chunjiang, Mao Bonian 657

Analysis of factors influencing clarity of hyper spectral images in long time series
Shang Ren, Hei Baoqin, Li Shengyang, Qin Bangyong, Zhang Jiuxing 665

Variogram-based rural build-up area extraction from middle and high resolution SAR images
Lin Chenxi, Zhou Yi, Wang Shixin, Liu Wenliang, Tian Ye, Zhang Yannan 674

E-mail: jig@radi.ac.cn
Website: www.cjig.cn
Tel: 010-64807995

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号：TN911.73　　**文献标识码**：A　　**文章编号**：1006-8961(2016)05-574-14

论文引用格式：Wang X N, Shu Y Y. K-steps stabilization based on automatic clustering of shoeprint images[J]. Journal of Image and Graphics, 2016, 21(5):574-587.[王新年,舒莹莹. K 步稳定的鞋印花纹图像自动聚类[J]. 中国图象图形学报,2016,21(5):574-587.][DOI:10.11834/jig.20160505]

K 步稳定的鞋印花纹图像自动聚类

王新年，舒莹莹

大连海事大学信息科学技术学院，大连　116026

摘　要：**目的**　鞋印是刑事侦查的重要物证之一，如何对积累的大量鞋印花纹图像进行自动归类管理是刑事技术迫切需要解决的问题之一。与其他类图像不同，鞋印花纹图像具有种类多但数目未知、同类花纹分布不均匀且同类花纹数目少的特点。基于鞋印花纹图像的这些特点，用目前典型的聚类算法对鞋印花纹图像集进行聚类，并不能取得很好的效果。在对鞋印花纹图像进行分析的基础上，提出一种 K 步稳定的鞋印花纹图像自动聚类算法。**方法**　对已标记的鞋印花纹图像进行统计发现，各类鞋印花纹之间在特征空间上存在互不相交的区域（本文称为隔离带）。算法的核心思想是寻找各类鞋印花纹之间的隔离带，来将各类分开。过程为：以单调递增或递减的方式调整特征空间中判定两点为一类的阈值，得到数据集的多次划分；若在连续 K 次划分的过程中，某一类的成员不发生变化，则说明这 K 次调整是在隔离带中进行的，即聚出一类，并从数据集中删除已标记的数据；选择下一个阈值对剩余的数据集进行划分，输出 K 步不变的类；依此类推，直到剩余数据集为空，聚类完成。**结果**　在两类公开测试数据集和实际鞋印花纹数据集上进行实验，本文算法的主要性能指标都超过典型算法，其中在包含 5 792 枚实际鞋印花纹数据集上的聚类准确率和 F-Measure 值分别达到了 99.68% 和 95.99%。**结论**　针对鞋印花纹图像特点，提出了一种通过寻找各类之间的隔离带进行自动聚类的算法，并在实际应用中取得了很好的效果。且算法性能受参数的变化以及类的形状影响较小。本文算法同样适用于具有类似特点的其他数据集的自动聚类。

关键词：鞋印花纹图像；聚类；隔离带；K 步稳定；可达半径；类集成树；任意形状类

K-steps stabilization based on automatic clustering of shoeprint images

Wang Xinnian, Shu Yingying

Information Science and Technology College, Dalian Maritime University, Dalian　116026, *China*

Abstract: **Objective** Shoeprints serve as vital evidence in forensic investigations, and determining how a massive amount of shoeprint images can be clustered automatically through long-term accumulation has become one of the urgent tasks of criminal technology. Unlike other image sets, the number of shoeprint categories is not only considerable but also unknown. Shoeprint images in the feature space are distributed inhomogeneously and sparsely, but the quantity of each class is low. For these reasons, most existing clustering algorithms cannot satisfactorily cluster shoeprints. In this study, an automatic clustering method is proposed to divide shoeprint sets effectively based on an analysis of the distribution of shoeprint images in the feature space. **Method** Through statistics on labeled shoeprint-image databases, we found that shoeprint sets of different

收稿日期：2015-08-17；**修回日期**：2016-01-11

基金项目：国家高技术研究发展计划(863)基金项目(2010AA710＊＊)

第一作者简介：王新年(1973—　)，男，副教授，2005 年于大连海事大学获通信与信息系统专业工学博士学位，主要研究方向为数字图像处理。E-mail：wxn@dlmu.edu.cn

通信作者：舒莹莹，硕士研究生，E-mail：syydip@163.com

Supported by: National High Technology Research and Development Program of China (2010AA710＊＊)

patterns do not intersect, and a blank region, where no shoeprints exist, is present between every two classes. Blank regions are called margins in this paper. The core objective of the proposed algorithm is to determine the margins between classes and use them to divide a shoeprint set. The process involves the following steps: 1) dividing the shoeprint set with monotonically increasing or descending thresholds, which are used to classify two shoeprint images into the same cluster; 2) searching for the cluster that does not change with K consecutive partitions; 3) outputting the stable cluster and removing the shoeprints belonging to the output stable cluster from the dataset; 4) choosing the next threshold and dividing the remaining dataset; 5) returning to step 2) until the remaining set is empty. **Result** Experimental results on two kinds of publicly available databases and one real shoeprint database which comprises 5792 images, have shown that the proposed algorithm outperforms state-of-the-art clustering algorithms on common clustering evaluation measures. The precision and F-measure of the proposed algorithm on the real shoeprint database are approximately 99. 68 and 95. 99 percent, respectively. **Conclusion** In this study, based on the distribution of shoeprint images in the feature space, an automatic clustering algorithm that searches for margins between clusters to divide a dataset is proposed. The proposed algorithm achieves a comparable or even better performance on clustering a shoeprint dataset than its competitors. Experiments have also shown that the performance of the algorithm is less sensitive to the parameter and shape of the clusters. The algorithm can also be applied to clustering other datasets of images with characteristics similar to those of shoeprint images.

Key words: shoeprint image; clustering; margin; K-steps stabilization; reachable radius; cluster aggregation tree; clusters of arbitrary shape

0 引　言

鞋印是刑事侦查的重要物证之一,其在案发现场的出现率和提取率都非常高。目前,每年各地提取的鞋印大约数千枚,甚至数万枚,且提取的鞋印多数以图像形式保存,久而久之,形成海量的鞋印花纹图像库,为使这些资源得到有效地利用,如何对鞋印花纹图像进行归类管理成了一个迫切需要解决的问题。

目前,鞋印花纹图像的归类管理主要是通过记录案情信息和人工标注花纹类型来实现的[1]。人工标注的优点是具有语义信息,缺点是工作量巨大且存在语义的二义性。本文的目的是研究基于图像内容的鞋印花纹图像自动聚类算法。

鞋印花纹图像自动聚类算法是数据聚类算法的一种,目前聚类算法主要分为:

1)基于划分的方法。核心思想是将给定的一个数据集合通过分裂的方法分成 K 组。算法首先是将数据集初步划分成 K 组,之后通过迭代改变分组直到达到最优的划分[2]。该类算法需要事先给定分组数目 K,典型代表算法有 K-means、CLARANS(clustering large applications based on randomized search)[3]等。

2)基于层次化的方法。核心思想是将数据集划分成一个树状的聚类集合,主要包括凝聚算法和分裂算法两种[4]。凝聚算法是自下向上进行的,即由最开始的一个数据一类到最后的所有数据形成一类;分裂算法是自上向下进行的,即由最开始的所有数据一类到最后一个数据一类。典型代表算法有 BIRCH(balanced iterative reducing and clustering using hierarchies)[5]、CURE(clustering using representatives)[6]等。

3)基于密度的方法。核心思想是从数据对象的分布密度出发,把密度大于一定阈值的区域连接起来形成任意形状的类。该类方法的结果对密度阈值很敏感[2]。典型代表算法有 DBSCAN(density based spatial clustering of applications with noise)[7]、OPTICS(ordering points to identify the clustering structure)[8]等。

4)基于网格的方法。核心思想是先将数据空间划分为有限个单元的网络结构,之后在构成的量化空间进行聚类[9]。典型代表算法有 STING(statistical information grid)[10]、WaveCluster[11]等。

5)基于模型的方法。核心思想是给每一个聚类假定一个模型,并寻找数据对给定模型的最佳拟合。该算法通过构建反映数据点空间分布的密度函数来实现聚类。这种聚类方法试图优化给定的数据和某些数学模型之间的适应性[2]。典型代表算法有 COBWEB[12]等。

