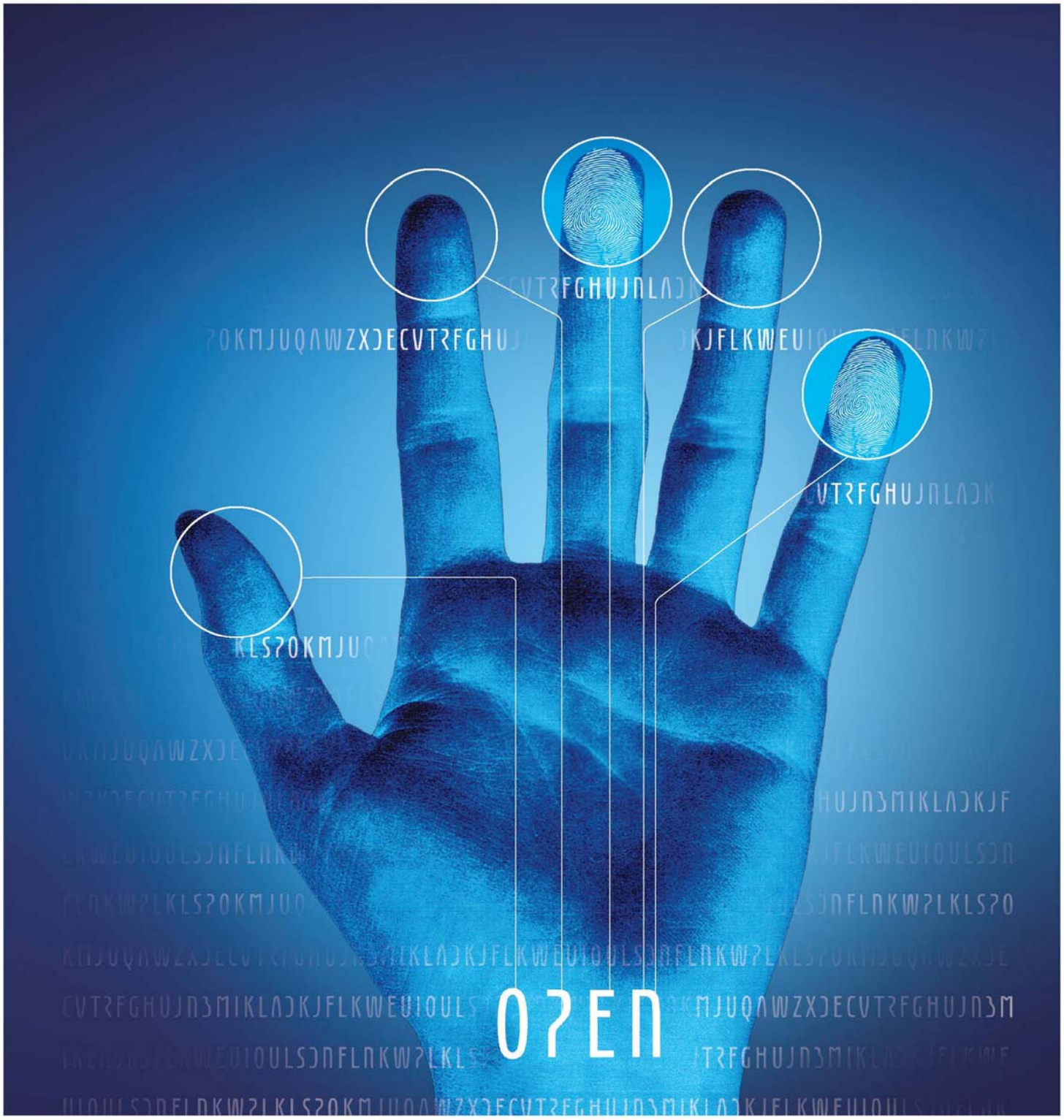
JIG
JOURNAL OF IMAGE AND GRAPHICS
主办：中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2013
07
VOL.18
ISSN1006-8961
CN11-3758/TB



中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第18卷第7期（总第207期）
2013年7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

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@irsa.ac.cn
电　　话 010-64807995 010-82614429
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
国外发行 中国国际图书贸易总公司
（中国国际书店，邮政信箱：北京399信箱　邮编：100044）
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@irsa.ac.cn
Telephone 010-64807995 010-82614429
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Foreign China International Book Trading Corporation
(P.O.Box 399, Beijing 100044,P.R.China)
Printed by Beijing Kexin Printing Co., Ltd.

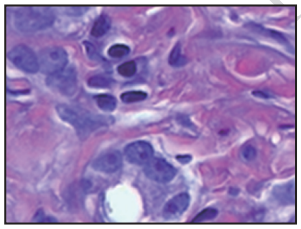
CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 50.00元

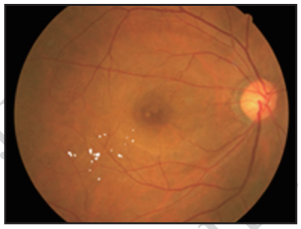
2013年7月 第18卷 第7期（总第207期）
中国图象图形学报 Zhongguo Tuxiang Tuxing Xuebao

目次

彩色纹理图像恢复的非局部TV模型（第753页）



基于Nyström密度值逼近的减法聚类（第790页）

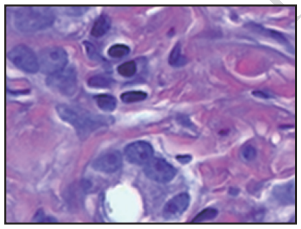


RBF神经网络和阈值分割实现视网膜硬性渗出自动检测（第859页）

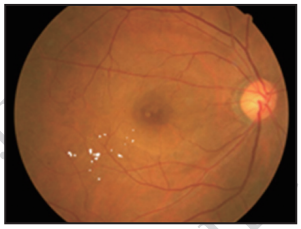
Volume 18, Number 7
Published July 16, 2013

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Non-Local TV models for restoration of color texture images (P753)



Density value approximation for subtractive clustering based on Nyström method (P790)



Automatic detection of hard exudates based on RBF neural network and threshold segmentation (P859)

E-mail: jig@irsa.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：TP391.41　　**文献标识码**：A　　**文章编号**：1006-8961(2013)07-0790-09
论文引用格式：孙志海，孔万增．基于 Nyström 密度值逼近的减法聚类[J]．中国图象图形学报，2013，18(7)：790-798.[DOI：10.11834/jig.20130707]

