计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190300125

图像的扩散界面无监督聚类算法

王成章[1]　白晓明[2]　杜金栗[1]
1 中央财经大学统计与数学学院　北京 100081
2 首都经济贸易大学信息学院　北京 100070

摘　要　图像的无监督聚类就是基于图像数据,在无任何先验信息的情况下将整个图像集合划分成若干子集的过程。由于图像的本征维度很高,在图像处理中会遇到"维数灾难"问题。针对图像无监督聚类的特点,提出了一种图像的扩散界面无监督聚类算法,将图像编码成高维观测空间中的点,再通过投影变换映射到低维特征空间,在低维特征空间中构建扩散界面无监督聚类模型,并在模型中引入维度约简算子,采用循环迭代算法优化扩散界面模型的能量函数。基于最优的扩散界面,将整个图像集合聚类成不同的子集。实验结果表明,扩散界面无监督聚类算法优于传统聚类算法中的 K-means 算法、DBSCAN 算法和 Spectral Clustering 算法,能够更好地实现图像的无监督聚类,在相同条件下具有更高的准确度。

关键词:扩散界面;无监督学习;图像聚类;维度约简;最优化
中图法分类号　TP391

Diffuse Interface Based Unsupervised Images Clustering Algorithm

WANG Cheng-zhang[1],BAI Xiao-ming[2] and DU Jin-li[1]
1 School of Statistics and Mathematics,Central University of Finance and Economics,Beijing 100081,China
2 Information School,Capital University of Economics and Business,Beijing 100070,China

Abstract　Unsupervised clustering of images aims to partition the whole image set into several subsets on the basis of image data itself,while without any priori information. As dimensionality of an image is usually very high,curse of dimensionality arises during the image processing. Having analyzed the problem of images clustering,a novel unsupervised image clustering algorithm is proposed. The proposed algorithm is based on diffused interface model on graph. Images were encoded as the data points in high dimensional observing space,and then were projected into low dimensional feature space. Diffuse interface model based unsupervised clustering algorithm was constructed in feature space,and dimension reduction operator was introduced into the model. Loop iterative algorithm was employed to optimize the energy function of diffuse interface model. The optimized diffuse interface was adopted to cluster images into different subsets. Experimental results show that the proposed algorithm is superior to traditional K-means,DBSCAN and Spectral Clustering algorithm. It achieves better clustering results and lower error rates.

Keywords　Diffuse interface,Unsupervised learning,Image clustering,Dimension reduction,Optimization

1　引言

聚类分析就是根据目标对象数据集的模式特征(观测值、数据项、特征向量等)将数据集分成不同的组,其中组内样本间的相似度较高,组间样本间的差异度较高。图像作为重要的非结构化数据,对其进行聚类分析一直是机器学习、计算机视觉领域的研究热点之一。基于内容的图像检索[1-3]、图像标注[4]等都是图像聚类的典型应用。

图像的无监督聚类就是在没有任何先验信息的条件下,根据预先定义好的准则或相似性度量,将图像数据集中的样本分成不同的子集。一般来说,传统方法中预先定义好的聚类准则都是基于某种假定建立起来的,而实际图像数据集中包含有大量的噪声数据,难以满足准则成立的条件;图像数据集的高维度特点会使图像处理面临维数灾难问题,同时高维数据中的冗余和噪声信息会造成误差的累积,无益于错误率的降低[5]。近年来,基于图的高维数据处理方法显示出了其独特的优势[6-7]。该方法首先将数据样本抽象成图中的点,然而基于样本点之间的相似性度量建立点之间的连接模式,从而构成一个图,最后在图上对样本点进行分类或者聚类。

扩散界面模型[8]在流体力学领域被用来建模两种流体在同一空间中扩散时所形成的界面的特性。该模型后来被引入计算数学领域,并被推广到图领域[6],用于对构成图的节点集

到稿日期:2019-03-25　返修日期:2019-06-12　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
基金项目:国家自然科学基金(71571197);北京市自然科学基金(9152016)
This work was supported by the National Natural Science Foundation of China (71571197) and Natural Science Foundation of Beijing,China (9152016).
通信作者:王成章(czwang@cufe.edu.cn)