6)基于集成的方法。核心思想是用多种方法对数据集分别进行聚类,从多个聚类结果中得到一个能够准确反映数据集结构的数据划分[2]。典型

代表算法有 CSPA(cluster-based similarity partitioning algorithm)[13]等。

这些算法在各自适用领域都取得了较大的成功,但对参数和初值依赖性很大,直接应用于鞋印花纹图像的自动聚类效果不好,主要原因在于:

1)在各地收集到的鞋印花纹图像库中,无法知道花纹的种类,因此需要事先知道类别数目的聚类算法如 K-means 等性能会受到影响。

2)由于采集的鞋印花纹图像质量较低,造成同类花纹之间的相似性差别较大,分布不集中,这就给基于密度的聚类算法如 DBSCAN 等性能带来影响。

3)鞋印花纹图像数据量大,通过人工调整参数并核实聚类结果的好坏,以确定较优聚类结果不现实,因此基于多个算法集成或参数较多的聚类算法可能取得较好的效果,但在现实中应用有困难。

根据鞋印花纹图像数据的分布特点,提出了一种 K 步稳定的鞋印花纹图像聚类算法(KSSC)。算法假设在特征空间中,各类鞋印花纹图像存在互不相交的区域,该区域在文中称为隔离带。K 步稳定是指在隔离带内进行连续 K 次划分,类的成员稳定不变。算法的核心思想是:不断调整特征空间中依据判定两点为一类的阈值,划分数据集,若在连续 K 次调整范围内,某一类的成员不发生变化,则说明这 K 次调整是在隔离带中进行的,即聚出一类,依此类推,便可将整个数据集划分完成。实验结果表明本文算法的主要性能指标都超过典型比较算法,并已经应用于实际系统。

1 问题描述与鞋印花纹图像分析

1.1 问题描述

设鞋印花纹图像数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_N\}$,其中 $\boldsymbol{x}_i=[x_{i,1},x_{i,2}\cdots,x_{i,d}]^{\mathrm{T}}\in\mathbf{R}^d$ 表示第 i 幅鞋印花纹图像的特征,特征是 d 维实数。矩阵 $\boldsymbol{W}\in\mathbf{R}^{N\times N}$ 表示的是数据集中各幅图像之间的距离,其中元素 $\boldsymbol{W}_{i,j}$ 表示图像 i 与 j 的距离。本文的目的是根据给定距离矩阵 $\boldsymbol{W}$ 将鞋印花纹图像数据集 $\boldsymbol{X}$ 划分为 K 类,其中 K 是未知的,需要从数据集中估计,每类成员的集合记为 $\boldsymbol{C}_i,i=1,2,\cdots,K$,所有类的集合为 $\boldsymbol{C}=\{\boldsymbol{C}_1,\cdots,\boldsymbol{C}_i,\cdots,\boldsymbol{C}_K\}$。且满足如下条件:

1) $\boldsymbol{C}_i\neq\varnothing,i=1,\cdots,K$ 且 $K<N$;

2) $\bigcup_{i=1}^{K}\boldsymbol{C}_i=\boldsymbol{X}$

3) $\boldsymbol{C}_i\cap\boldsymbol{C}_j=\varnothing,i,j=1,\cdots,K$ 且 $i\neq j$;

4)若数据集 $\boldsymbol{X}$ 的真实划分记为 $\boldsymbol{L}=\{\boldsymbol{L}_1,\cdots,\boldsymbol{L}_j,\cdots,\boldsymbol{L}_M\}$,其中 M 表示数据集的真实类别数,则希望 $P_i=\max\limits_{j=1,2,\cdots,M}\dfrac{|\boldsymbol{L}_j\cap\boldsymbol{C}_i|}{|\boldsymbol{L}_j|}$ 接近于1,即聚类的各类成员与真实的各类成员相同,其中 $|\cdot|$ 表示集合的势。

1.2 鞋印花纹图像的分布特点分析

由于鞋印花纹图像的数目及特征维数都很大,因此采取统计的方法来分析其分布特点。样本来源于某市刑警队提供的 5 792 枚数据,且由专家已做好标记,共 3 256 种花纹,即已知数据集的真实划分。鞋印花纹图像的分布特点与其采用的特征紧密相关,本文采用的是文献[14]所提出的基于小波傅里叶梅林变换的特征以及相似度计算与估计方法。经统计分析可知鞋印花纹图像集具有以下特点:

1)鞋印花纹图像集花纹种类多,而同类花纹图像较少。在已标记的 5 792 枚鞋印花纹图像中,各类花纹数目的统计如图 1 所示。为便于观看,将各类别按成员数目由小到大排序,可以看出,平均类成员个数为 2 幅,最大类的图像数目为 81 幅。其中有 2 338 个单幅图像类别(即该类别中只有 1 幅图像),多幅图像类别的平均类成员个数为 4 幅。可见花纹种类远远大于同类花纹数目,且各类成员个数分布不均匀。

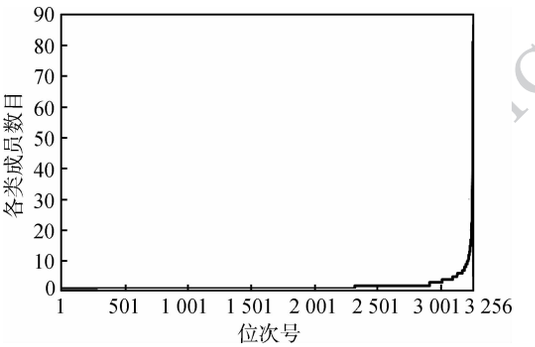


图 1 鞋印花纹图像各类别花纹数目统计图

Fig. 1 Distribution of the member number of each class

2)同类花纹图像分布不均匀。对于图像来说,相似性更便于理解,由此在已标记的 5 792 枚鞋印花纹图像集上,对各类类内成员之间的相似性分布进行分析。首先将数据点与其所在类其他数据点相似度的最大值称为数据点的相似性,其值越大,说明该数据点与所在类的成员的相似性就越大。基于数据点相似性定义以下统计指标分析同类花纹的分

布,分别是:

(1)类内成员相似性的标准差。即统计类内每个成员的数据点相似性的标准差,该值越大,说明类内成员分布得越分散。

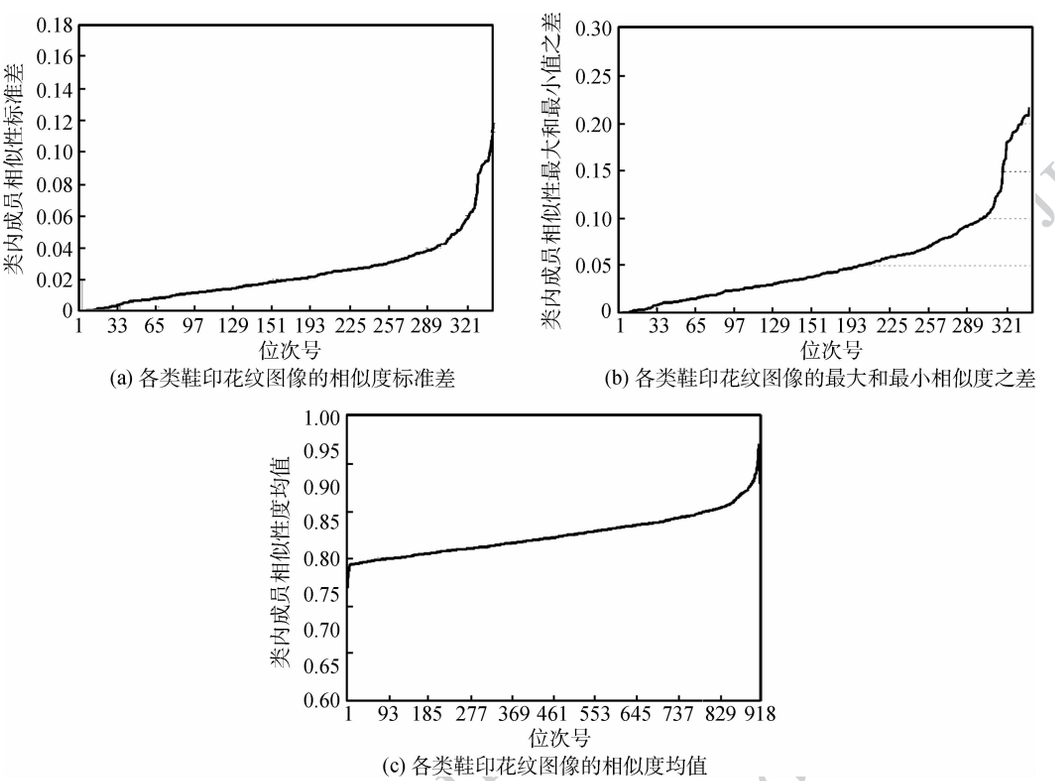
(2)类内成员相似性的最大值与最小值之差。即统计类内所有成员数据点相似性的最大值和最小值之差,该值越大,说明成员分布得越分散。