基于 Nyström 密度值逼近的减法聚类

孙志海，孔万增
杭州电子科技大学计算机学院，杭州　310018

摘　要：针对大规模数据集减法聚类时间复杂度高的问题，提出一种基于 Nyström 密度值逼近的减法聚类方法。特别适用于大规模数据集的减法聚类问题，可极大程度降低减法聚类的时间复杂度。基于 Nyström 逼近理论，结合经典减法聚类样本密度值计算的特点，巧妙地将 Nyström 理论用于减法聚类未采样样本之间密度权值矩阵的逼近，从而实现了对所有样本的密度值逼近，最后沿用经典减法聚类修正样本密度值的方法，实现整个减法聚类过程。将本文算法在人工数据、标准彩色图像及 UCI 数据集上进行了实验，详细说明了本文算法利用少数采样样本逼近多数未采样样本密度权值、密度值以及进行减法聚类的详细过程，并给出了聚类准确率、耗时及算法性能加速比。实验结果表明，与经典的减法聚类相比，本文算法在不影响聚类结果的情况下，对于较大规模数据集，可显著降低减法聚类的时间复杂度，极大程度地提高减法聚类的实时性能。
关键词：减法聚类；Nyström；密度值逼近；时间复杂度

Density value approximation for subtractive clustering based on Nyström method

Sun Zhihai, Kong Wanzeng
College of Computer Science, Hangzhou Dianzi University, Hangzhou 310018, *China*

Abstract: Subtractive clustering based methods have been well known for data clustering problems. However, due to the computational demands of these approaches, clustering for large scale datasets, such as spatio-temporal data and images, have been slow to appear. A novel subtractive clustering method based on Nyström approximation is proposed. The proposed method is based on the famous Nyström method. Combined with the density value computation characteristics for each sample of the classical subtractive clustering method, we apply Nyström theory to approximate the density value for each data point which has not been sampled. Finally, we complete the whole clustering procedure using classical subtractive clustering method in modifying the density values in each circulation. The proposed method substantially reduces the computational requirements of subtractive clustering based algorithms, making it feasible to use subtractive clustering to large scale subtractive clustering problems. Density value of samples could be approximated quickly using only a small number of samples. The experiment results on artificial datasets, color images, and UCI machine learning repository show efficiency in comparing with classical subtractive clustering method.
Key words: subtractive clustering; Nyström; density value approximation; time complexity

收稿日期：2012-09-14；**修回日期**：2012-11-16
基金项目：国家自然科学基金项目(61102028)；浙江省自然科学基金项目(Y1100086)
第一作者简介：孙志海(1981—　)，男，2009 年于浙江大学获控制理论与控制工程专业博士学位，主要研究方向为模式识别。E-mail：szh@hdu.edu.cn

0 引　言

在机器学习和模式识别领域，聚类是一种进行数据分组的有效方法。聚类算法就是按照一定的相似性对样本进行分组的过程，是一种应用广泛的数据分析工具。聚类方法大致可分为：划分法、层次法、网格法、密度法以及基于模型的方法。目前比较常见的聚类算法有：k 均值聚类、模糊 c 均值聚类、减法聚类[1]、层次聚类以及谱聚类[2]等。k 均值聚类及模糊 c 均值聚类算法在使用时均需要指定聚类个数和初始聚类中心，当聚类数目与真实情况不符时，聚类失败，初始聚类中心的设置也对算法有着很大影响[3]。层次聚类在使用时也需指定类别数，当指定类别数目满足时，算法停止。谱聚类虽能保证收敛至全局最优解，但使用时仍存在聚类个数确定，相似度矩阵构造及大规模数据的聚类问题。减法聚类算法是一种基于密度，用于大致估计聚类中心的相对简单且有效的方法。它在使用时需指定密度权值函数的邻域半径及终止条件系数，谱聚类算法与其类似，高斯核函数尺度参数的选取对聚类效果具有很大影响。终止条件系数则影响着聚类中心数目。然而，与谱聚类[2]不同，减法聚类一般适用于紧凑的超球体流形结构的数据集，不适用于具有任意分布的复杂形式的聚类问题，但由于常常被用于 k 均值或模糊 c 均值聚类的初始中心设置，使得减法聚类算法获得广泛应用。

近几年，减法聚类比较常见的应用是与模糊 c 均值聚类结合[3]，根据具体应用对数据进行聚类并对算法做相应的改进。Bataineh 等人[4]将减法聚类与模糊 c 均值聚类做性能比较，得出减法聚类在模型估计精度上高于模糊 c 均值聚类的结论。Sun 等人[5]将减法聚类与本征间隙结合用于视频运动目标个数的自动确定及目标中心位置的确定，为减法聚类中心个数确定提供了一种新思路，但本征间隙往往不稳定，且算法相对繁琐。Chen 等人[6]提出了一种基于加权平均的减法聚类算法，他们认为减法聚类算法以样本作为选定的聚类中心可能存在误差，故提出在采用传统减法聚类算法初步获取聚类中心后，利用 k 近邻样本的密度值求得加权平均后的聚类中心。Ngo 等人[7]提出一种区间 2 型模糊减法聚类算法，将高斯核函数指数项欧氏距离范数引入模糊因子，进一步模糊化样本密度权值。Kim 等人[8]引入核函数距离测度替换了原减法聚类算法采用范数平方距离的度量。然而，以上这些相对典型的减法聚类改进方法，都未涉及减法聚类针对大规模数据集的时间复杂度问题，即实验所选用的数据集一般都是样本数为几十个的小型样本集，回避了减法聚类时间复杂度的问题。然而，虽然减法聚类与原来的山峰聚类比更简单有效，但减法聚类在构造样本的密度值指标时，其算法时间复杂度为 $O(d^2N^2)$，即减法聚类所有样本密度指标随着样本维数 d 及样本个数 N 的平方乘积递增，对于大规模的聚类问题（如视频、图像分割），计算量往往难以接受，从而限制了减法聚类算法在大规模数据聚类、图像及视频分割等方面的应用。

在对减法聚类算法研究过程中发现：减法聚类算法在建立所有样本密度值指标时，与谱聚类算法具有相似之处，即都需要计算样本两两之间的某种相关系数。谱聚类算法通过基于高斯核函数的相似度矩阵构建度矩阵以及拉普拉斯矩阵，通过分析拉普拉斯矩阵的主特征向量实现聚类[2]。而减法聚类则通过所有样本间基于欧拉指数函数的密度权值累积构建样本的密度值指标，通过聚类中心邻域密度指标衰减逐步获取样本集的聚类中心。而谱聚类的时间复杂度则更高，为 $O(d^2N^3)$。Belongie 等人[9]最早提出用于谱聚类的 Nyström 逼近方法，即采用少量样本求解特征向量，然后将其特征向量扩展到所有样本集合相似度矩阵的特征向量。受 Belongie等人[9]工作的启发：在 Nyström 逼近谱聚类主特征向量过程中首先需要对相似度矩阵进行逼近。结合减法聚类与谱聚类的共性，同样可以采用少量样本实现对减法聚类所有样本密度指标的逼近，从而设计了一种新的基于 Nyström 密度值逼近的减法聚类算法。本文算法在 UCI 数据集及标准测试图像的聚类实验中取得了良好效果，并体现出了快速的聚类特性。