合进行分割。在呈现非线性复杂分布的数据集合(如双月亮数据集)上,扩散界面模型展现出了很好的分类性能。针对图像处理中的维数灾难问题,通常先对图像数据进行降维,然后基于降维后的数据对目标对象进行聚类或分类。常用的降维方法有主成分分析、非负矩阵分解、奇异值分解等,这些方法的本质是基于某种准则建立从高维观测空间到低维特征空间的映射。值得注意的是,数据降维的目标函数与样本聚类或分类的目标函数并不完全一致,这种分开处理的模式势必会降低算法的性能。鉴于此,本文提出了一种新的图像无监督聚类算法,该算法将图像编码成几何图上的顶点,图中连接顶点的边采用图像之间的相似度进行度量。本文在几何图上构建扩散界面模型,并在模型中引入维度约简算子,基于最优的扩散界面对图像集合进行无监督聚类。

2 相关工作

从某种意义来看,聚类研究是对目标数据集进行数据建模,并给出总结性的描述。一般来说,现有的聚类算法大致可以分成 3 类:基于划分的聚类算法、基于密度的聚类算法和基于图的聚类算法。

基于划分的聚类算法是根据数据样本之间不同的距离度量模式来描述样本之间的关系,将数据集进行分割,并以分割准则目标函数最优作为数据集的聚类结果。典型的聚类算法有 K-means 算法[9-10]和聚合算法[11]。

基于密度的聚类算法是根据数据样本分布的密度函数对数据集合中的样本点进行划分,一般要求组内点的分布密度大于一定的阈值,或者组内点的邻域半径内分布的点的个数大于一定的阈值。典型的聚类算法有 DBSCAN 算法[12-13]。

基于图的聚类算法是将样本数据集中的目标对象抽象成几何图中的顶点,顶点数据间的相似度量作为相应顶点连接边的权值,基于图的最优划分准则对样本数据集进行分组。典型的聚类算法有 Spectral Clustering 算法[14-15]。

3 扩散界面无监督聚类算法

本文将一幅图像抽象为高维观测空间中的点,假定整个图像集合一共有 K 幅图像,每幅图像的大小为 $r\times c$,则该图像可记为 $\boldsymbol{x}=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$,其中 $n=r\times c$。

针对图像处理中的“维数灾难”问题,本文将图像从高维观测空间映射到低维特征空间,即:

$$f:\boldsymbol{x}\mapsto\boldsymbol{y}$$

其中,$\boldsymbol{x}\in\mathbb{R}^n$ 为高维观测空间中图像对应的点,$\boldsymbol{y}\in\mathbb{R}^d(d<n)$ 为该图像在对应的低维特征空间中的映射点,f 为映射函数。

在图像的低维特征空间$\mathbb{R}^d$上,本文将映射点 $\boldsymbol{y}$ 编码成几何图中的顶点,顶点之间连接边的权值则定义为:

$$w(\boldsymbol{y}_i,\boldsymbol{y}_j)=\exp\left(\frac{-\|\boldsymbol{y}_i-\boldsymbol{y}_j\|^2}{\tau}\right)$$

其中,τ 为调节因子。本文在低维特征空间的几何图上构建扩散界面模型,其能量函数为:

$$E(\varphi)=\frac{\mu}{2}\int|\nabla\varphi|^2\mathrm{d}\boldsymbol{y}+\frac{1}{\mu}\int F(\varphi)\mathrm{d}\boldsymbol{y}\tag{1}$$

其中,$F(\cdot)$为双井势函数,本文定义为 $F=k\cdot(u^2-1)^2$,k 为非零常数,该函数在±1 点处取得最小值;∇为空间上的梯度算子;$\int|\nabla\varphi|^2\mathrm{d}\boldsymbol{y}$ 为实值变量 φ 的 H^1 半范数;参数 μ 为扩散界面尺度。该模型被称为扩散界面,是因为能量函数的两个构成部分之间存在“竞争”关系[6]。在最小化该能量函数的要求下,作为第 2 个组成因子的双井势函数会迫使变量 u 向 $+1$ 或者 -1 无限靠近;而作为第 1 个组成因子的半范数 H^1 则会要求变量 φ 具有一定的平滑性,从而消除其在 F 的两个最小值点处的剧烈跳跃。最小化能量函数的结果为一个区域接近$+1$,另外一个区域接近-1,在这两个区域之间存在一个非常“薄”的过渡区域,该过渡区域称为界面。

本文提出的扩散界面无监督聚类算法就是要寻求最优的低维特征子空间,使得扩散界面模型的能量函数达到最小化,即:

$$\arg\min_{\varphi^*,f^*}E(\varphi,f)=\frac{\mu}{2}\int|\nabla\varphi[f(\boldsymbol{x})]|^2\mathrm{d}f(\boldsymbol{x})+\frac{1}{\mu}\int F(\varphi[f(\boldsymbol{x})])\mathrm{d}f(\boldsymbol{x})\tag{2}$$