(3)类内成员相似性的均值。即类内每个成员的数据点相似性的均值,该值越大,说明类内成员相似性就越大。

图 2 所示的是各类别按指标由小到大排列后的分布图。

图 2(a)(b)是对成员个数超过 2 个的 343 组数据统计的结果。从图 2(a)可以看出,各类内相似度的标准差分布范围较广并且最大标准差达到 0.16;图 2(b)所示的是各类成员相似性的最大值和最小值之间的差,可以看出,有的类差异较小,接近于 0,有的差异比较大,达到 0.3,并且没有哪个值占优势。这两个指标说明各类内部花纹分布不均匀,很难通过一个阈值来判定两个点是否属于同一类。相似性阈值设得过大可能将本来应是一类的给分开了;而设得过小,可能把不是一类的却给分成一类。所以通过一个全局阈值来进行数据的划分的方法不能得到较好的聚类结果。

图 2(c)显示的是成员个数超过 1 个的 918 组数据的各类相似度均值,可以看出,各类的相似度均值分布范围都很广,从最小值 0.746 到最大值 0.962,而且也并没有哪个值占多数,这说明各类内成员的分布不均匀,有的比较紧密,如相似性均值为 0.962 的类,而有的类成员分布的比较松散,如相似度均值为 0.746 的类。这样的分布特性就会对基于密度的聚类算法的性能造成影响。



(a) 各类鞋印花纹图像的相似度标准差　(b) 各类鞋印花纹图像的最大和最小相似度之差

(c) 各类鞋印花纹图像的相似度均值

图 2　各类鞋印花纹图像相似度值统计分布图

Fig. 2　Statistical distributions of the shoeprint similarity values of each class ((a) distribution of the shoeprint similarity standard deviations of each class; (b) distribution of the differences between the maximum and minimum shoeprint similarity values of each class; (c) distribution of the shoeprint similarity means of each class)

3)各类花纹在特征空间上的分布形状不固定。由于现场背景和采集方式的多样性,采集的鞋印花纹图像受干扰程度以及残缺程度使同类图像之间的分布并不是球形,因此基于球形的聚类算法并不适合。部分鞋印花纹图像如图 3 所示。

4)各类花纹之间存在隔离带。统计 3 256 类鞋印花纹图像之间相似性的最大值和各类类内成员相似性的最小值,其中对于只有一个成员的类,其类内

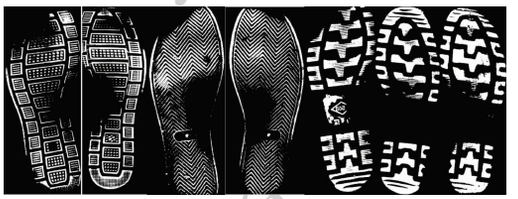


图 3　3 类鞋印花纹图像示意

Fig. 3　Three sample classes of the shoeprints

成员相似性最小值设为 1。可以发现，多数情况下，类间相似性的最大值都会小于类内各成员相似性的最小值。该现象说明各类之间存在一个空隙，称之为隔离带，通过这个隔离带，能将两个类的数据分开。为便于观看，按照类内相似性的最小值由小到大顺序进行排序，将各类的类内相似性的最小值和类间相似性的最大值分别显示如图 4 所示。两个空心圆之间位次号的类只有一个成员。可以发现，多数情况下，三角标记的数据位于圆形标记的数据的下方，且有一定的距离。这说明鞋印花纹图像各类之间存在明显的隔离带。

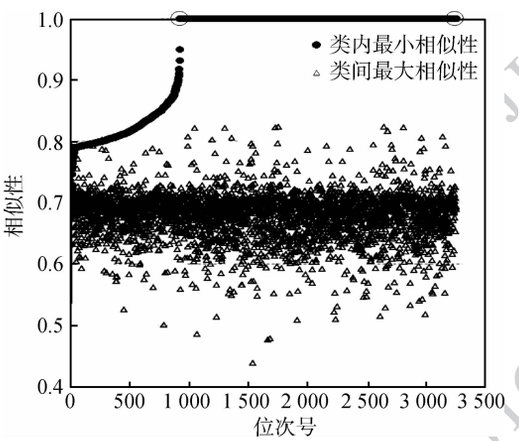


图 4　鞋印花纹图像类间隔离带示意图

Fig. 4　Illustration of the margins of the shoeprint classes

本文正是基于鞋印花纹图像分布的这个特点，提出了 K 步稳定的聚类算法。

2　算法原理与描述

2.1　一个直观的例子

以图 5 所示的数据分布为例说明本文算法的思想。在图 5 中，共有 27 个数据点，这些数据点的标准分类为 5 类。图中各点所在圆的半径大小说明其与最近的同类点的距离，在半径范围内，该点至少与所在类内一个点相连。半径越大，说明其与最近同类点的距离就越大。虚线表示的是各类的边界线，即隔离线，该线将类内成员包围住；边界线之间的区域称为隔离带。可以发现通过增加各点所在圆的半径，能使类内任意两点之间总存在一条可达的路径。若任意两点之间存在一条可达的路径，则认为这两点属于同一类。当任意点所在圆越过隔离线，处于隔离带内时，在一定范围内，其所在类成员不发生变化，我们称其所在的类稳定，即聚出一类，依此类推，便可将整个数据集划分完成。

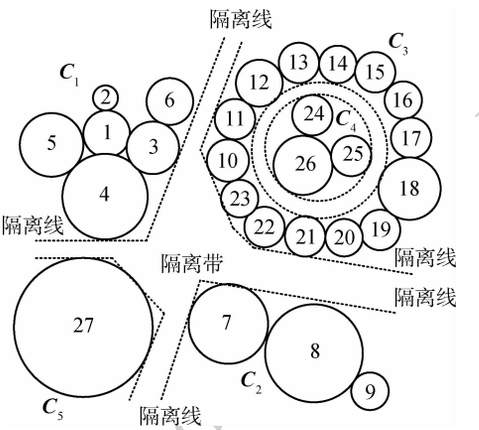


图 5　本文算法的直观例子

Fig. 5　An intuitive example of the proposed algorithm

2.2　定义

定义 1　直接可达半径。定义为直接判定两点为同一类的距离阈值，记为 δ。若点 $\boldsymbol{p}$ 和 $\boldsymbol{q}$ 之间的距离小于等于 δ，则可判定点 $\boldsymbol{p}$ 和 $\boldsymbol{q}$ 是同类点，称两点直接可达。图 5 中，1 与 2、3、4 和 5 点都是直接可达。另外将一个点与其同类别中最近点的距离定义为该点的最小直接可达半径。

定义 2　间接可达半径。若点 $\boldsymbol{p}$ 和 $\boldsymbol{q}$ 之间的距离大于直接可达半径，但它们之间存在一条链 $[\boldsymbol{p}, \boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_n, \boldsymbol{q}]$，在该链上相邻两点之间的距离皆小于直接可达半径，则也可判定点 $\boldsymbol{p}$ 和 $\boldsymbol{q}$ 是同类点，且两点间接可达。并且将这条链上相邻两点之间距离的最大值称为间接可达半径，记为 δ_I，即 $\delta_I = \max_i(\delta_i), i = 0, 1, \cdots, n$，其中 δ_0 表示的是 $\boldsymbol{p}$ 与 $\boldsymbol{p}_1$ 点的距离，δ_n 表示的是 $\boldsymbol{p}_n$ 与 $\boldsymbol{q}$ 点的距离，δ_i 表示链上 $\boldsymbol{p}_i$ 与 $\boldsymbol{p}_{i+1}$ 点的距离。如图 5 中，5 点和 6 点可以通过 3 点和 4 点或者 3 点和 1 点间接可达。

定义 3　类可达半径。定义为类内所有点的最小直接可达半径的最大值，记为 δ_c，即

$$\delta_c = \max_{p \in \boldsymbol{C}} \min_{q \in \boldsymbol{C}} \delta(\boldsymbol{p}, \boldsymbol{q})$$