1 考虑不同维度邻域的减法聚类

减法聚类其实是山峰聚类的一种替代[1]。在该方法中，聚类中心的候选集为样本点而非网格点。考虑 d 维空间的 N 个样本点 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N$，不失一般性，假定所有样本点均已归一化到一个超立方体中。由于每个样本点都是聚类中心的候选者，定义对角阵

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 18, No. 7, July 2013

$$\boldsymbol{M}_a=\begin{bmatrix} r_{a1}^2 & 0 & \cdots & 0 \\ 0 & r_{a2}^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & r_{ad}^2 \end{bmatrix}$$

$$\boldsymbol{M}_a^{-1}=\begin{bmatrix} 1/r_{a1}^2 & 0 & \cdots & 0 \\ 0 & 1/r_{a2}^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1/r_{ad}^2 \end{bmatrix}$$

$r_{ak}(k=1,2,\cdots,d)$为正常数,为样本点不同维度的邻域半径。因此,样本点$\boldsymbol{x}_i$处的密度指标$D(\boldsymbol{x}_i)$定义为

$$D(x_i)=\sum_{j=1}^{N}\mathrm{e}^{-4(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{M}_a^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j)} \tag{1}$$

显然,如果某个样本点有多个邻近的样本点,则该样本点具有高密度值。半径r_{ak}定义了该样本点在不同维度一个邻域,即影响半径。

在获得这N个样本点的密度值后,选择最大密度值点$\boldsymbol{x}_{c1}$为第1个聚类中心,$D(\boldsymbol{x}_{c1})$为对应的最大密度值。为找出新的聚类中心,需对每个样本点的密度值进行衰减,即在每个样本点$\boldsymbol{x}_i$的密度值$D(\boldsymbol{x}_i)$上减去一定的密度值[1],即

$$D(\boldsymbol{x}_i)\leftarrow D(\boldsymbol{x}_i)-D(\boldsymbol{x}_{c1})\cdot\mathrm{e}^{-4(\boldsymbol{x}_i-\boldsymbol{x}_{c1})^{\mathrm{T}}\boldsymbol{M}_b^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_{c1})} \tag{2}$$

式中,$\boldsymbol{M}_b$为另一对角阵,$r_{bk}(k=1,2,\cdots,d)$为正常数,定义了密度值在不同维度显著衰减的邻域,一般$r_{bk}=\eta r_{ak}$,η为大于1的正常数,η的经验值为$1.25\leqslant\eta\leqslant1.5$,这样可避免出现相离很近的聚类中心。在修正了每个数据点密度值后,将具有最大密度值的数据点选为第2个聚类中心$\boldsymbol{x}_{c2}$。该过程不断迭代,直到找出所有的聚类中心。减法聚类时可以$D_c^*<\varepsilon\cdot D_1^*(0<\varepsilon\leqslant1)$作为迭代终止条件,$D_1^*$为第1个聚类中心密度值,$D_c^*$为第$c$个聚类中心密度值。式(1)中的$r_{ak}$也称为窗宽参数,它的取值对减法聚类算法的性能有很大的影响。怎么选取该参数,仍然是个公开难题。ε参数则影响着聚类中心个数,如何自动获得聚类中心数也是减法聚类有待解决的问题。一般情况下,ε经验值为0.5。

2 Nyström样本密度值逼近

第1节的减法聚类算法在计算每个样本的密度值前,需先计算每个样本点与N个样本点基于欧拉指数函数的密度权值,该权值定义与谱聚类中样本点之间相似度的定义近似。即减法聚类N个样本点两两之间的密度权值组成一个密度权值矩阵,且该矩阵的所有元素都是正的并具有对称特性,而且是一个正定矩阵,这是本文算法的前提条件。所以,如引言所述,减法聚类在计算样本的密度值时其时间复杂度为$O(d^2N^2)$,对于大规模聚类问题,耗时往往难以接受。因此,引入Nyström方法以降低减法聚类的时间复杂度,提出了一种基于Nyström密度值逼近的适合快速处理大规模数据聚类问题的新型减法聚类算法。

2.1 Nyström样本相关系数加权

Nyström是一种专门用于寻找积分特征函数问题的数值近似技术[9-11],即如式(3)的特征函数求其数值逼近解。

$$\int_0^1 G(x,y)\psi(y)\mathrm{d}y=\lambda\psi(x) \tag{3}$$

式中,$\psi(x)$表示特征函数,$G(x,y)$是x与y的某种关系函数,可理解为类似式(1)中的欧拉指数函数。利用定积分求解原理,假设$x_1,x_2,\cdots,x_N$是在[0,1]区间上的N个数据点,在这N个数据点里等间隔采样n个数据点记为$y_1,y_2,\cdots,y_n(n\ll N)$,用简单的积分求解方式逼近式(3),得

$$\frac{1}{n}\sum_{j=1}^{n}G(x,y_j)\hat{\psi}(y_j)=\lambda\hat{\psi}(x) \tag{4}$$

式中,$\hat{\psi}(x)$为真实$\psi(x)$的逼近。为了进一步求解式(3),取$x=y_i$代入式(4),并设对应的λ值为λ_i,得

$$\frac{1}{n}\sum_{j=1}^{n}G(y_i,y_j)\hat{\psi}(y_j)=\lambda_i\hat{\psi}(y_i) \tag{5}$$

结合式(5),则对于所有的$i\in\{1,2,\cdots,n\}$,可获得

$$\begin{bmatrix} G(y_1,y_1) & G(y_1,y_2) & \cdots & G(y_1,y_n) \\ G(y_2,y_1) & G(y_2,y_2) & \cdots & G(y_2,y_n) \\ \vdots & \vdots & & \vdots \\ G(y_n,y_1) & G(y_n,y_2) & \cdots & G(y_n,y_n) \end{bmatrix}\cdot\begin{bmatrix} \hat{\psi}(y_1) \\ \hat{\psi}(y_2) \\ \vdots \\ \hat{\psi}(y_n) \end{bmatrix}=$$

$$n\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}\cdot\begin{bmatrix} \hat{\psi}(y_1) \\ \hat{\psi}(y_2) \\ \vdots \\ \hat{\psi}(y_n) \end{bmatrix} \tag{6}$$

$$\boldsymbol{K}\hat{\boldsymbol{\psi}}=n\boldsymbol{\Lambda}\hat{\boldsymbol{\psi}} \tag{7}$$

式中,$K_{ij}=G(y_i,y_j)$,$\hat{\boldsymbol{\psi}}=[\hat{\psi}(y_1),\hat{\psi}(y_2),\cdots,\hat{\psi}(y_n)]^{\mathrm{T}}$是与矩阵$\boldsymbol{K}$的特征值$[\lambda_1,\lambda_2,\cdots,\lambda_n]$相对

