计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190400123

复杂高维数据的密度峰值快速搜索聚类算法

陈俊芬　张　明　赵佳成
河北大学数学与信息科学学院河北省机器学习与计算智能重点实验室　河北 保定 071002

摘　要　机器学习的无监督聚类算法已被广泛应用于各种目标识别任务。基于密度峰值的快速搜索聚类算法(DPC)能快速有效地确定聚类中心点和类个数,但在处理复杂分布形状的数据和高维图像数据时仍存在聚类中心点不容易确定、类数偏少等问题。为了提高其处理复杂高维数据的鲁棒性,文中提出了一种基于学习特征表示的密度峰值快速搜索聚类算法(AE-MDPC)。该算法采用无监督的自动编码器(AutoEncoder)学出数据的最优特征表示,结合能刻画数据全局一致性的流形相似性,提高了同类数据间的紧致性和不同类数据间的分离性,促使潜在类中心点的密度值成为局部最大。在4个人工数据集和4个真实图像数据集上将AE-MDPC与经典的K-means,DBSCAN,DPC算法以及结合了PCA的DPC算法进行比较。实验结果表明,在外部评价指标聚类精度、内部评价指标调整互信息和调整兰德指数上,AE-MDPC的聚类性能优于对比算法,而且提供了更好的可视化性能。总之,基于特征表示学习且结合流形距离的AE-MDPC算法能有效地处理复杂流形数据和高维图像数据。

关键词:聚类;密度峰值;DPC算法;特征表示;流形距离
中图法分类号　TP181

Clustering Algorithm by Fast Search and Find of Density Peaks for Complex High-dimensional Data

CHEN Jun-fen, ZHANG Ming and ZHAO Jia-cheng
Hebei Key Laboratory of Machine Learning and Computational Intelligence, College of Mathematics and Information Sciences, Hebei University, Baoding, Hebei 071002, China

Abstract　Unsupervised clustering in machine learning is widely applied in various object recognition tasks. A novel clustering algorithm based on density peaks (DPC) can find out cluster center points quickly in decision graph and the number of clusters. However, when dealing with the data of complex distribution shape and high-dimensional image data, there are still some problems in DPC algorithm, such as difficult to determine the cluster center points and few clusters. In order to improve its robustness in dealing with complex high-dimensional data, an improved DPC clustering algorithm (AE-MDPC) was presented, which employs an autoencoder, a kind of unsupervised learning method, to obtain the optimal feature representation from input data, and manifold similarity of pairwise data to describe the global consistence. The autoencoder can reduce feature noises via reducing dimension of the high-dimensional image data, whilst manifold distance can lead to the densities of the potential cluster centers become global peaks. AE-MDPC algorithm was compared with K-means, DBSCAN, DPC and DPC combined PCA on four artificial datasets and four real face image datasets. The experimental results demonstrate that AE-MDPC outperforms the other clustering algorithms on clustering accuracy, adjusted mutual information and adjusted rand index, meanwhile AE-MDPC provides better clustering visualization. Overall, the proposed AE-MDPC algorithm can effectively handle complex manifold data and high-dimensional image data.

Keywords　Clustering, Density peaks, DPC algorithm, Features representation, Manifold distance

1　引言

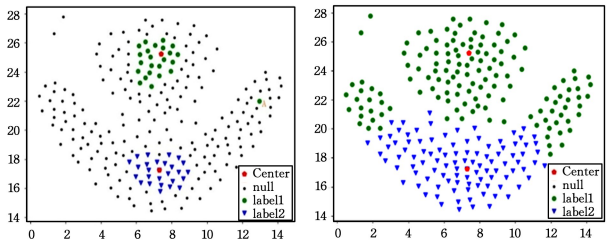
聚类的本质是依据数据的相似性将数据划分为若干类簇,使得同一类数据的相似性最大而不同类数据的相似性较小。Queen结合前人的研究成果,在1967年提出了基于空间划分的K-means聚类算法[1],该算法成为了最具代表性的聚类算法之一,直到2018年仍有许多工作围绕K-means算法展开。比如,Ismkhan[2]提出了结合类的信息增益和信息损失的I-K-means-+算法;Jia等[3]提出了能自动获取初始聚类中心和类数目的改进的K-means算法;Ester等于1996年提出了