其中，$\boldsymbol{C}$ 表示 $\boldsymbol{p}$ 和 $\boldsymbol{q}$ 所在的类别。

定义 4　类集成树，记为 $\boldsymbol{T}$。利用一个直接可达半径 δ_h 就可完成对数据集的一次划分，记为 $\boldsymbol{\pi}_h = [\boldsymbol{C}_1^{(h)},\cdots,\boldsymbol{C}_s^{(h)},\cdots,\boldsymbol{C}_{n_h}^{(h)}]$，其中，$\boldsymbol{C}_s^{(h)}$ 表示利用直接可达半径 δ_h 对数据集划分所得的第 s 类，n_h 表示本次划分类别的总数目。将直接可达半径由小到大逐渐增加，则每次划分的类的数目不变或者减小，即第 $h+1$ 次划分的类数目 n_{h+1} 要小于或等于第 h 次划分的类数目，亦 $\boldsymbol{\pi}_{h+1}$ 中的类至少包含 $\boldsymbol{\pi}_h$ 中的一个类。将每次划分的结果作为一层按序排放，该层中的每个节点即是该次划分的各个类别。每个节点的父节点是上一层中与其交集的势最大的那个节点，即 $r^* = \arg\max_r |\boldsymbol{C}_s^{(h)} \cap \boldsymbol{C}_r^{(h+1)}|$，其中 r^* 表示第 h 层的 s 节点的父节点，据此建立父子节点关系，这样就形成了一个树状结构，称之为类集成树。形式如图 6 所示。

定义 5　类 K 步稳定系数。对于数据集第 h 次划分中的 $\boldsymbol{C}_s^{(h)}$ 类，其 K 步稳定系数定义为以其为中心连续 k 层对应类别紧密度的均值，即

$$q(h,k) = \frac{\sum_{l=h-\frac{k}{2}}^{h+\frac{k}{2}} A(h,l)}{k}, A(h,l) = \frac{|\boldsymbol{C}_s^{(h)} \cap \boldsymbol{C}_{s^*}^{(l)}|}{|\boldsymbol{C}_s^{(h)}|}$$

$\boldsymbol{C}_{s^*}^{(l)}$ 表示在第 l 层与 $\boldsymbol{C}_s^{(h)}$ 对应的类别，即

$$s^* = \arg\max_r |\boldsymbol{C}_s^{(h)} \cap \boldsymbol{C}_r^{(l)}|$$

$A(h,l)$ 表示 $\boldsymbol{C}_{s^*}^{(l)}$ 与 $\boldsymbol{C}_s^{(h)}$ 的紧密度。若一个类别的 K 步稳定系数为 1，说明该类别在 k 次划分中没有发生变化，该类即可作为最终聚类结果中的一类。

在图 6 所示的类集成树中，若设 k 为 3，点 27 在前 3 次划分中都是稳定的，即点 27 可单独作为一类；点 1 至 6，10 至 23，24 至 26 在 2 到 4 次划分中分别是稳定的，因此各自单独作为一类；而点 7 至 9 在第 3 到 5 次划分是稳定的，可以作为一类。

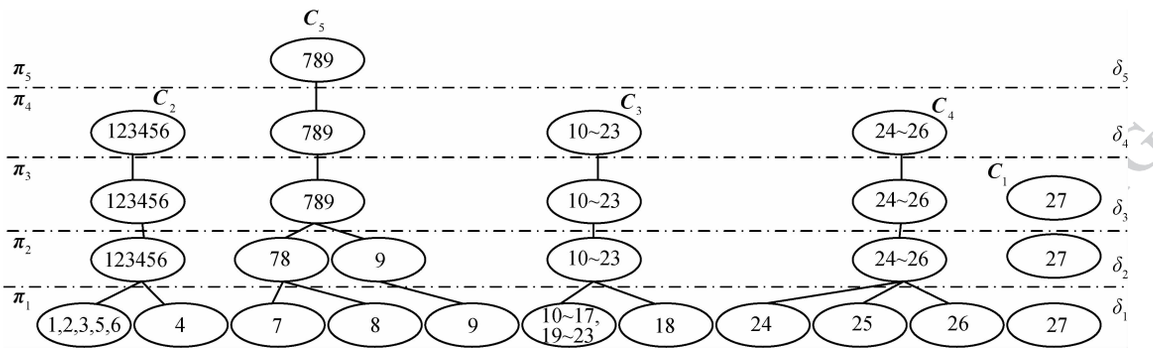


图 6　类集成树示意图

Fig. 6　Illustration of the cluster aggregation tree

定义 6　隔离带宽度，记为 W_{m}。将两个数据集 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 之间的隔离带宽度定义为两个数据集之间的最短距离，即 $W_{\mathrm{m}} = \min_{a\in\boldsymbol{A}} \min_{b\in\boldsymbol{B}} (dist(\boldsymbol{a},\boldsymbol{b}))$，其中 $\boldsymbol{a}$ 和 $\boldsymbol{b}$ 分别是数据集 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 中的元素，$dist(\cdot)$ 表示距离函数。

定义 7　对峙点。若集合 $\boldsymbol{A}$ 中的 $\boldsymbol{a}^*$ 与集合 $\boldsymbol{B}$ 中的点 $\boldsymbol{b}^*$ 之间的距离等于 W_{m}，则将 $\boldsymbol{a}^*$ 和 $\boldsymbol{b}^*$ 称为对峙点。

定理 1　若集合中存在成员个数多于 1 的类，则各类的可达半径一定位于数据集中各点直接可达半径的最小值和最大值之间。

证明：以反证法证明。

若设置的类可达半径小于等于数据集中各点之间的最小距离，则任何一点都不会存在可达点，因此就不会存在多于 1 个点的类。同理，若设置的类可达半径大于等于数据集中各点的直接可达半径的最大值，则集合中各点是其他点的直接可达点，整个数据集变为一类。从而得证。

定理 2　利用 K 步稳定系数对数据集进行正确划分的必要条件是：数据集两个类别之间的隔离带宽度要大于这两个类的可达半径的最大值。

证明：设 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 为构成待划分数据集 $\boldsymbol{X}$ 的两个类别，即 $\boldsymbol{X} = \boldsymbol{A} \cup \boldsymbol{B}$ 且 $\boldsymbol{A} \cap \boldsymbol{B} = \varnothing$。同时设 δ_A 和 δ_B 分别为集合 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的类可达半径。令 δ_{AB} 为 δ_A 和 δ_B 的较大者，即 $\delta_{AB} = \max(\delta_A,\delta_B)$。

若 $W_{\mathrm{m}} \leqslant \delta_{AB}$，则在 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 集合的对峙点 $\boldsymbol{a}^*$ 和 $\boldsymbol{b}^*$ 处，二者一定是直接可达的，从而将 $\boldsymbol{A}$ 与 $\boldsymbol{B}$ 链接形成一类。在图 7 中，网状标记的两个点是 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的对峙点 $\boldsymbol{a}^*$ 和 $\boldsymbol{b}^*$，圆的半径是各点的可达半径。如图 7(a)所示，在 $\boldsymbol{a}^*$ 处，集合 $\boldsymbol{A}$ 的可达半径 δ_A 大于带宽 W_{m}，从而有 $\boldsymbol{a}^*$ 和 $\boldsymbol{b}^*$ 之间的距离小于 $\boldsymbol{a}^*$ 点

的可达半径,从而使 $\boldsymbol{a}^*$ 和 $\boldsymbol{b}^*$ 成为同类点,这样集合 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 就分不开。

同理,若 $W_{\mathrm{m}} > \delta_{AB}$,则 $\boldsymbol{A}$ 中任意点与 $\boldsymbol{B}$ 中任意点的距离都会大于可达半径,从而使 $\boldsymbol{B}$ 中的点都不会在 $\boldsymbol{A}$ 中点的可达范围内,同理 $\boldsymbol{A}$ 中的点也不会在 $\boldsymbol{B}$ 中点的可达范围内,因此两个集合之间存在空隙,具备分开的条件,如图7(b)所示。

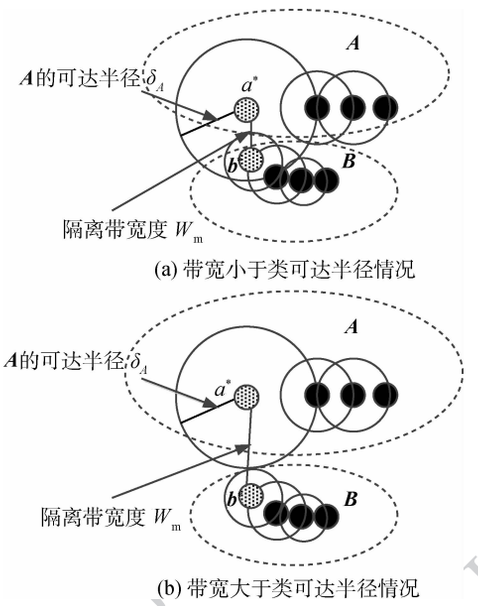


图7 K步稳定聚类方法的必要条件示意图

Fig. 7 Illustration of the necessary condition of the K steps stabilization clustering method ((a) the case that the width of the margin is less than the class reachable radius; (b) the case that the width of the margin is greater than the class reachable radius)