其中,f 定义为维度约简算子。

为了计算式(2)所示的目标函数的最优解,本文采用循环迭代法进行最优化计算。

Step1 给定映射函数 f^*,计算低维特征空间上的能量函数:

$$E(\varphi,f^*)=\frac{\mu}{2}\int|\nabla\varphi[f^*(\boldsymbol{x})]|^2\mathrm{d}f^*(\boldsymbol{x})+\frac{1}{\mu}\int F(\varphi[f^*(\boldsymbol{x})])\mathrm{d}f^*(\boldsymbol{x})\tag{3}$$

然后采用文献[6]给出的基于 PDE 的方法求解:

$$\arg\min_{\varphi^*}E(\varphi,f^*)$$

其中:

$$\varphi_t=-\mu\cdot\Delta\varphi-\frac{1}{\mu}\cdot F'(\varphi)$$

Step2 给定变量 φ^*,计算低维特征空间上的能量函数:

$$E(\varphi^*,f)=\frac{\mu}{2}\int|\nabla\varphi^*[f(\boldsymbol{x})]|^2\mathrm{d}f(\boldsymbol{x})+\frac{1}{\mu}\int F(\varphi^*[f(\boldsymbol{x})])\mathrm{d}f(\boldsymbol{x})\tag{4}$$

采用梯度下降法求解目标函数:

$$\arg\min_{f^*}E(\varphi^*,f)$$

其中,迭代方向为:

$$\nabla f=-\frac{\mu}{2}|\nabla\varphi^*[f(\boldsymbol{x})]|^2-\frac{1}{\mu}F(\varphi^*[f(\boldsymbol{x})])$$

Step3 重复 Step1 和 Step2,直到满足收敛条件为止。

值得注意的是,在算法的 Step2 中给定变量 φ^*,采用梯度下降法求解$\arg\min_{f^*}E(\varphi^*,f)$时,最优迭代步长的选择会耗费巨大的计算资源。如果从图像的高维空间$\mathbb{R}^n$到低维特征空间$\mathbb{R}^d$上的映射 f 为线性映射,即 $\boldsymbol{y}=\boldsymbol{W}\cdot\boldsymbol{x}$,其中 $\boldsymbol{W}$ 为 $d\times n$ 维的映射矩阵,则算法的 Step2 中可以采用简单高效的两阶段网格优化算法来求解。

第 1 阶段 对 $\boldsymbol{W}$ 的可行域空间进行粗粒度的网格划分,在每一个网格点处计算待优化的目标函数(4),并比较各个网格点处的目标函数值。记最优值对应的网格点为 $\boldsymbol{W}^a$。

第 2 阶段　在网格点 $\boldsymbol{W}^{\alpha}$ 的附近做细粒度的网格划分，进一步在各个细粒度网格划分的网格点处计算待优化的目标函数(4)，并比较各个网格点处的目标函数值。记最终最优值对应的网格点为 $\boldsymbol{W}^{*}$，则最优的映射函数为：

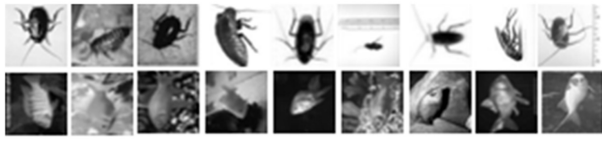
$$f^{*}=\boldsymbol{W}^{*}\cdot\boldsymbol{x}$$

最后，基于最优的实值函数 $\varphi^{*}(\cdot)$ 对图像数据集进行无监督聚类。

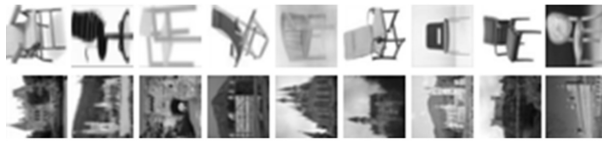
4　实验结果

为了验证本文提出的扩散界面无监督聚类算法在图像聚类上的有效性和准确性，分别在 CIFAR-100，CIFAR-10，MNIST，VOC，CVL 和 ORL 数据库上进行了实验，同时还选取了传统聚类算法中的 K-means，DBSCAN 和 Spectral Clustering 在相同的数据集上进行测试，并对最终的实验结果进行了比较分析。在聚类实验中，对于传统的聚类算法，本文首先采用 MDS(Multidimensional Scaling)算法进行维度约简，将图像从高维观测空间投影到同一维度的低维特征空间，然后在低维特征空间上进行图像的聚类分析。实验中采用了 Python 语言的第三方支持库 sklearn[1)] 中的聚类算法包。K-means 算法和 DBSCAN 算法基于图像数据之间的欧氏距离来计算样本间的相似度，Spectral Clustering 算法基于图像数据之间的高斯核函数来计算样本间的相似度。