到稿日期:2019-04-22　返修日期:2019-07-18　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:河北省自然科学基金(F2016201161);高层次创新人才科研启动经费项目
This work was supported by the Natural Science Foundation of Hebei Province, China (F2016201161) and Research Foundation for Advanced Scholars Program of Hebei University, China.
通信作者:陈俊芬(chenjunfen2010@126.com)

一种基于数据密度的聚类方法——DBSCAN，该算法能在包含噪声的数据集中识别出任意形状的簇，但是该算法的超参数和 MinPts 并不容易设置，导致其实用性受限[4-6]。

2014 年，*Science* 发表了一种新颖的能自动确定类数目和类中心的聚类算法[7]（为了叙述简便，简称为 DPC 算法）。该算法思路新颖，能快速发现任意形状数据的密度峰值点（即类中心）并完成聚类指派，在一些有挑战的数据集上取得了很好的聚类结果。DPC 算法中，数据间的相似度是基于欧氏距离的高斯核函数值（见第 2 节中的式(1)），并且截断距离 d_c 值的选取在一定程度上决定了聚类结果的优劣。这种相似性度量对凸分布或近似凸分布的数据能提供好的聚类效果，而对于一些特殊分布的数据，则需要精心选择 d_c 值，DPC 算法可能产生好的聚类结果。

图 1 展示了 DPC 算法在 Flame 数据集上的聚类中间结果和最终结果，可以看出聚类效果并不理想。由于 Flame 形状特殊，中间部分较紧凑地分布成圆形的点属于一类（花），向左右延伸较远的点属于另一类（叶子）。当设置 d_c 为所有样本点对的欧氏距离（从小到大排序后）的 2% 位置值时，聚类先按照正确的方向进行，但某一时刻样本点 A（见图 1(a)的右边叶子上的某点）被错误分类，接着离 A 最近且局部密度比点 A 小的样本点也被错误分类，这个错误会持续下去，从而得到最终的聚类结果（见图 1(b)），这就是所谓的“多米诺骨牌效应”[8]。造成这种错误的原因可能是归类原则“欧氏空间中两点距离最近时就将这两点归为同一类”，距离点 A 最近的是类 1 的数据点（圆点表示）而不是类 2 的数据点（三角点表示），因此点 A 被错误归类，并且还会继续扩散这种错误。



(a)聚类中间结果 (b)聚类最终结果

图 1 DPC 算法在 Flame 上的聚类中间结果和最终结果

Fig. 1 Intermediate and final results of clustering visualization using DPC algorithm on Flame dataset

另外，DPC 算法在处理高维图像数据时也表现出了一些力不从心，以像素为特征的图像数据维度很高，以像素为描述子会导致了图像间的差异不易察觉，相当于存在很多特征噪音，这种情况下的密度值分布比较均匀，很难出现局部峰值，导致了精准挑选聚类中心点困难和类个数偏少等问题。

为了弥补上述 DPC 算法的不足，本文提出一种基于流形距离且结合特征表示学习的密度峰值聚类算法（AE-MDPC）。其主要思想是将数据点看成图的顶点，位于同一流形上的顶点间有连接，权重是流形距离；而不同流形上的顶点间没有连接，其权重是无穷的。数据点的密度值等于到其他数据点的流形距离的高斯核函数值的和。将基于这个密度值和流形距离的密度峰值聚类算法记作 MDPC。自动编码器提取数据的最优特征表示，接着用 MDPC 算法进行聚类分析，该算法被记作 AE-MDPC。本文的贡献包括：