2.3 算法描述

对于给定的鞋印花纹图像集 $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_N\}$,根据上述定义进行聚类的步骤为:

1)计算各幅图像之间的距离,形成距离矩阵 $\boldsymbol{W} \in \mathbf{R}^{N \times N}$。

2)确定对数据集进行多次划分的直接可达半径集合 $\boldsymbol{\delta}$。根据定理1可知,类的可达半径一定位于集合内各点直接可达半径的最小值和最大值之间,因此将 $\boldsymbol{W}$ 中上三角或下三角矩阵(不包括对角线上的元素)中的所有值按由小到大排序,去掉最大值和最小值,取出其余不相同的数据构成可达半径的候选集,即 $\boldsymbol{\delta} = [\delta_1, \cdots, \delta_h, \cdots, \delta_L]$,其中 L 为互不相同数据的个数。

3)构建 k 层类集成树。根据判定类稳定的K步稳定系数中的参数 k,取可达半径集合中的前 k 项,根据定义4对数据集进行 k 次划分,得到 k 层类集成树。

根据给定的可达半径 δ_h,对数据集进行划分的方法有很多,如基于图的方法[15]和基于密度的方法[7]等,本文采用类似图像分割中的区域生长法进行数据集的划分,即:

(1)将鞋印花纹图像集 $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_i, \cdots, \boldsymbol{x}_N\}$ 作为初始未划分数据集,类别序号 s 的初始值为1;

(2)从未划分的数据集中找一个任意点作为种子点,并新建一个包含该点的类别,记为 $C_s^{(h)}$,然后将所有与其直接可达的点合并到种子点所在的类别 $C_s^{(h)}$ 中,并从未划分数据集中删除进入到新类别中的点。

(3)将这些新进入类别 $C_s^{(h)}$ 中的点当作新的种子点继续上面的过程,直到没有满足条件的点可被包括进来。这样就形成一个新的类别。

(4)类别序号 $s = s + 1$。

(5)执行步骤(2),直到未划分数据集为空,即所有数据都归到相应类别中便停止。

4)利用K步稳定系数对数据集进行聚类输出,步骤如下:

(1)从第 l 层节点开始,其中 l 的初始值为1,根据定义5分别计算各类的K步稳定系数。

(2)将满足K步稳定条件的类作为数据集最终划分的一类输出,记为 $\boldsymbol{C}_f$, f 为最终类别编号,初始值为1,每次输出类别时类别编号加1,且从数据集中删除所有已输出类别所包含的点,剩余的未划分的数据集记为 $\boldsymbol{X}_{\mathrm{left}}$。

(3)若存在不满足K步稳定条件的类,则取下一个直接可达半径 δ_{k+l},按步骤3)中的方法对 $\boldsymbol{X}_{\mathrm{left}}$ 进行划分,并将聚类集成树新增一层,且层次序号 l 加1。

(4)转向步骤(1),直到 $\boldsymbol{X}_{\mathrm{left}}$ 为空停止,即所有数据归到相应类别中。

2.4 算法特性分析

鞋印花纹图像的若干特性导致现有的聚类算法效果不佳,对本文算法适用于鞋印花纹图像的特性以及时间和空间复杂度进行如下分析:

1)不需要事先知道聚类的数目。本文算法通过K步稳定的方法依次寻找类别之间的隔离带,从而确定类别且能自动算出类别数目,因此不需要知道聚类的数目。

2)能够有效解决同类数据分布不均匀的问题。

若数据集中有的类成员之间分布地比较紧密,有的类成员之间分布地很稀疏,且各类成员数目不均匀,则现有基于密度的聚类算法在整个数据集上采用同一个密度阈值和距离阈值进行所属类别的判定方法就不会奏效,而本文算法通过由小到大依次遍历数据点之间的直接可达半径,寻找使类成员不发生变化的各类之间的隔离带,从而进行聚类。即各类所用的距离阈值(直接可达半径)是不同的,对分布稀疏的类,采用的直接可达半径较大,而对分布比较紧密的类,采用的直接可达半径比较小,这样,算法就可有效解决数据分布不均匀的问题。

3)本文算法不对花纹在特征空间上的分布形状进行假定。算法判定数据点是否属于某一类采用的是与类内任意一点直接可达方式判定,而不是如 K-means 等算法与中心点直接可达的方式,因此对数据点的分布形状没有如球形假设等限制。

4)不受鞋印花纹种类以及同类花纹数目限制。算法没有设定密度阈值,单独点亦可称为一类,因此对各类成员数目没有要求。另外本算法采用的是 K 步稳定方法确定类别,只要类别之间存在隔离带,均可得到很好的聚类效果。因此不会受种类数目影响。

5)算法时间和空间复杂度分析。算法各步的时间和空间复杂度分析如下:

(1)计算数据集上各点之间的距离矩阵 $\boldsymbol{W}$,需要进行 $(N-1)\times N/2$ 次操作,时间复杂度为 $O(N^2)$,距离矩阵需要 $(N-1)\times N/2$ 个存储单元。

(2)确定对数据集进行多次划分的直接可达半径集合 $\boldsymbol{\delta}$,需要对 $(N-1)\times N/2$ 个距离进行排序,采用箱排序的方法,时间复杂度为 $O(N^2)$,可达半径集合需要 $(N-1)\times N/2$ 个存储单元。

(3)构建 k 层类集成树主要进行的操作是将与种子点的距离小于直接可达半径的点确定为与种子点是一类。为提高速度,需要事先将各点与其他点的距离由小到大排序,每次排序的时间复杂度为 $O(N)$,N 个点需要进行 N 次排序,时间复杂度为 $O(N^2)$。根据排序后的数据,查找每个种子点的直接可达点需要的是线性时间。存储各个点的排序矩阵需要 $(N-1)\times N$ 个存储单元,存储每个点的所属类别标识需要 N 个单元。

(4)利用 K 步稳定系数对数据集进行聚类输出只需根据每层节点的对应 K 步稳定系数进行,所需要的时间与参数 K 和该层的节点个数有关,时间复杂度是线性的。

3　实验结果与分析

3.1　实验环境

3.1.1　测试数据集

为了验证本文算法的性能,针对提出的算法进行了 3 类数据集的测试,分别是合成数据集(http://cs.uef.fi/sipu/datasets/)、公共测试数据集(Olivetti Face 数据库,http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html)和鞋印花纹图像数据集。合成数据集一共包括 5 个不同形状的数据,分别反应了不同类型的问题;Olivetti Face 数据库中共有 40 个不同年龄、不同性别和不同种族的对象。每个对象有 10 幅不同表情的图像,整个数据库共计 400 幅人脸图像,本文所用的距离矩阵与文献[16]相同;鞋印花纹图像数据集来源于某市刑警队提供的 5 792枚数据,共包含 3 256 个不同类别的花纹。

3.1.2　比较算法

将本文算法和 3 个典型算法进行比较,分别是谱聚类、DBSCAN 和文献[16]中的方法,程序分别来源于相应文献作者网站。

3.1.3　评价指标

关于聚类性能的评价指标有很多种,本文进行各个算法的比较时主要采用文献[17]定义的 6 种评价指标纯度(*Purity*)、归一化互相关信息(*NMI*)、Rand 指数(*RI*)、F-measure、准确率(P)、召回率(R)和文献[18]定义的另一种形式的 F-measure。F-measure有两种形式,一种形式是根据整体的准确率和召回率计算,本文用 F_1 表示[17];另一种形式是根据每一类实际标签对应的聚类类别的 F-measure 值加权得到,本文用 F_1^* 表示[18]。