CIFAR-100 数据库[16]包含了水生哺乳动物类、鱼类、花类、食品容器类、水果蔬菜类、家用电器设备类、家具类、昆虫类等 20 个子类，共 100 类不同对象的彩色图像数据，每类对象包含 6000 幅彩色图像，图像大小为 32×32。本文在 CIFAR-100 数据库中随机挑选了类别标号为 1 和 24，17 和 20，24 和 73，52 和 23，71 和 52 共 5 组对象的图像数据进行聚类实验，实验中首先将彩色图像转换成灰度图像。用于实验的部分图像数据如图 1 所示。

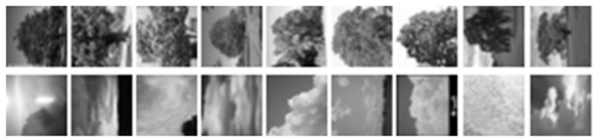


(a)类别标号为 1 和 24 的图像

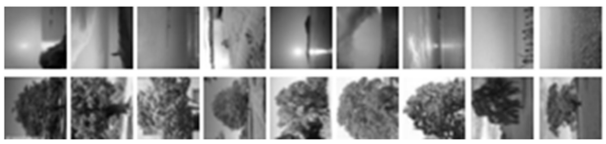


(b)类别标号为 17 和 20 的图像

(c)类别标号为 24 和 73 的图像



(d)类别标号为 52 和 23 的图像



(e)类别标号为 71 和 52 的图像

图 1　CIFAR-100 的部分实验图像

Fig. 1　Images from CIFAR-100

CIFAR-10 数据库[16]包含了飞机、小汽车、鸟、猫、鹿、狗、青蛙、马、轮船和卡车共 10 类不同对象的彩色图像数据，每个对象包含 6000 幅彩色图像，图像大小为 32×32。本文在 CIFAR-10 数据库中随机挑选了鹿和卡车的图像数据进行聚类实验，实验中首先将彩色图像转换成灰度图像。用于实验的部分图像数据如图 2 所示。

图 2　CIFAR-10 的部分实验图像

Fig. 2　Images from CIFAR-10

MNIST 数据库[17]包含 0 到 9 共 10 个数字的手写体图像，数据库共有 70000 幅图像，每个 MNIST 图像是一个单一的手写数字字符的数字化图片，每幅图像的大小为 28×28。MNIST 数据库的图像数据如图 3 所示。本文分别挑选了数字 1 和 3、数字 4 和 9 的图像进行了两组聚类实验。

图 3　MNIST 的图像数据

Fig. 3　Images from MNIST

PASCAL VOC 数据库[18]包含 20 个不同对象类别的彩色图像数据，具体的图像集合可分成人类、动物类(鸟、猫、牛、狗、马、羊)、交通工具类(飞机、自行车、船、公共汽车、小轿车、摩托车、火车)、室内物品类(瓶子、椅子、餐桌、盆栽植物、沙发、电视)。实验中首先将彩色图像转换成灰度图像，并将图像的大小归一化为 300×300。本文在 VOC 数据库中随机选择了动物类中的鸟和交通工具类中的公共汽车的图像进行了聚类实验，部分实验图像数据如图 4 所示。

1) https://scikit-learn.org/stable/index.html

图 4 VOC 的部分实验图像

Fig. 4 Images from VOC

CVL 数据库[19]包含 114 人的彩色人脸图像，每人 7 幅不同角度的面部图像，每幅图像的大小为 648×480。实验中首先将彩色图像转换成灰度图像，并在 CVL 数据库中随机选择了类别标号为 1 和 5、1 和 8、1 和 63 共 3 组人脸图像进行了聚类实验。数据库的部分图像数据如图 5 所示。