(1)用对数据分布敏感的流形距离替代欧氏距离，抓取有利于聚类的全局一致性信息，进而提高了 DPC 算法对复杂分布数据的聚类性能；

(2)用自动编码器提取图像数据的最优特征表示，减小特征属性的冗余信息对确定类中心点的影响，提高了同类数据的紧致性和不同类数据的分离性；

(3)改进了 DPC 算法在流形数据和图像数据上的聚类性能，进一步扩展了 DPC 算法的应用领域。

本文第 2 节简要介绍了 DPC 算法的思想和步骤，并且综述了相关文献；第 3 节详细描述了所提的 MDPC 算法和 AE-MDPC 算法；第 4 节给出了实验结果和实验分析；最后总结本文并展望下一步的工作。

2 相关工作

Rodriguez 等于 2014 年提出的密度峰值快速搜索聚类算法（DPC）[7]为聚类算法家族带来了一股清流。DPC 算法的技术核心包括以下 3 步。

(1)确定类中心点和类别个数。先根据式(1)—式(3)计算每一个数据点的密度值 ρ 和距离值 δ，然后绘制以 ρ 为横轴、δ 为纵轴的决策图（Decision Graph），手动框选出 ρ 和 δ 都较大的点作为聚类中心点，即选择位于决策图右上角的点作为类中心点，类别个数随即确定。

$$\rho_i=\sum_{j\neq i}\exp\left\{-\left(\frac{d_{ij}}{d_c}\right)^2\right\} \tag{1}$$

$$\rho_i=\sum_{j\neq i}\chi(d_{ij}-d_c) \tag{2}$$

$$\delta_i=\begin{cases}\min\limits_{j:\rho_j>\rho_i}(d_{ij}), & \exists j,\text{s.t. } \rho_j>\rho_i\\ \max\limits_{j}(d_{ij}), & \rho_i \text{ is the maximum}\end{cases} \tag{3}$$

其中，d_{ij} 是数据点 x_i 和 x_j 的欧氏距离，符号函数 $\chi(x)=\begin{cases}1, & x<0\\ 0, & \text{others}\end{cases}$，$d_c$ 为截断距离，手工设定其值时需考虑样本的近邻数大约是整个数据集规模的 1%～2%。密度最大的数据点的距离 δ 是数据点对欧氏距离的最大值，即密度最大的数据点，其距离也是最大的，该数据点一定是类中心点。其余点的距离 δ 等于到比自身密度值大的点的最近欧氏距离。

(2)对非类中心点进行类标分配。如果数据点 x_j 不是类中心点，则将其归入密度比 x_j 大且距离 x_j 最近的数据点 x_i 所在的类。该过程只需执行一次，没有迭代更新。

(3)判别类的核心和晕。对于类 C_k 内的数据点 x_j，如果它与类 $C_l(l\neq k)$ 内数据点的距离小于截断距离 d_c，则 x_j 称为 C_k 的边界点，C_k 的边界点中的最大密度值记为 ρ_b。类 C_k 中所有密度值大于 ρ_b 的数据点被称为该类簇的核心点，通过将其可视化为彩色点并与决策图中对应中心点的颜色一致来区别其类别；而密度值小于 ρ_b 的数据点为晕点，将其可视化为黑色点或灰色图像。

为了进一步提升 DPC 算法的聚类效果，研究者们提出了各种改进的 DPC 算法。改进主要集中在如何计算局部密度、如何分配类标以及噪声点的类标[9-14]。考虑到超参数 d_c 与