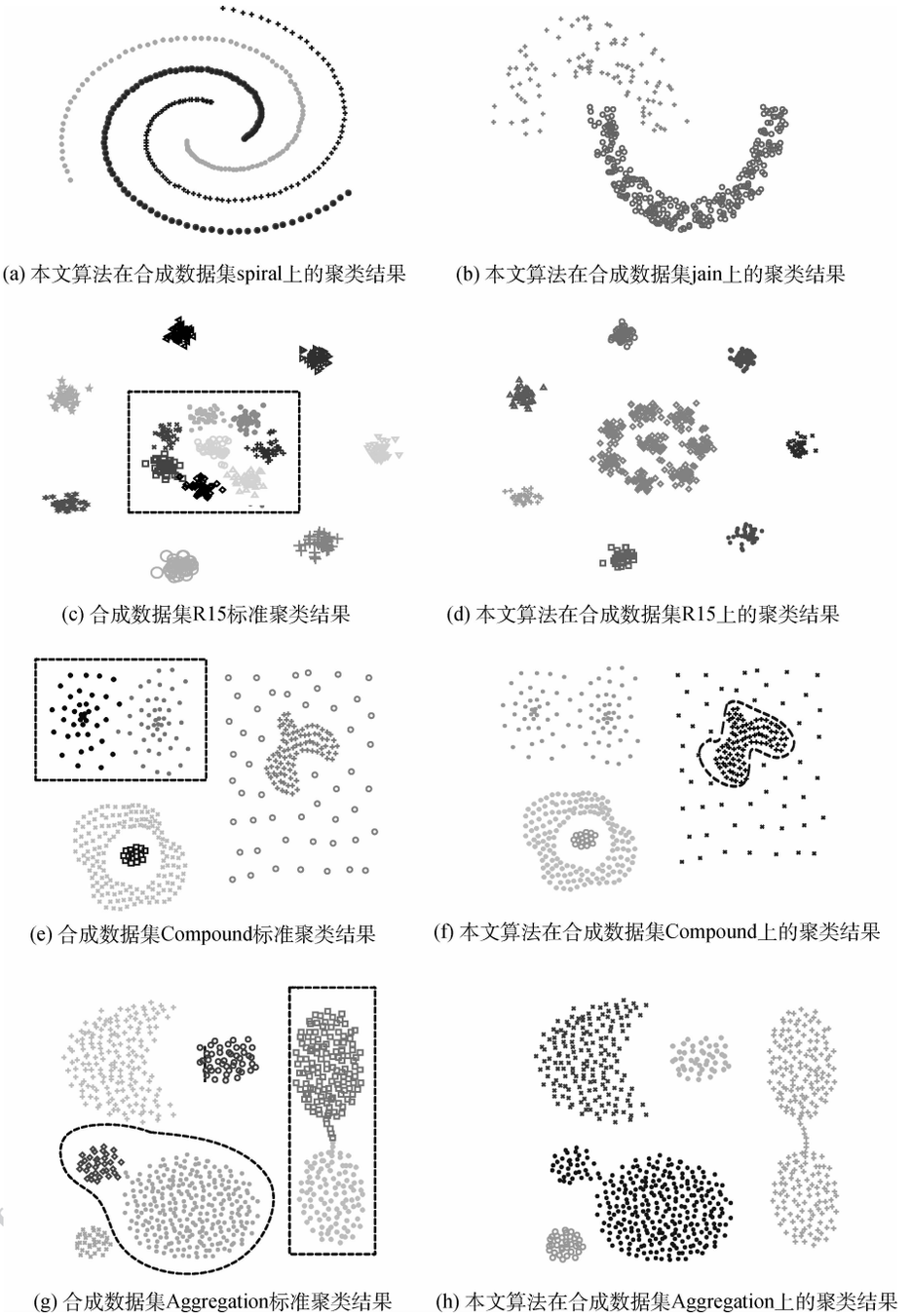
纯度反映的是聚类后的同一类别中同类标记的最大比例;归一化互相关信息从整体上反映了聚类结果和实际标记的一致情况;准确率反映了在聚为同类的数据中标记也为同类的数据所占的比例;召回率反映了在标记为同类的数据中聚类也为同类的比例;Rand 指数、F_1 和 F_1^* 则是折衷反映了准确率和召回率的综合情况,衡量的是聚类算法的综合性能。

这 7 个指标的取值范围都是 0 到 1,并且都是数值越大说明聚类效果越好。

3.2 合成数据集的实验结果与分析

为了说明算法的性能与适用条件,共对5个合成数据集进行实验,实验结果如图8和表1所示。

图8所示的是本文算法在5个典型数据集上的聚类结果。对于图8(a)(b)这种类别之间存在明显隔离带的数据,本文算法可以正确将其聚类;对于如图8(c)(e)(g)中虚线所围的两个类别之间并不存在明显隔离带的数据集,算法聚类结果分别如图8(d)(f)(h)所示,不存在隔离带的类别会被聚成一类;但只要存在隔离带即使两个类的密度以及分布情况不一致也可以准确聚类,如图8(f)中,虚线中的"+"型标记的类别与其周围环绕"×"型标记的类别,尽管在密度和分布上都存在很大差异,本文算法也能将其分开。



(a) 本文算法在合成数据集spiral上的聚类结果　(b) 本文算法在合成数据集jain上的聚类结果

(c) 合成数据集R15标准聚类结果　(d) 本文算法在合成数据集R15上的聚类结果

(e) 合成数据集Compound标准聚类结果　(f) 本文算法在合成数据集Compound上的聚类结果

(g) 合成数据集Aggregation标准聚类结果　(h) 本文算法在合成数据集Aggregation上的聚类结果

图8　本文算法在5种合成数据集上的聚类结果

Fig. 8　Clustering results of the proposed algorithm on five kinds of synthetic datasets ((a) clustering result on the spiral dataset; (b) clustering result on the jain dataset; (c) the ground truth on the R15 dataset; (d) clustering result on the R15 dataset; (e) the ground truth on the Compound dataset; (f) clustering result on the Compound dataset; (g) the ground truth on the Aggregation dataset; (h) clustering result on the Aggregation dataset)

表 1 所示的是各算法在 5 种合成数据集上的聚类结果的平均性能指标。可以看出,本文算法的性能总体要高于谱聚类和 DBSCAN 方法和文献[16]的算法相比,本文算法的纯度、准确率和 F-measure 的平均指标略低,原因在于将不存在隔离带的类别聚成了一类,如图 8 中(c)(e)(g)所示,这样使算法的纯度、准确率和 F-measure 指标降低。若不考虑无隔离带的情况,本文算法均能给出完全正确的聚类结果。

综合图 8 和表 1,说明本文算法能够有效解决同类数据分布不均匀的问题,且不对花纹在特征空间上的分布形状进行假定。

表 1　4 类算法在合成数据集上的平均性能

Table 1　Average performance of four kinds of algorithms on the synthetic datasets

/%

指标项	文献[16]算法	谱聚类	DBSCAN	本文算法
Purity	92.50	63.07	82.55	82.80
NMI	82.21	45.38	86.05	90.71
RI	90.92	74.68	90.84	91.97
P	87.85	53.93	73.31	74.25
R	86.20	53.56	98.49	100
F_1	86.96	52.07	79.90	81.22
F_1^*	90.96	60.10	83.06	85.11

3.3　公共测试数据集的实验结果与分析

采用本文算法和比较算法在 Olivetti Face 数据库上分别进行聚类,计算的性能指标如表 2 所示。其中文献[16]的各项指标是按论文中所给的参数计算的,谱聚类和 DBSCAN 的指标是根据遍历各个参数所获的最好结果统计的。从表中可以看出本文算法明显要优于谱聚类和 DBSCAN 方法,只在召回率指标上低于文献[16]算法,原因在于本文算法聚出类的数目多于文献[16]设置的聚类数目,出现将同类数据给分开的情况,造成召回率降低。

表 2　4 类算法在 Olivetti Face 数据集上的平均性能

Table 2　Average performance of four kinds of algorithms on the Olivetti Face dataset

/%

指标项	文献[16]算法	谱聚类	DBSCAN	本文算法
Purity	58.75	51.5	35.25	86
NMI	81.05	69.48	57.06	87.04
RI	96.08	96.47	57.1	98.47
P	32.97	27.88	4.12	74.55
R	71.56	35.72	80.83	48.67
F_1	45.14	31.32	7.83	58.89
F_1^*	61.28	50.49	33.6	74.66

文献[16]也给出了前 100 幅图像进行聚类的结果,如图 9(a)所示,其中相同的颜色代表聚到同一类中,灰色的图是没有给出相应的聚类类别的图像。为了便于比较,本文算法也在同样数据上进行了聚类,结果如图 9(b)所示。对比图 9(a)(b)所示的聚类结果可以看出,本文算法的结果会将所有的图像给出相应的聚类类别;大多数类别中的图像都是同一人的脸,并且有 4 类将同一人的 10 幅图像完全聚到一类中;只有两组把不是同一类的脸合并成一类的情况。因此本文算法的效果要明显优于文献[16]。