图 5 CVL 数据库的部分人脸图像

Fig. 5 Images from CVL

ORL 数据库[20]包含从 1992 年 4 月到 1994 年 4 月期间拍摄的 40 个不同年龄、不同性别和不同种族的人脸对象的面部灰度图像，每人 10 幅图像，每幅图像的大小为 92×112。从 ORL 数据库中随机选择类别标号为 3 和 35、3 和 17 共 2 组人脸图像进行了聚类实验。数据库的部分图像数据如图 6 所示。

图 6 ORL 数据库的部分人脸图像

Fig. 6 Images from ORL

本文在每一个数据集上分别采用本文提出的聚类算法、K-means 聚类算法、DBSCAN 算法和 Spectral Clustering 算法进行了图像的聚类实验，并统计了各种聚类算法的错误率(%)。实验结果如表 1 所列。

表 1 聚类错误率

Table 1 Error rate of clustering

(单位：%)

数据集	本文算法	K-means	DBSCAN	Spectral
Cifar-I	**5.10**	10.50	49.10	6.65
Cifar-II	**3.74**	9.54	49.60	4.29
Cifar-III	7.94	13.10	49.20	**7.89**
Cifar-IV	**9.49**	12.30	49.40	11.40
Cifar-V	**6.88**	12.10	49.60	9.91
Cifar-10	**18.70**	25.47	46.40	20.48
MNIST-I	**1.13**	4.33	5.20	8.07
MNIST-II	**21.30**	46.30	46.20	31.00
VOC	**25.50**	42.09	49.70	32.61
CVL-I	**6.42**	14.30	21.50	33.57
CVL-II	**3.74**	9.54	35.80	4.29
CVL-III	**7.85**	22.86	35.80	19.29
ORL-I	**5.00**	18.00	10.00	17.00
ORL-II	**3.74**	9.54	40.00	4.29

实验结果表明，本文提出的图扩散界面无监督聚类算法在 CIFAR-100 数据库上的 5 组图像子集 Cifar-I(标号为 1 和 24 的图像)、Cifar-II(标号为 17 和 20 的图像)、Cifar-III(标号为 24 和 73 的图像)、Cifar-IV(标号为 52 和 23 的图像)、Cifar-V(标号为 71 和 52 的图像)上均取得了较好的聚类效果。与其他 3 种聚类算法相比，本文的错误率除了在 Cifar-III 图像集合上比 Spectral Clustering 算法略高外(高 0.05%)，在其余 4 组实验中均低于其他 3 种聚类算法。

本文分别在 Cifar-10 数据库和 VOC 数据库上进行聚类实验，在 MNIST 数据库上进行 2 组实验(MNIST-I(数字 1 和 3 的图像)、MNIST-II(数字 4 和 9 的图像))，在 CVL 数据库上进行 3 组实验(CVL-I(标号为 1 和 5 的图像)、CVL-II(标号为 1 和 8 的图像)、CVL-III(标号为 1 和 63 的图像))，在 ORL 数据库上进行两组实验(ORL-I(标号为 3 和 35 的图像)、ORL-II(标号为 3 和 17 的图像))。实验结果表明，本文算法的聚类错误率均低于相同条件下其余 3 种聚类算法的错误率。在 MNIST 数据库上，4 种聚类算法对相似度较高的 2 个数字 4 和 9 的聚类错误率都较高。在所有的实验数据集中，VOC 数据库上的图像数据的聚类效果不佳，4 种聚类算法的聚类错误率都较高。

本文还统计了各种算法在各个数据集上进行聚类分析时所需的时间，计量单位为单幅图像上算法所需的时间(单位为 s)，即算法对图像数据集进行聚类所需的总时间除以图像样本的个数。实验结果如图 7 所示。

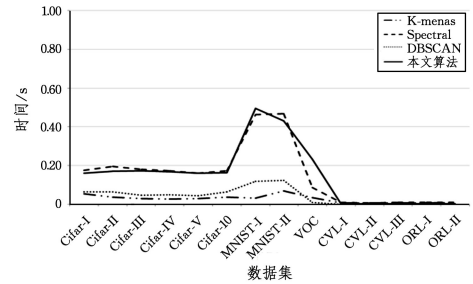


图 7 算法运行时间的比较

Fig. 7 Comparison of consuming time

实验结果表明，4 种算法在各个数据集上运行所需的时间均未超过 0.5 s，本文提出的聚类算法与 Spectral Clustering 算法的运行时间基本一致，K-means 算法与 DBSCAN 算法的运行时间基本一致。总的来看，前两种算法运行所需时间长于后两种算法运行所需时间，最大时间差(本文算法所需时间与 K-means 算法所需时间的差)为 0.45 s。