数据集密切相关，文献[15]基于信息熵理论提出了一种从原始数据集中自动获取截断距离 d_c 的新方法，而另外几种改进工作通过引入 k 近邻思想来计算局部密度值，即密度值的计算由局部 k 近邻点代替了全局 $n-1$ 个样本点[14,16-17]。文献[18]基于共享最近邻计算相似度，进一步改进了密度值的计算方法。基于 k 近邻的改进 DPC 算法尽管考虑了数据的局部结构信息对密度值的影响，能有效地解决变密度聚类问题，但是在消除 d_c 的同时却引入了新的超参数——近邻个数 k，该参数仍旧依赖于经验和交叉验证才能确定。另外，这种只用部分数据的亲和度来刻画一个数据的密度值的方式，会增加伪聚类中心点出现的概率。文献[19-20]在类标分配阶段也利用 k 近邻和模糊加权 k 近邻来改进 DPC 算法中简单的分配策略，这可能会引起错分累加的“多米诺骨牌现象”问题。文献[21]将流形距离和 DPC 算法相结合来处理复杂分布数据的聚类问题，当近邻连接 $k=3$ 时，新算法在具有挑战的 4 个数据集上得到了很好的聚类结果；而当 k 较大时，数据过于连通，会导致不同流形上的点也有最短路径，从而影响聚类性能。另外，文献[22]主要研究了属性噪音对聚类性能的影响，实验结果也表明聚类数目随着属性噪音的增加而减少。基于改进 DPC 算法对图像数据进行聚类时[8,14]，用主成分分析技术(PCA)进行降维处理。由于 PCA 通过线性变换(特征矩阵)将原始数据映射到特征空间，有限的特征表示能力会导致实验结果不甚理想。

3 AE-MDPC 算法

流形是局部具有欧几里得空间性质的空间，流形上两点间的距离并不遵守“欧氏空间中两点间的直线距离最短”这一准则。当两点非常接近（比如是近邻点）时，其流形距离等于欧氏距离；当两点相对较远时，流形距离是近邻点测地距离的累加[23]。

3.1 MDPC 算法

密度 ρ 和距离 δ 是 DPC 算法的两个关键指标。为了减小参数 d_c 对密度值的影响，也为了使密度值能反映出数据的全局一致性和局部拓扑性，本文采用式(4)、式(5)来计算这两个指标。

$$\rho_i=\sum_{j\neq i}\exp\{-dm_{ij}^2\} \tag{4}$$

$$\delta_i=\begin{cases}\min\limits_{j:\rho_j>\rho_i}(dm_{ij}^2), & \exists j,\text{s.t. } \rho_j>\rho_i\\ \max\limits_{j}(dm_{ij}^2), & \rho_i \text{ is the maximum}\end{cases} \tag{5}$$

其中，dm_{ij}^2 为数据点 x_i 和 x_j 的流形距离。式(4)相当于式(1)中令 $d_c=1$ 并用流形距离替换欧氏距离。

基于流形距离的 MDPC 算法如算法 1 所示。

算法 1 基于流形距离的 MDPC 算法

输入：数据集 $\mathbf{X}=\{x_1,x_2,\cdots,x_N\}$ 和近邻点个数 k

输出：聚类后的类标向量，聚类中心点

初始化：连接矩阵 **D** 和路径矩阵 **P**

　　调用弗洛伊德算法计算任意两点间的最短路径，从而得到任意两点间的流形距离和最优路径矩阵；

for 每一个数据点

　　计算 ρ，δ 值；

end

　　画出决策平面图；

　　框选出类中心点，确定类别个数；

for 每一个非聚类中心点

　　将其归入密度比它大且距离它最近的数据点所在的类；

end

3.2 AE-MDPC 算法

图像通常是高维数据，图像的分辨率决定其维度，像素特征很难表示出同类图像的相似性和不同类图像的差异性。因此，聚类图像数据的技术点包括图像数据的降维和特征抽取/表示。近年来，大量研究表明深层网络可以从数据中自动学习到更好的特征表示[24]。但是，随着网络层数的增多，基于 BP 算法的训练难度大大增加。为了降低训练复杂度，拥有对称结构的自动编码器(AE)得到广泛的应用。这种特殊结构的神经网络使用无监督分层训练策略得到编码器的网络参数，利用对称性得到解码器的网络参数[25]。本文利用 AE 提取图像数据的特征，同时对其进行降维。首先，设计一个对称结构的神经网络；其次，用贪婪的逐层训练策略来获取 AE 的未知参数($\boldsymbol{W}$，$\boldsymbol{b}$)，逐层训练的本质是用 BP 算法训练一个两层的神经网络；最后，在训练好的编码器后，连接 MDPC 聚类算法。基于自动编码器的聚类框架如图 2 所示。