3.4　实际鞋印花纹图像数据集的实验结果与分析

采用本文方法与其他 3 种典型方法分别对鞋印花纹图像数据集进行聚类,并根据 7 种评价指标来比较本文算法和其他算法的性能。各类方法的参数设置如下:

1)谱聚类属于基于划分的聚类方法,实验中将计算相似度值的高斯函数的标准差 σ 设为 0,聚类数目根据实际标签情况设置,即针对鞋印花纹图像数据集将聚类数目设成 3 256。

2)DBSCAN 是基于密度的聚类方法的典型代表,将其密度阈值 MinPts 设为 1,对距离参数 ε 则从小到大进行遍历,从各次聚类结果中选效果最好的那次结果进行比较(即 ε = 0.3 时的实验结果)。

3)文献[16]中的方法需要根据生成的决策图进行选定类别中心,鞋印花纹图像数据集的决策图如图 10 所示,该方法是通过选取决策图中 ρ(点的局部密度)和 δ(点与高密度点的距离)都较大的数据点作为类别中心进行聚类的。从图 10 可以看出,除了右上角一个点外,没有其他明显满足条件的点。由于实验是在已标记的数据集上进行的,类的数目已知,因此为了获得该算法的最好效果,从决策图的右上角开始向左下角依次选 3 256 ± 100 个点作为中心进行聚类,选择性能指标最好的结果作为最终结果比较。选择的类中心范围如图 10 中方框部分所示,此时聚类数目为 3 312。

(a) 文献[16]算法结果

(b) 本文算法结果

图 9　Olivetti Face 数据库结果对比图

Fig. 9　Performance results of both the algorithm in reference [16] and the proposed algorithm on the Olivetti face database ((a) performance result of the algorithm in reference [16] on the Olivetti Face database; (b) performance result of the proposed algorithm on the Olivetti face database)

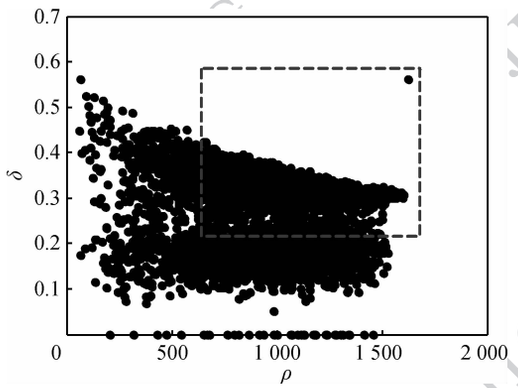


图 10　文献[16]算法的决策图及聚类中心选择示意图

Fig. 10　Illustration of the decision graph of the algorithm in reference[16] and the chosen central points of the clusters

表 3 所列的是各算法的性能指标，从中可以看出，本文算法的各项指标均明显优于其他算法，具体分析如下：

1）本文算法聚出的类别数目和真实类别数目比较接近。本文算法聚出的类别数目与真实类别数目差 172 类，偏离 5.28%；而 DBSCAN 算法聚的类别数目差了 725 类，偏离 22.3%；文献[16]聚类数目多了 56 个，偏离为 1.7%；谱聚类算法需要事先设置类别数目，我们直接设成真实类别数目。若仅从类别数目偏离程度上来说，本文算法不如文献[16]算法，但文献[16]的类别数目是在已知真实类别数目的前提下，在其上下范围内，遍历选择聚类效果最好的数目为最终的结果，仅从原算法的决策图上是无法设置类别数目的。因此本文算法要优于其他算法。

2）本文算法的纯度、归一化互相关信息、Rand 指数指标都远远超过比较算法，并且比较算法的指标还是在参数进行过多次调整后选择的最好结果。

3）本文算法的准确率指标远超过比较算法，但召回率指标低于 DBSCAN 算法。这是因为本文算法把两幅图像成功分成一类的概率要远大于比较算法，但也可能把同类图分开；而 DBSCAN 易把不是一类的合成一类。

4）本文算法的 F-measure 值要远大于其他算法。由于本文算法的准确率要远大于其他算法，但召回率略低于 DBSCAN 算法，因此衡量二者综合性能的 F-measure 指标要远大于比较算法。表中 F_1 是根据聚类的整体结果的准确率和召回率计算的 F-measure 值，比较算法的准确率都很低，尤其 DBSCAN算法的准确率最低，仅为 2.69%，所以比较算法的 F_1 值都很低，其中 DBSCAN 算法的 F_1 值最低。表 3 中 F_1^* 指标是根据与每一类结果的 F-measure

值与该类成员所占比例的加权和得到。在这种基于每一类结果的 F-measure 计算的指标上，各算法都有所提高，尤其 DBSCAN 提高的幅度很大，但也不如本文算法。

表 3　4 类算法在鞋印花纹图像数据集上的性能

Table 3　Performance of four kinds of algorithms on the real shoeprint image dataset

/%

指标项	算法(类别数目)			
	文献[16](3 312)	谱聚类(3 256)	DBSCAN(2 531)	本文(3 428)
Purity	73.24	66.49	76.67	99.71
NMI	91.61	90.19	91.29	99.16
RI	98.09	99.86	96.19	99.98
P	25.87	13.72	2.69	99.68
R	22.92	4.38	97.87	85.10
F_1	24.31	6.64	5.23	91.81
F_1^*	56.60	54.02	75.58	95.99

从表 3 也可以看出，文献[16]算法、谱聚类算法以及 DBSCAN 算法在鞋印花纹图像数据集上的性能指标非常低，主要是由于鞋印花纹图像各类特征分布的不均匀，现有算法存在大量的将同类的分成多类或多类合并成一类的情况。一类分成多类会使算法的召回率和准确率降低。多类合成一个类会使算法纯度和准确率降低。从而 F_1 的值因受召回率和准确率下降的影响也会相应降低。具体分析如下：

1)文献[16]算法将局部密度 ρ_i 和到高局部密度点的距离 δ_i 都很大的点作为类的中心，但由于鞋印花纹图像各类分布的不均匀，有的类密度高，有的类密度低，因此该算法会将分布比较紧凑的类边缘部分的点分出单独作为一类，因为这些点的局部密度也相对较大且与高密度点的距离也大；而对于分布比较稀疏的类，其中各点的局部密度比较小，这些点会被分到密度高的其他类中去。

2)谱聚类采用的是图划分理论，划分准则选取的是同类之间的相似性和最大，类间的相似性和最小。因此只有当同一类中两两图像相似性都大时才会取得很好的效果，对于本文鞋印花纹图像的同类数据分布不均匀的现象不能给出很好的聚类。

3)DBSCAN 算法需要两个参数，一个是密度阈值，另一个是判定两幅图像为一类的距离阈值。由于鞋印花纹图像分布的不均匀，因此在整个数据集上采用一个阈值会出现大量将类分开或合并的现象，从而造成准确率和召回率下降的情况。

3.5　参数设置对算法性能影响的实验结果与分析

通过改变参数 K，比较每次聚类的指标的变化以分析算法参数对性能的影响，测试数据集为包含 5 792 枚鞋印花纹图像集。K 的取值，从 10 开始一直取到 25。表 4 所列的是部分指标的统计结果，其中类别数目偏离度是指聚出的类的数目与真实类数目之差所占真实类数目的百分比。图 11 所列的是各指标随着 K 变化的示意图。从表 4 和图 11 可以看出，本文算法的性能受参数 K 的影响，但不敏感。

表 4　算法参数对性能指标的影响

Table 4　The effects of parameters on the performance of the proposed algorithm

/%

指标项	K/(类别数目)			
	10/(3 588)	15/(3 499)	20/(3 428)	25/(3 373)
类别数目偏离度	10.20	7.46	5.28	3.59
Purity	99.78	99.78	99.71	97.72
NMI	98.56	98.91	99.16	98.93
RI	99.97	99.98	99.98	99.94
P	99.76	99.77	99.68	64.63
R	72.92	80.18	85.10	88.90
F_1	84.26	88.91	91.81	74.85
F_1^*	93.16	94.81	95.99	94.91