应的 n 个特征向量，$\boldsymbol{\Lambda}=\mathrm{diag}\{\lambda_1,\lambda_2,\cdots,\lambda_n\}$ 即以矩阵 $\boldsymbol{K}$ 的特征值为对角元素的对角阵。将式(7)反代入式(4)，则可产生对每个 $\hat{\psi}_i(x)$ 的 Nyström 扩展形式[9]，即

$$\hat{\psi}_i(x)=(\lambda_i n)^{-1}\sum_{j=1}^{n}G(x,y_j)\hat{\psi}_i(y_j) \tag{8}$$

式(8)表示任一未被采样点 x 的第 i 个特征向量可由函数 $G(x,y_j)$ 对应的相关系数与 $\hat{\psi}_i(y_j)$ 加权获得逼近特征向量。

2.2　减法聚类样本密度值逼近

Nyström 利用样本相关系数加权近似估计未采样样本的特征向量。为了进一步解释如何将这种逼近的特征向量用于减法聚类未采样样本的密度值逼近，假设原减法聚类样本 (x_i,x_j) 两两之间基于欧拉指数函数的密度权值记为 d_{ij}，由所有 d_{ij} 组成的密度权值矩阵用 $\boldsymbol{G}$ 表示。为使用 Nyström 方法，假设从待聚类的 N 个样本中等间隔采样 n 个样本($n\ll N$)，剩余样本数为($N-n$)。则可把密度权值矩阵 $\boldsymbol{G}$ 分块表示为

$$\boldsymbol{G}=\begin{bmatrix}\boldsymbol{G}_{n\times n} & \boldsymbol{G}_{n\times(N-n)}\\ \boldsymbol{G}_{n\times(N-n)}^{\mathrm{T}} & \boldsymbol{G}_{(N-n)\times(N-n)}\end{bmatrix} \tag{9}$$

假设 $\boldsymbol{G}_{n\times n}$ 被对角化为 $\boldsymbol{EAE}^{\mathrm{T}}$，则根据式(8)可获得密度权值矩阵 $\boldsymbol{G}$ 的 Nyström 特征向量逼近，记为

$$\hat{\boldsymbol{U}}=\begin{bmatrix}\boldsymbol{E}\\ \boldsymbol{G}_{n\times(N-n)}^{\mathrm{T}}\boldsymbol{EA}^{-1}\end{bmatrix} \tag{10}$$

用 $\hat{\boldsymbol{G}}$ 表示对密度权值矩阵 $\boldsymbol{G}$ 的逼近，于是有

$$\hat{\boldsymbol{G}}=\hat{\boldsymbol{U}}\boldsymbol{A}\hat{\boldsymbol{U}}^{\mathrm{T}}=\begin{bmatrix}\boldsymbol{E}\\ \boldsymbol{G}_{n\times(N-n)}^{\mathrm{T}}\boldsymbol{EA}^{-1}\end{bmatrix}\cdot \boldsymbol{A}\begin{bmatrix}\boldsymbol{E}^{\mathrm{T}} & \boldsymbol{A}^{-1}\boldsymbol{E}^{\mathrm{T}}\boldsymbol{G}_{n\times(N-n)}\end{bmatrix}=\begin{bmatrix}\boldsymbol{G}_{n\times n} & \boldsymbol{G}_{n\times(N-n)}\\ \boldsymbol{G}_{n\times(N-n)}^{\mathrm{T}} & \boldsymbol{G}_{n\times(N-n)}^{\mathrm{T}}\boldsymbol{G}_{n\times n}^{-1}\boldsymbol{G}_{n\times(N-n)}\end{bmatrix} \tag{11}$$

式(11)表明密度权值矩阵 $\boldsymbol{G}$ 仅需计算 n 个样本两两之间的权值矩阵 $\boldsymbol{G}_{n\times n}$ 和 n 个样本与剩余 $N-n$ 个样本之间的权值矩阵 $\boldsymbol{G}_{n\times(N-n)}$，结合 Nyström 特征向量逼近方法即可获得对 $\boldsymbol{G}_{(N-n)\times(N-n)}$ 矩阵的逼近，记为

$$\hat{\boldsymbol{G}}_{(N-n)\times(N-n)}=\boldsymbol{G}_{n\times(N-n)}^{\mathrm{T}}\boldsymbol{G}_{n\times n}^{-1}\boldsymbol{G}_{n\times(N-n)} \tag{12}$$

由于采样样本点数 $n\ll N$，因此可以省去对大量剩余样本之间的密度权值计算，即算法时间复杂度由原来的 $O(d^2N^2)$ 降为 $O(nd^2N)$，从而极大程度降低减法聚类的时间复杂度。此时，如果将 $\hat{\boldsymbol{G}}$ 矩阵的所有元素按行累加，即可获得各个样本点的密度值，实现基于 Nyström 的减法聚类样本密度值逼近。

经过上述的推导，通过逼近的 $\hat{\boldsymbol{G}}$ 矩阵显然可以直接计算样本的密度值。然而，在针对大规模样本集考虑降低减法聚类时间复杂度的同时，忘记随之产生的空间复杂度 $O(N^2)$。即当 N 比较大时，$\hat{\boldsymbol{G}}$ 矩阵对存储的需求往往超出物理容量，直接导致算法异常中断。因此，换一种思路，计算所有样本点的密度值

$$\boldsymbol{D}=\begin{bmatrix}\boldsymbol{G}_{n\times n}\boldsymbol{I}_n+\boldsymbol{G}_{n\times(N-n)}\boldsymbol{I}_{N-n}\\ \boldsymbol{G}_{n\times(N-n)}^{\mathrm{T}}\boldsymbol{I}_n+\boldsymbol{G}_{n\times(N-n)}^{\mathrm{T}}\boldsymbol{G}_{n\times n}^{-1}(\boldsymbol{G}_{n\times(N-n)}\boldsymbol{I}_{N-n})\end{bmatrix} \tag{13}$$

式中，$\boldsymbol{D}=[D(x_1),D(x_2),\cdots,D(\boldsymbol{x}_N)]^{\mathrm{T}}$，$\boldsymbol{I}_n$ 和 $\boldsymbol{I}_{N-n}$ 表示全体元素均为 1 的单列向量。

2.3　Nyström 减法聚类小结

考虑 d 维空间的 N 个样本点 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N$，将所有样本点归一化到一个超立方体中。基于 Nyström 样本密度值逼近的减法聚类算法可描述如下：

1)在 N 个样本点中等间隔 $\delta(1\leqslant\delta\leqslant N)$ 采样 n 个样本点，计算这 n 个样本点两两之间的密度权值矩阵 $\boldsymbol{G}_{n\times n}$、逆阵 $\boldsymbol{G}_{n\times n}^{-1}$ 以及 n 个样本与剩余 $N-n$ 个样本之间的密度权值矩阵 $\boldsymbol{G}_{n\times(N-n)}$；