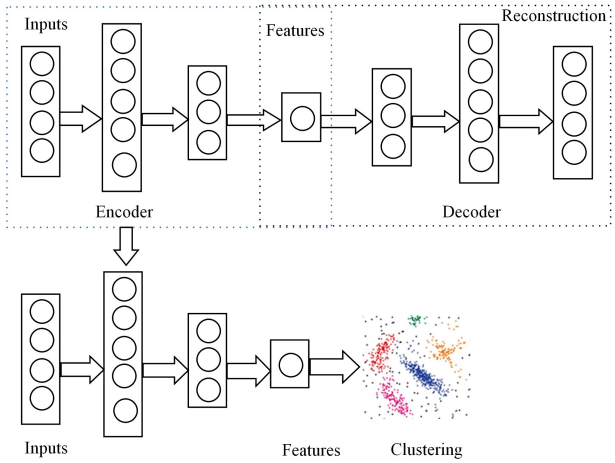


图 2 基于自动编码器的图像聚类框架

Fig. 2 Framework of image clustering based on autoencoder

基于自动编码器的密度峰值聚类算法 AE-MDPC 如算法 2 所示。

算法 2 基于自动编码器的 AE-MDPC 算法

输入：图像数据集

输出：聚类后的类标向量，聚类中心点

预处理：将每张图片转化成向量 $\mathbf{X}=\{x_1,x_2,\cdots,x_N\}$，进行数据缺失处理和去中心化处理。

训练 AE：设计一个对称结构的 AE；

　　利用 BP 算法逐层训练 AE；

　　得到最优特征表示 $\mathbf{Z}=\{z_1,z_2,\cdots,z_N\}$。

聚类：用 MDPC 算法对 $\mathbf{Z}=\{z_1,z_2,\cdots,z_N\}$ 进行聚类。

DPC 算法的第 3 步撤销了被判为噪声的类标，在聚类可视化图中将其标记为灰色。这一步可能把正常数据误判为噪声，如文献[7]提供的实验结果中 100 张 Olivetti 人脸图片中有 59 张被当成噪声而没有类标。这种宁忽略一百而不错分

一张的强硬策略提高了聚类的精准率和召回率(分配类标的41张中没有错分的图片,只是第6个人的8张图片被聚成了两类)。在识别或者标定应用中,很多数据没有给出归类类标,这种结果不被接受。因此,本文的MDPC和AE-MDPC算法主要包含以下两步。

(1)确定类中心点和类别个数。首先,对数据进行缺失值处理和数据归一化。将数据集看成图上的顶点集,计算顶点间的测地距离。根据式(4)和式(5)计算每个数据点的 ρ 值和 δ 值。绘制决策图,手动框选出位于右上角(即 ρ 和 δ 的高值)的点作为类中心点,同时确定类别个数。这种根据数据启发式确定类别个数的方式,大大降低了使用者的主观性。

(2)对非类中心点进行类标分配。分配非类中心点到局部密度比它高且与它距离最近的点所属的类。

为了实验的公平性,本文也只使用原DPC的前两步。假设本文实验中所用数据都是干净的,而对噪声数据的聚类是未来工作的一部分。

4 实验与分析

本节在人工数据集和真实数据集上来验证文中所提算法的聚类性能,并将其与K-means,DBSCAN和DPC算法进行比较。实验所需数据的详细信息如表1所列。实验环境:硬件平台为Intel®Core™ i5-4590CPU@3.3 GHz处理器,8.0 GBRAM;编程环境为Anaconda3-5.1.0。