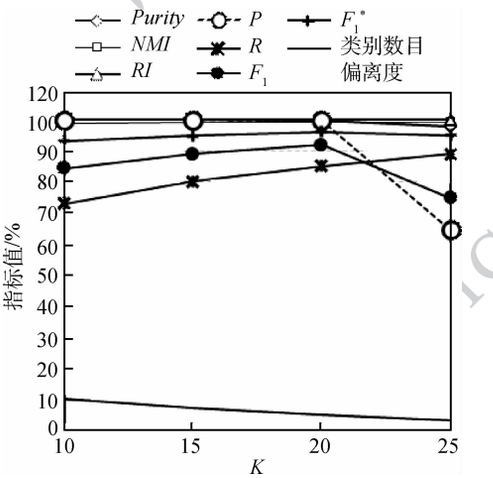


图 11　算法参数对性能指标的影响示意图

Fig. 11　Illustration of the effects of varying parameters on the performance of the proposed algorithm

在本文中将指标的变化量与参数 K 的变化量的比值称为敏感度。具体分析如下：

1）随着 K 值的增大，本文算法聚出的类数目在减少，与真实类别数目的偏离度会出现先减小后增大的现象。原因在于：较小的 K 值能将较窄的隔离带分开，聚出的数目多；随着 K 值增大，K 次聚类会越过隔离带，将多个类合并成一个类，聚出的类数目逐渐减小，因而会出现先靠近真实类数目，然后再远离真实数目的现象。但聚类数目的敏感度都低于 0.55%，这说明算法性能对参数的变化并不敏感。另外图 11 中最下方的类别数目偏离度直线的斜率很小也说明了这个问题。

2）本文算法的纯度、归一化互信息、Rand 指数和 F_1^* 这 4 项指标受参数变化影响较小。图 11 所示的这 4 条曲线变化幅度很小。

3）当 K 值位于 10 到 20 时，算法所有指标受参数影响较小。当 K 值超过 20 时，聚类准确率下降，原因在于算法将不在一个类别的数据分成一类，从而准确率下降，而召回率上升，因此 F_1 值也要下降。其中准确率的敏感度的最大值为 7.01%，召回率的敏感度最大值为 3.39%。图 11 中准确度和 F_1 值曲线在 K 大于 20 时都开始下降。

通过上述分析可知，参数对本文算法性能影响较小，但并不说明算法的参数可以任意设置，必须满足 K 次稳定的阈值变化长度在隔离带的范围内才能得到较好的稳定的聚类结果。

3.6 种子点的选取对结果的影响

为了验证 2.3 节算法步骤 3）中不同种子点对聚类效果的影响，进行了 10 次聚类实验。每次初始种子点是在 N 个数据点中以正态分布形式随机选取的不同点，分别计算聚类结果的各项指标，如图 12 所示。可以看出，各个指标完全没有变化，这说明本文算法的结果并不受种子点选取的影响。

3.7 不满足适用条件的数据集的聚类结果与分析

根据定理 2 可知，本文算法适用于类别之间存在隔离带的数据集。但在实际应用时，有时并不知道该数据集各类之间是否具有隔离带。对于不存在隔离带的数据集采用本文算法进行划分，会出现的现象及判定方法如下：

1）若整个数据集各类之间都不存在隔离带，则本文算法会将所有的数据分到一类中，因此可根据算法是否将所有数据分成一类判定本文方法适用于该数据集与否。若以图 8（c）中 R15 数据集中间的 8 类数据作为算法的输入，算法会将所有数据分成一类，则可直接断定方法不适用于该数据集；

2）若只有少数几个类别间不存在隔离带，此时算法会把不存在隔离带的类合并成一类，但合并成的类成员个数不会特别多，则需要用引入新的先验知识和其他衡量类内一致性的方法来判断其是否有效或者根据应用直接认定其有效。如图 8（e）中方框所围的数据、图 8（g）中曲线所围的数据以及矩形所围的数据聚成单独一类也有一定的道理。

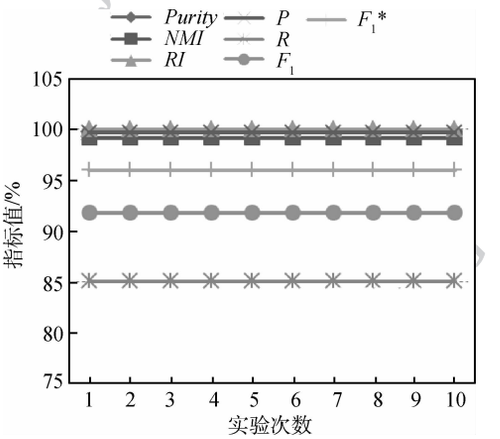


图 12 选取不同种子点的聚类结果示意图

Fig. 12 Illustration of the clustering results on different seed points

4 结 论

在分析鞋印花纹图像的特点基础上提出了一种 K 步稳定的鞋印花纹图像自动聚类算法。该算法的特点是通过自动寻找类别之间的隔离带来进行聚类。在两类公共数据集和实际鞋印花纹图像数据集上的实验结果表明，本文算法在使用一个不敏感的参数的情况下，各项聚类指标都要优于目前典型的聚类算法。但并不是所有数据集都满足本文算法的应用条件，接下来的研究工作将重点考虑对于不满足类别之间存在明显隔离带的数据集的聚类问题以及参数 K 的自适应设置问题。

参考文献（References）

[1] Bodziak W J. Footwear Impression Evidence: Detection, Recovery, and Examination[M]. New York: CRC press, 2000.

[2] Jain A K. Data clustering: 50 years beyond K-means[J]. Pattern Recognition Letters, 2010, 31(8): 651-666. [DOI: 10.1016/j.patrec.2009.09.011]

[3] Ng R T, Han J W. CLARANS: a method for clustering objects for spatial data mining[J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14(5): 1003-1016. [DOI: 10.1109/TKDE.2002.1033770]

[4] Jain A K, Murty M N, Flynn P J. Data clustering: a review[J]. ACM Computing Surveys, 1999, 31(3): 264-323. [DOI: 10.1145/331499.331504]

[5] Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data. New York, NY: Association for Computing Machinery, 1996: 103-114. [DOI: 10.1145/233269.233324]

[6] Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for large databases[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data. Seattle, WA, USA: Association for Computing Machinery, 1998: 73-84. [DOI: 10.1145/276304.276312]

[7] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. California, USA: Association for Computing Machinery, 1996: 226-231.

[8] Ankerst M, Breunig M M, Kriegel H P, et al. OPTICS: ordering points to identify the clustering structure[C]//Proceedings of the ACM SIGMOD International Conference on Management of Data. Philadelphia PA, USA: Association for Computing Machinery, 1999: 49-60. [DOI: 10.1145/304182.304187]

[9] Xu R, Wunsch II D. Survey of clustering algorithms[J]. IEEE Transactions on Neural Networks, 2005, 16(3): 645-678. [DOI: 10.1109/TNN.2005.845141]

[10] Wang W, Yang J, Muntz R R. STING: a statistical information grid approach to spatial data mining[C]//Proceedings of the 23rd International Conference on Very Large Data Bases. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 1997: 186-195.

[11] Sheikholeslami G, Chatterjee S, Zhang A D. WaveCluster: a wavelet-based clustering approach for spatial data in very large databases[J]. The VLDB Journal, 2000, 8(3-4): 289-304. [DOI: 10.1007/s007780050009]

[12] Fisher D H. Knowledge acquisition via incremental conceptual clustering[J]. Machine Learning, 1987, 2(2): 139-172. [DOI: 10.1007/BF00114265]

[13] Strehl A, Ghosh J. Cluster ensembles-a knowledge reuse framework for combining partitionings[J]. Journal of Machine Learning Research, 2002, 3: 583-617.

[14] Wang X N, Sun H H, Yu Q, et al. Automatic shoeprint retrieval algorithm for real crime scenes[C]//Proceedings of the 12th Asian Conference on Computer Vision. Singapore: Springer, 2015: 399-413. [DOI: 10.1007/978-3-319-16865-4_26]

[15] Xu X W, Yuruk N, Feng Z D, et al. SCAN: a structural clustering algorithm for networks[C]//Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Jose, California, USA: Association for Computing Machinery, 2007: 824-833. [DOI: 10.1145/1281192.1281280]

[16] Rodriguez A, Laio A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496. [DOI: 10.1126/science.1242072]

[17] Manning C D, Raghavan P, Dan Schütze H. An Introduction to Information Retrieval[M]. Cambridge, England: Cambridge University, 2009: 356-360.

[18] Larsen B, Aone C. Fast and effective text mining using linear-time document clustering[C]//Proceedings of the fifth ACM SIGKDD international conference on Knowledge Discovery and Data Mining. New York, NY: Association for Computing Machinery, 1999: 16-22. [DOI: 10.1145/312129.312186]