2)计算 N 个样本的密度值

$$\begin{bmatrix}D(\boldsymbol{x}_1)\\ D(\boldsymbol{x}_2)\\ \vdots\\ D(\boldsymbol{x}_N)\end{bmatrix}=\begin{bmatrix}\boldsymbol{G}_{n\times n}\boldsymbol{I}_n+\boldsymbol{G}_{n\times(N-n)}\boldsymbol{I}_{N-n}\\ \boldsymbol{G}_{n\times(N-n)}^{\mathrm{T}}\boldsymbol{I}_n+\boldsymbol{G}_{n\times(N-n)}^{\mathrm{T}}\boldsymbol{G}_{n\times n}^{-1}(\boldsymbol{G}_{n\times(N-n)}\boldsymbol{I}_{N-n})\end{bmatrix} \tag{14}$$

3)在获得这 N 个样本的密度值后，选择最大密度值点 $\boldsymbol{x}_{c1}$ 为第 1 个聚类中心，$D(\boldsymbol{x}_{c1})$ 为对应的最大密度值。循环衰减每个数据点的密度值，根据式(15)修正所有样本点的密度值

$$D(\boldsymbol{x}_i)\leftarrow D(\boldsymbol{x}_i)-D(\boldsymbol{x}_{c*})\cdot\mathrm{e}^{-4(\boldsymbol{x}_i-\boldsymbol{x}_{c*})^{\mathrm{T}}\boldsymbol{M}_b^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_{c*})} \tag{15}$$

式中，$\boldsymbol{x}_{c*}$ 代表当前最大密度值中心位置，$D(\boldsymbol{x}_{c*})$ 代表与 $\boldsymbol{x}_{c*}$ 对应的最大密度值；

4)将步骤 3)循环迭代，直至找出所有有效聚类中心。以式(16)作为迭代终止条件，ε 经验值为 0.5。$D_1^*=D(\boldsymbol{x}_{c1})$ 代表第 1 个最大密度值，D_c^* 代表第 c 个密度值。

$$D_c^*<\varepsilon\cdot D_1^*\ (0<\varepsilon\leqslant 1) \tag{16}$$

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 18, No. 7, July 2013

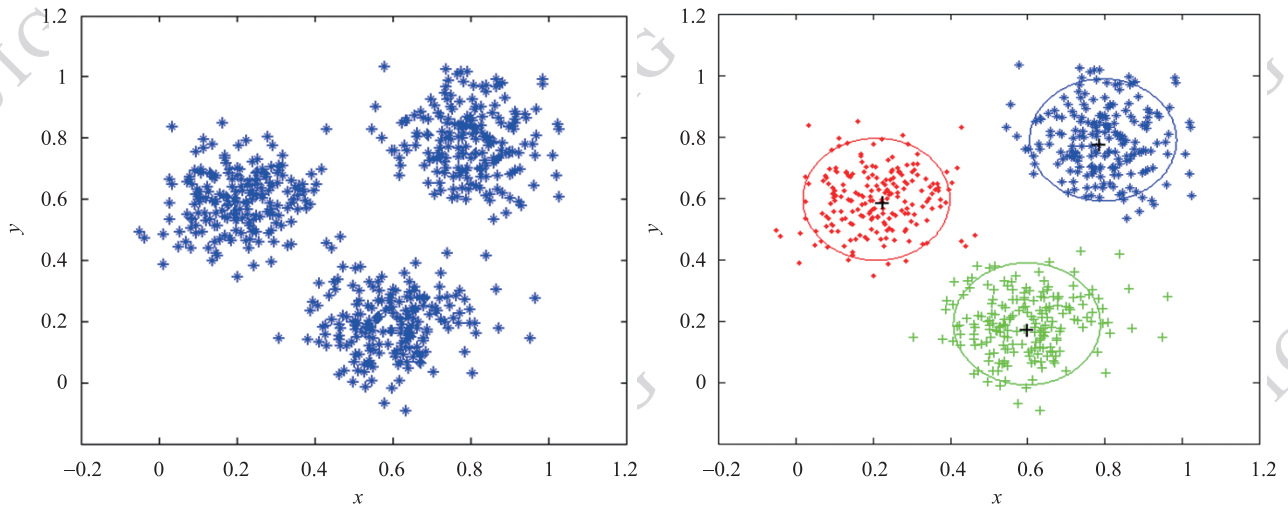
3 实验结果和分析

为了说明基于 Nyström 密度值逼近的减法聚类算法过程并验证其针对大规模数据样本的聚类性能，首先针对人工样本集详细说明本文算法过程，然后进一步针对标准彩色图像聚类及 UCI 标准数据库部分样本集的聚类验证本文算法性能。算法测试平台：处理器 Intel(R) Core(TM)2 Duo E8400 @3.00 GHz，内存 2.99 GHz，3.25 GB，操作系统 Windows XP SP2，编程环境 Matlab7.9.0(R2009b)。

3.1 人工数据实验

3.1.1 人工数据密度权值逼近实验

首先对 Matlab7.9.0(R2009b)所提供的一组人工数据进行实验。该数据命名为“clusterdemo.dat”，是 Matlab 专门用于数据分析的通用测试集。该数据集共有 600 个样本，即 $N=600$，样本维数 $d=3$，样本已按类别顺序排序，原始数据集无归一化。为便于直观说明本文算法与经典减法聚类算法具有同样的聚类过程，先对该测试样本集的第 1 和第 2 维样本进行聚类实验。在对该样本集采用本文算法进行聚类时，样本集采样间隔为 $\delta=10$，不同维度的邻域参数 r_{a1} 和 r_{a2} 统一设为 0.5，η 取 1.25，ε 取 0.5。采样样本数 $n=60$，剩余样本数 $N-n=540$。在用本文算法对样本集进行聚类前，首先对样本进行重组，即采样样本放新样本集前面，剩余样本统一放于新样本集后面。图 1(a)为该测试样本集 2 个维度的原始样本，图 1(b)给出的是本文算法的最终聚类结果。



(a) 人工样本集　　(b) 本文算法聚类结果

图 1　本文算法对人工样本聚类实验结果

Fig. 1　Clustering result of artificial samples

图 2 示意的是本文减法聚类所获得的经重采样后样本两两之间的样本密度权值矩阵。图 2 中的每个像素点的亮度值对应于两个样本之间的关系权值。图 2 中的像素点越亮代表两样本之间的关系权值越大，反之则越小。本文算法基于 Nyström 逼近理论，即利用少数采样样本两两之间的密度权值矩阵 $\boldsymbol{G}_{n\times n}$ 和采样样本与未采样样本两两之间的密度权值矩阵 $\boldsymbol{G}_{n\times(N-n)}$ 去逼近未采样样本两两之间的密度权值矩阵 $\boldsymbol{G}_{(N-n)\times(N-n)}$。从图 2 可以直观地看出上述逼近。