结束语 本文提出了一种新的图像无监督聚类算法，首先将图像编码成几何图中的顶点，图中连接边的权值以图像的相似度为度量，然后在几何图上构建扩散界面模型，并在模型中引入维度约简算子，采用循环迭代算法优化目标函数，基于最优的扩散界面完成图像的无监督聚类。在多组图像数据集上的实验结果表明，本文提出的聚类算法在图像无监督聚类方面的性能优于 K-means，DBSCAN 和 Spectral Clustering。

目前，本文提出的图像聚类算法还只能实现对数据集的

二值聚类,如何将算法有效地推广到多值聚类问题是今后需要进一步研究的重要方向。

参 考 文 献

[1] ALZU'BI A, AMIRA A, RAMZAN N. Semantic content-based image retrieval: A comprehensive study [J]. Journal of Visual Communication and Image Representation, 2015, 32: 20-54.

[2] ZHOU J X, LIU X, XU T W, et al. A new fusion approach for content based image retrieval with color histogram and local directional pattern [J]. International Journal of Machine Learning and Cybernetics, 2018, 9(4): 677-689.

[3] ZHANG F, KONG X W, NING F, et al. Image Retrieval by Extended Attribute Based on Web Search Amount [J]. Computer Engineering, 2017, 43 (9): 276-280, 287.

[4] CHENG Q, ZHANG Q, FU P, et al. A survey and analysis on automatic image annotation [J]. Pattern Recognition, 2018, 79: 242-259.

[5] FAN J, FAN Y. High dimensional classification using features annealed independence rules [J]. Annals of Statistics, 2008, 36(6): 2605-2637.

[6] BERTOZZI A L, FLENNER A. Diffuse interface models on graphs for classification of high dimensional data [J]. SIAM Review, 2016, 58(2): 293-328.

[7] BERTOZZI A L, LUO X, STUART A M, et al. Uncertainty quantification in graph-based classification of high dimensional data [J]. SIAM/ASA Journal on Uncertainty Quantification, 2018, 6(2): 568-595.

[8] ANDERSON D, MCFADDEN G B, WHEELER A A, et al. Diffuse-interface methods in fluid mechanics [J]. Annual Review of Fluid Mechanics, 1997, 30(1): 139-165.

[9] AGRAWAL A, KARNICK H. Unsupervised Image clustering [D]. Kanpur: Indian Institute of Technology, 2009: 1-6.

[10] WANG J, WANG J, SONG J, et al. Optimized cartesian k-means [J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(1): 180-192.

[11] GOWDA K C, KRISHNA G. Agglomerative clustering using the concept of mutual nearest neighbourhood [J]. Pattern Recognition, 1978, 10(2): 105-112.

[12] WANG W T, WU Y L, TANG C Y, et al. Adaptive density-based spatial clustering of applications with noise (DBSCAN) according to data[C]// 2015 International Conference on Machine Learning and Cybernetics (ICMLC). IEEE, 2015: 445-451.

[13] TRON R, ZHOU X, ESTEVES C, et al. Fast multi-image matching via density-based clustering[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 4057-4066.

[14] LEI J, RINALDO A. Consistency of spectral clustering in stochastic block models [J]. The Annals of Statistics, 2015, 43(1): 215-237.

[15] CHEN J, LI Z, HUANG B. Linear spectral clustering superpixel [J]. IEEE Transactions on Image Processing, 2017, 26(7): 3317-3330.

[16] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[R]. University of Toronto, 2009.

[17] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[C]// Proceedings of the IEEE. 1998: 2278-2324.

[18] EVERINGHAM M, VAN GOOL L, WILLIAMS C K I, et al. The pascal visual object classes (voc) challenge [J]. International Journal of Computer Vision, 2010, 88(2): 303-338.

[19] CVL Face Database. Computer vision laboratory, University of Ljubljana, Slovenia [EB/OL]. http://www.lrv.fri.uni-lj.si/facedb.html, 2005.

[20] SAMARIA F S, HARTER A C. Parameterisation of a stochastic model for human face identification[C]// Proceedings of 1994 IEEE Workshop on Applications of Computer Vision. IEEE, 1994: 138-142.

WANG Cheng-zhang, born in 1977, Ph.D, associate professor, master supervisor. His main research interests include machine learning, pattern recognition and big data analysis.