已知数据集合 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$,真实类划分 $U=\{U_1,U_2,\cdots,U_T\}$,而 $V=\{V_1,V_2,\cdots,V_C\}$ 是聚类结果。本文采用聚类精度(*ACC*)、调整互信息系数(*AMI*)和调整兰德系数(*ARI*)来度量聚类性能[26],它们的定义分别如下。

定义1(聚类精度) *ACC* 是正确聚类的数据点个数 N_1 与数据点总数 N 的比值,即:

$$ACC=\frac{N_1}{N} \tag{6}$$

定义2(调整互信息) 基于信息论的 *AMI* 的定义如下:

$$AMI(U,V)=\frac{MI(U,V)-E\{MI(U,V)\}}{\max\{H(U),H(V)\}-E\{MI(U,V)\}} \tag{7}$$

其中,$H(U)$和$H(V)$是U和V的熵,$MI(U,V)$是互信息,$E\{MI(U,V)\}$是互信息的期望值,每一部分的计算详见文献[8]。

定义3(调整兰德系数) 设a是属于U的同类且属于V的同类的数据对数目,b是属于U的同类但属于V的不同类的数据对数目,c是属于U的不同类而属于V的同类的数据对数目,d是属于U的不同类且属于V的不同类的数据对数目,则调整兰德系数 *ARI* 的定义为:

$$ARI=\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)} \tag{8}$$

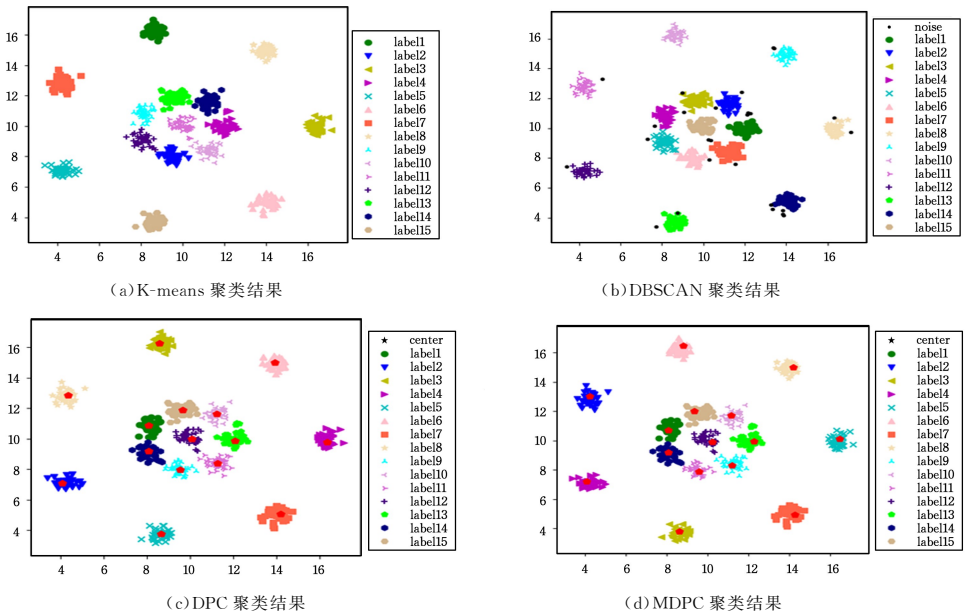
表1 4个人工数据集和4个图像数据的信息

Table 1 Information description of four artificial datasets and four image databases

数据集	样例个数	属性个数	类别数
Flame	240	2	2
Path-based2	312	2	3
Jain	373	2	2
R15	600	2	15
CAS-PEAL-R1	200	480×360	40
IMM	240	640×480	40
BioID	156	384×286	26
MNIST	5000	28×28	10