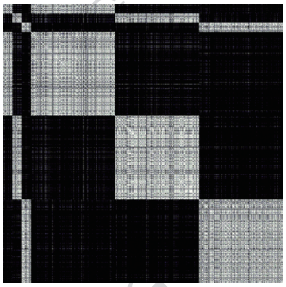


图 2　本文减法聚类样本密度权值矩阵

Fig. 2　The density weighted matrix using proposed subtractive clustering method

为便于进一步解释，将图 2 的样本密度权值矩阵根据采样样本数 n 及未采样样本数 $N-n$ 拆分为 4 个矩阵，分别为 $n\times n$、$n\times(N-n)$、$(N-n)\times n$ 以及 $(N-n)\times(N-n)$。由于重采样后的样本是按 n 个采样样本在前，$(N-n)$ 个采样样本在后的顺序排列的，则对应于式(9)，上述的 4 个矩阵分别代表 $\boldsymbol{G}_{n\times n}$、$\boldsymbol{G}_{n\times(N-n)}$、$\boldsymbol{G}^{\mathrm{T}}_{n\times(N-n)}$ 以及 $\boldsymbol{G}_{(N-n)\times(N-n)}$。

与图 2 对应的 4 个拆分矩阵如图 3 所示。严格地讲，图 3(d) 应该对应于 $\boldsymbol{G}_{(N-n)\times(N-n)}$ 的逼近 $\hat{\boldsymbol{G}}_{(N-n)\times(N-n)}$。

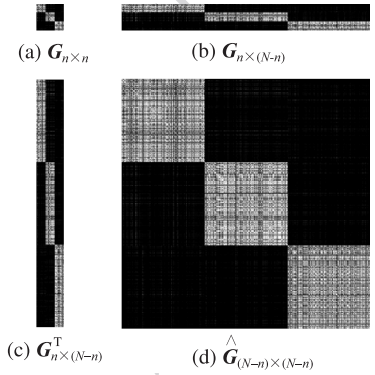


图 3　样本密度权值拆分矩阵

Fig. 3　The density weighted segmentation matrix of samples

本文算法与经典的减法聚类算法明显不同之处在于，本文算法仅仅用少数采样样本之间的密度权值矩阵 $\boldsymbol{G}_{n\times n}$ 和少数采样样本与剩余未采样样本之间的密度权值矩阵 $\boldsymbol{G}_{n\times(N-n)}$ 直接逼近 $\boldsymbol{G}_{(N-n)\times(N-n)}$。而传统减法聚类则是通过计算所有样本两两之间的密度权值矩阵，当用于大规模多维度数据聚类时，算法的时间复杂往往让人难于接受。引入 Nyström 逼近后，减法聚类算法的时间复杂度从 $O(d^2N^2)$ 降至 $O(nd^2N)$，又因为 $n<<N$，因此，本文算法可极大程度降低经典减法聚类的时间复杂度。图 4 给出了这种直观的逼近策略。

为进一步和经典减法聚类算法进行比较，将图 2 的权值矩阵根据重采样时所记录的样本索引进行还原，得到如图 5(a) 所示的效果，图 5(b) 对应于经典减法聚类算法所获得的权矩阵。从图 5 可看出，本文算法的逼近结果与经典减法聚类的结果基本一致。图 5(a) 和图 5(b) 均成对角方块分布，由于原始样本点按类别为单位经过了排序，且同类样本两两之间的密度权值比较大，所以图 5 呈现出对角方块状规律性分布。由于采用等间隔采样的方

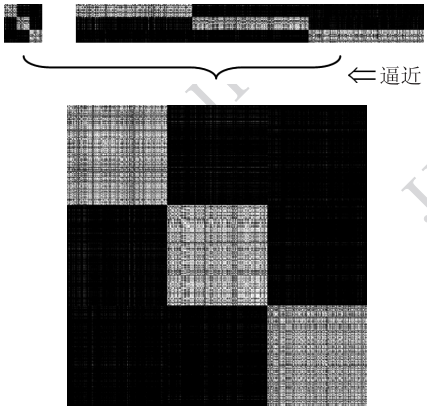
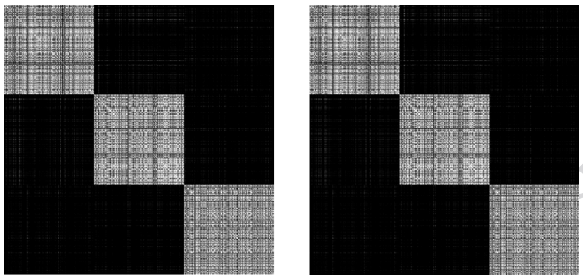


图 4　用少数密度权值逼近多数密度权值

Fig. 4　Majority density weighted value approximation using few density weighted value



(a) 本文减法聚类　　(b) 经典减法聚类

图 5　本文及经典减法聚类所获取的样本密度权值矩阵对比

Fig. 5　Density weighted matrix comparison between proposed method and classical subtractive clustering method

式，所以图 3(a) 的纹理和图 5 的纹理相似。在获得样本的密度权值后，则可利用式(14)对所有样本的密度值进行计算，而无须对权值矩阵进行存储(注：对大规模样本集直接存储可能会发生内存溢出问题)。

图 6 给出了对应于图 5(a) 和图 5(b) 的减法聚类样本密度值的计算结果。其中红色圆圈点密度值分布为对应于图 5(a) 由本文减法聚类所计算及估计的样本密度值。蓝色实心点密度值分布对应于图 5(b) 由经典减法聚类直接计算的样本密度值。从图 6 可以更加直观地看出，本文算法和经典减法聚类算法所获取的样本密度值基本是重合的。

3.1.2　人工数据密度值分布衰减过程实验

3.1.1 节详细解释了如何以少数采样样本的密度权值获取未采样样本的密度权值，以及如何最终获得所有样本的密度值。为确保实验的整体性，在 3.1.1 节的基础上，继续解释减法聚类样本密度值分布的衰减过程。

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 18, No. 7, July 2013

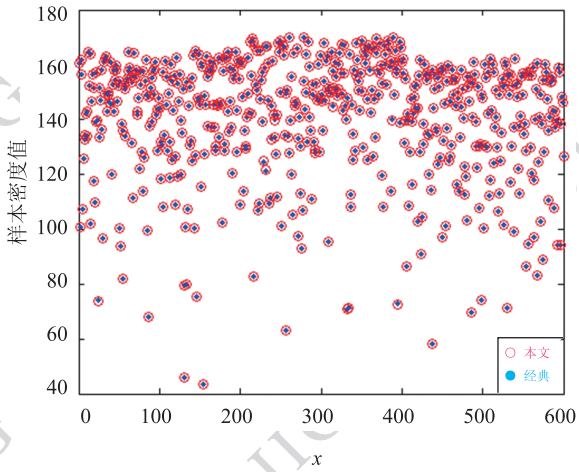
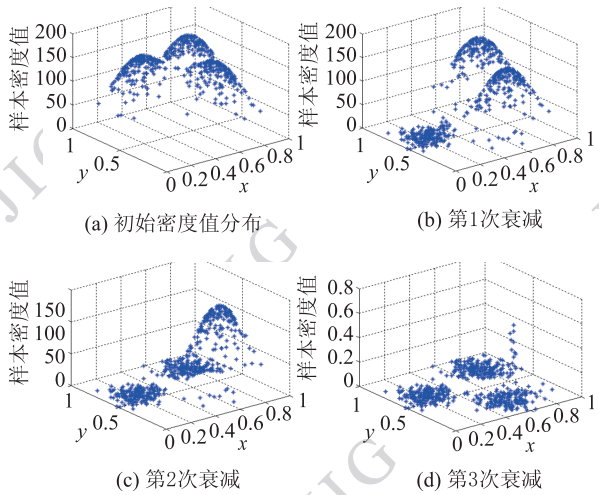


图 6　本文及经典减法聚类获取的样本密度值分布

Fig. 6　Density value distribution of proposed method and classical subtractive clustering method

用本文算法获取所有样本的密度值后，由于选用的人工数据样本是 2 维数据，因此可以直接对样本的密度值分布进行显示。图 7 给出了所有样本密度值分布及衰减过程示意。图 7(a)为对应于图 6 中 600 个样本的密度值空间分布，显现出 3 类数据样本密度值分布的山峰特性。在获得最大密度值 D_1^* 及对应的密度中心 $\boldsymbol{x}_{c1}$(0.222 1, 0.586 1)后，采用式(15)对图 7(a)的分布作第 1 次衰减得到图 7(b)所示的分布。在图 7(b)中计算此时的最大密度值 D_2^* 并获取与其对应的聚类中心 $\boldsymbol{x}_{c2}$(0.784 3, 0.771 5)，并做第 2 次衰减，如图 7(c)所示。在图 7(c)中计算此时的最大密度值 D_3^* 并获取与其对应



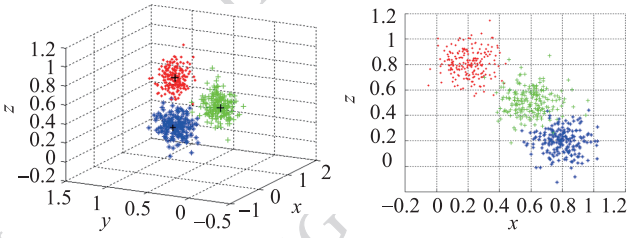
(a) 初始密度值分布　(b) 第1次衰减

(c) 第2次衰减　(d) 第3次衰减

图 7　本文减法聚类样本密度值分布衰减过程

Fig. 7　Density value decaying procedure of samples using proposed subtractive clustering method

的聚类中心 $\boldsymbol{x}_{c3}$(0.596 5, 0.169 3)，并做第 3 次衰减。做完 3 次衰减后，所有密度值符合式(16)的终止条件而完成减法聚类。因此，从图 7 可以直观地看出减法聚类做“减法”的整个过程，也可从中看出，减法聚类很适合用于紧凑的超球形分布的样本集。Kim 等人[8]虽在计算样本密度权值时引入了核函数，但由于“减法”聚类自身的特性，核函数减法聚类对非凸流形样本集的聚类性能还是比较有限的。图 8 展示了本文算法对“cluster-demo. dat”中所含 600 个样本 3 维样本进行聚类的结果。



(a) 聚类结果　(b) x-z平面

图 8　本文算法的聚类结果

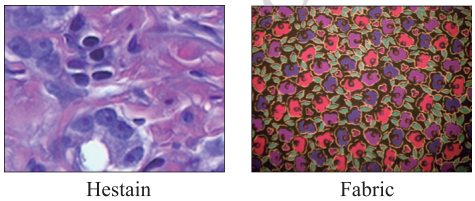
Fig. 8　Clustering result using proposed method

3.2　彩色图像聚类实验

针对人工数据集，4.1 节利用丰富的实验结果详细分析了本文算法对未采样样本之间密度权值的逼近、样本密度值计算及减法聚类的整个过程。为进一步说明本文算法相对经典减法聚类时间复杂度的降低程度，对 Matlab 提供的两幅公用测试彩色图像结合 k-均值算法进行了聚类实验。图 9(a)图像格式 152×114×3，类别数为 3。图 9(b)图像格式 320×240×3，类别数为 6。图像原为 RGB 空间的彩色图像，本文算法聚类时将其转至 Lab 空间，并对图像像素 Lab 空间中的 a、b 色度分量进行聚类。因此，调整后 Hestain 图像样本总个数为 17 328，维度为 2。同理，Fabric 图像样本总个数为 76 800，维度为 2。

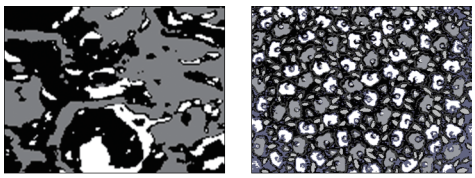
对于 Hestain 图像，样本总数为 17 328，采样间隔 δ 为 50，采样样本数 n 为 347，剩余样本数 $N-n$ 为 16 981，r_{a1} 和 r_{a2} 设置为 0.3，η 取 1.25，ε 取 0.5。对于 Fabric 图像，样本总数为 76 800 个，采样间隔 δ 为 500，样本数 n 为 154，剩余样本数 $N-n$ 为 76 646，r_{a1} 和 r_{a2} 设置为 0.15，η 取 1.25，ε 取 0.5。图 9 给出了本文算法与经典减法聚类算法的实验对比。从图 9(b)与图 9(c)的聚类结果来看，本文算

法与经典减法聚类算法对 Hestain 与 Fabric 两幅图像具有几乎一致的聚类效果。

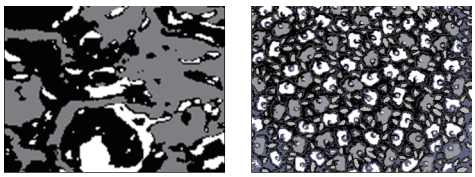


Hestain　　Fabric

(a) 两幅图像



(b) 经典减法聚类



(c) 本文减法聚类

图 9　彩色图像减法聚类实验

Fig. 9　Subtractive clustering experiments for colour images

表 1 和表 2 分别给出了 Hestain 和 Fabric 对应的聚类参数。从聚类中心及其对应的密度值上看，个别聚类中心的密度值有细微差异，但本文算法与经典减法聚类算法均具有同样的聚类中心。在聚类效果一样的情况下，观察两种减法聚类算法的平均耗时情况（重复 10 次的平均值），可以发现本文算法的加速效果（加速比——算法改进前、后的耗时比）非常明显，特别是对样本数达 76 800 个的情况，性能加速比高达 252. 56。