4.1 人工数据集

本组实验分别在Path-based2,Jain,Flame和R15这4个人工数据集上进行,主要检验基于流形距离的MDPC算法的聚类性能,结果如图3所示。



(a)K-means聚类结果 (b)DBSCAN聚类结果

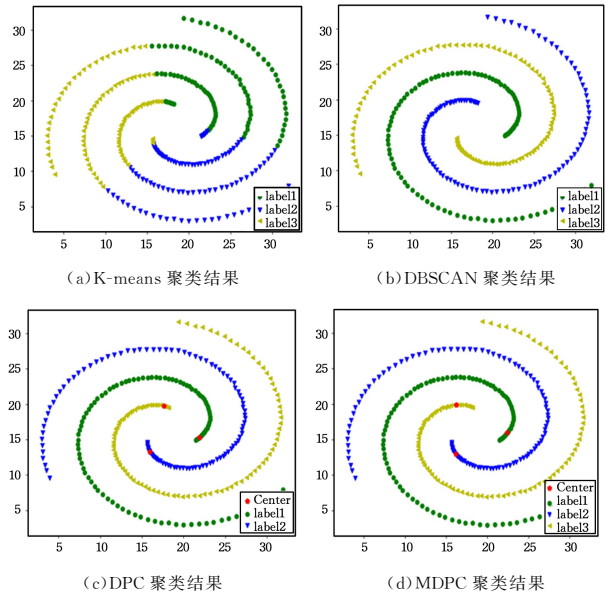
(c)DPC聚类结果 (d)MDPC聚类结果

图3 4种算法在R15数据集上的聚类结果

Fig. 3 Clustering results of four algorithms on R15 dataset

从图3中可以看出,这4种算法都提供了很好的聚类可视化结果,其中K-means,DPC和MDPC算法的 *ACC*,*AMI* 和 *ARI* 都为1,只是DBSCAN算法将一些样本点归为了噪声(见图3(b)中的小圆点)。

如图 4 所示，K-means 算法对 Path-based2 的聚类效果不佳，未能正确聚类，而 MDPC，DPC 和 DBSCAN 算法都能正确聚类。这一结果显示：基于密度的 MDPC，DPC 和 DBSCAN 算法对流形数据的聚类性能优于基于空间划分的 K-means 算法。