表 1　Hestain 图像聚类参数

Table 1　The parameters using subtractive clustering and proposed methods for Hestain

方法	聚类中心	密度值	平均耗时/s
经典减法聚类	(168,81)	6 816. 2	19. 587
	(418,84)	2 359. 6	
	(153,92)	1 685. 6	
本文算法	(168,81)	6 816. 2	0. 107
	(418,84)	2 359. 7	
	(153,92)	1 685. 4	
算法改进前后时间性能加速比 S_p			183. 06

表 2　Fabric 图像聚类参数

Table 2　The parameters using subtractive clustering and proposed methods for Fabric

方法	聚类中心	密度值	平均耗时/s
经典减法聚类	(139,146)	14 263. 0	523. 309
	(148,159)	7 054. 7	
	(137,134)	4 284. 1	
	(195,157)	3 014. 0	
	(111,142)	2 881. 0	
	(176,112)	2 174. 0	
本文算法	(139,146)	14 262. 0	2. 072
	(148,159)	7 054. 8	
	(137,134)	4 284. 1	
	(195,157)	3 034. 8	
	(111,142)	2 858. 9	
	(176,112)	2 163. 7	
算法改进前后时间性能加速比 S_p			252. 56

3. 3　UCI 数据集聚类实验

为进一步评价本文算法的加速性能，及针对不同规模数据集的泛化性能。选取 UCI 机器学习数据库（http://archive. ics. uci. edu/ml/）中部分不同样本数和维数的数据集对算法进行验证。表 3 给出了所选数据集的特性，将聚类结果与真实已知分类信息匹配后，给出了聚类的准确率及平均运行时间。不同数据集的参数设置、聚类准确率及平均耗时（重复 10 次的平均值）如表 4 所示。从表 4 的时间性能参数可以发现，对于样本数不多的数据集，本文减法聚类耗时基本与经典减法聚类算法一样，体现不出加速性。然而，随着数据

表 3　实验中所选 UCI 数据集特性

Table 3　Characteristics of the UCI dataset used for experiment

数据集	类数	维数	点数	归一化
Iris	3	4	150	否
Wine	3	13	178	是
X8D5K	5	8	1 000	否
Normal7	7	2	7 000	是

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 18, No. 7, July 2013

表 4 UCI 数据集减法聚类实验结果比较

Table 4 Experiment result comparison of the UCI dataset

方法	参数	Iris	Wine	X8D5K	Normal7
经典减法聚类	邻域 r_{ak}	0.6	1.1	0.46	0.4
	准确率/%	89.33	70.22	97.4	99.69
	耗时/s	0.023 9	0.025 6	0.160 4	3.123 5
本文算法	邻域 r_{ak}	0.6	1.0	0.46	0.4
	采样数	14	17	10	70
	准确率/%	89.33	70.22	97.4	99.69
	耗时/s	0.023 6	0.025 0	0.025 5	0.080 1
	加速比	1.013	1.024	6.290	38.995

集规模的增大,本文减法聚类的加速性能也越明显。在聚类准确率上,由于两种方法最后所获得的聚类中心一致,也就是说做进一步分类时,k-均值算法的 k 取值相同,在聚类中心初始值也相同的情况下,k-均值算法将会得到同样的准确率。

4 结 论

基于 Nyström 逼近理论,结合经典减法聚类样本密度值计算的特点,巧妙地将 Nyström 逼近用于减法聚类未采样样本之间密度权值矩阵的逼近,提出了一种新的减法聚类方法—基于 Nyström 密度值逼近的减法聚类。利用 Nyström 逼近理论对未采样样本之间的密度权值矩阵进行逼近,从而获取未采样样本的密度值逼近。实验结果表明,针对本文所用的样本数据和实验参数,与经典减法聚类相比,本文算法在不影响聚类结果的情况下,对于较大规模数据集(如图像数据),本文算法可从原来的时间复杂度 $O(d^2N^2)$ 降为 $O(nd^2N)$,以提高减法聚类的实时性。

众所周知,减法聚类对于邻域半径参数十分敏感,对于邻域半径参数的自适应选取至今仍未得到很好地解决;减法聚类中心个数受 ε 参数的影响,如何自适应的计算聚类中心数也是值得研究的问题;引入 Nyström 对减法聚类样本密度值进行逼近,既然是逼近则必定存在逼近误差,同时,采样间隔对逼近误差的影响也可以同时考虑。Zhang 等人[11]对 Nyström 在谱聚类中的逼近做了分析,给出了逼近误差的分析方法。同样,Nyström 对减法聚类的逼近分析也将是进一步研究的内容。

参考文献(References)

[1] Chiu S L. Fuzzy model identification based on cluster estimation [J]. Journal of Intelligent and Fuzzy Systems, 1994, 2(3): 267-278.

[2] Luxburg U V. A Tutorial on Spectral Clustering [J]. Statistics and Computing, 2007, 17(4): 395-416.

[3] Das S, Bora P K, Gogoi A K. Subtractive clustering of vertices for CPCA based animation geometry compression [C]//Proceedings of the 7th Indian Conference on Computer Vision, Graphics and Image Processing. New York, NY, USA: ACM, 2010: 205-210.

[4] Bataineh K M, Naji M, Saqer M. A comparison study between various fuzzy clustering algorithms [J]. Jordan Journal of Mechanical and Industrial Engineering, 2011, 5(4): 335-343.

[5] Sun Z H, Zhou W H. Li E T, et al. Study on subtractive clustering video moving object locating method with introduction of eigengap [C]//Proceedings of the 9th International Conference on Fuzzy Systems and Knowledge Discovery. Chongqing, China: IEEE, 2012: 626-629.

[6] Chen J Y, Zheng Q, Ji J. A weighted mean subtractive clustering algorithm [J]. Information technology Journal, 2008, 7(2): 356-360.

[7] Ngo L T, Pham B H. A type-2 fuzzy subtractive clustering algorithm [J]. Advances in Intelligent and Soft Computing, 2012, 125: 395-402.

[8] Kim D, Lee K, Lee D, et al. A kernel-based subtractive clustering method [J]. Pattern Recognition Letters, 2005, 26(7): 879-891.

[9] Fowlkes C, Belongie S, Chung F, et al. Spectral grouping using the Nyström method [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, 26(2): 214-225.

[10] Jia J H. Research on Spectral Clustering Ensemble Algorithm [M]. Tianjin: Tianjin University Press, 2011. [贾建华. 谱聚类集成算法研究[M]. 天津: 天津大学出版社, 2011.]

[11] Zhang X C, You Q Z. Clusterability analysis and incremental sampling for Nyström extension based spectral clustering [C]// Proceedings of the 11th IEEE International Conference on Data Mining. Vancouver, Canada: IEEE, 2011: 942-951.