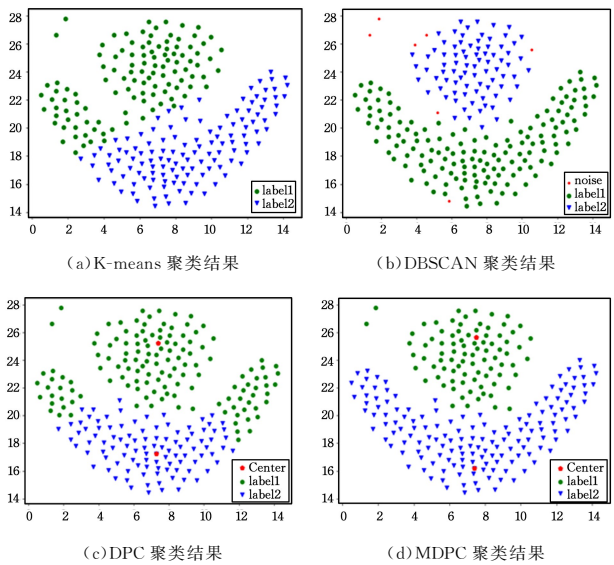
(a)K-means 聚类结果 (b)DBSCAN 聚类结果

(c)DPC 聚类结果 (d)MDPC 聚类结果

图 4 4 种算法在 Path-based2 数据集上的聚类结果

Fig. 4 Clustering results of four algorithms on Path-based2 dataset

图 5 给出了这 4 种算法在 Flame 数据集上的聚类结果。



(a)K-means 聚类结果 (b)DBSCAN 聚类结果

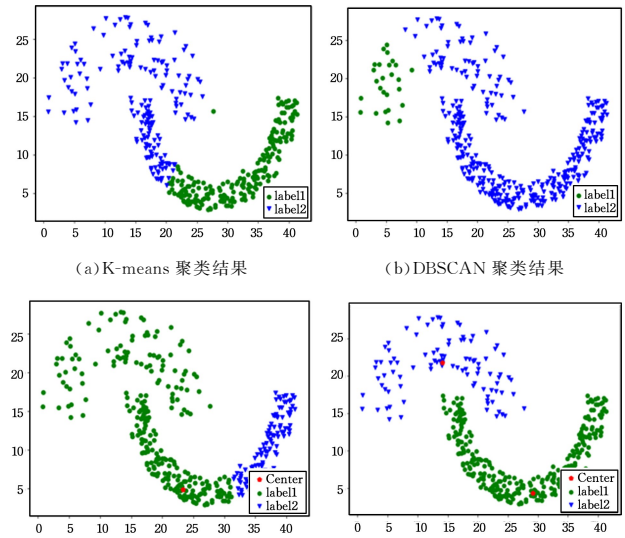
(c)DPC 聚类结果 (d)MDPC 聚类结果

图 5 4 种算法在 Flame 数据集上的聚类结果

Fig. 5 Clustering results of four algorithms on Flame dataset

K-means 依然不能正确聚类，DBSCAN 算法能提供正确聚类，但是一些点被判别为噪声点(标注为小圆点)。正如之前的分析，DPC 算法不能正确聚类，产生了“多米诺骨牌效应”。MDPC 算法用流形距离的高斯核函数值度量两个数据点的相似性，避免了原 DPC 算法的错误划分，得到了正确的聚类结果。

由图 6 可知，K-means 和 DBSCAN 算法在 Jain 数据集上的聚类结果不好，DPC 算法也不尽人意(包括没有找到正确的类中心点)，只有 MDPC 算法提供了正确聚类。这表明基于流形距离的 MDPC 算法比基于欧氏距离的 DPC 算法更能处理这种分布的数据。



(a)K-means 聚类结果 (b)DBSCAN 聚类结果

(c)DPC 聚类结果 (d)MDPC 聚类结果

图 6 4 种算法在 Jain 数据集上的聚类结果

Fig. 6 Clustering results of four algorithms on Jain dataset

由表 2 可以看出，MDPC 算法在这 4 个数据集上的聚类指标值都是 1，聚类性能最好；DPC 算法在 R15 和 Path-based2 这 2 个数据集上的聚类指标值都是 1；DBSCAN 只在 Path-based2 上的聚类指标值为 1；K-means 算法虽然只在 R15 上的指标值为 1，但是在 Flame 和 Jain 上的聚类指标均高于 DPC 算法。从聚类数值指标来看，MDPC 算法最优，DBSCAN 和 K-means 算法次之，而 DPC 算法最差。

本文使用的 DPC 算法由于没有取消一些疑似噪音的类标，也就是说没有执行原 DPC 算法的第 3 步，导致聚类精度大大降低，如在 Flame 和 Jain 上的指标值都低于 K-means。本文的 MDPC 算法也没有取消一些疑似噪音的类标，但是在这 4 个数据集上的聚类性能都优于 DPC 算法。可见，用流形距离替代欧氏距离大大提高了聚类性能，验证了流形距离的有效性，即本文的 MDPC 算法比 DPC 和 DBCSAN 算法更擅长对任意形状的数据进行聚类。

表 2 4 种聚类算法在 4 个人工数据集上的性能对比

Table 2 Comparison performance of four algorithms on four artificial datasets

Algorithms	R15			Path-base2			Jain			Flame		
	ACC	*AMI*	*ARI*	*ACC*	*AMI*	*ARI*	*ACC*	*AMI*	*ARI*	*ACC*	*AMI*	*ARI*
K-means	1.000	1.000	1.000	0.433	0.005	0.005	0.789	0.336	0.324	0.863	0.386	0.453
DBSCAN	0.998	0.928	0.932	1.000	1.000	1.000	0.739	0.173	0.256	0.971	0.773	0.917
DPC	1.000	1.000	1.000	1.000	1.000	1.000	0.550	0.175	−0.034	0.787	0.403	0.327
MDPC	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

4.2 人脸数据集

在3个人脸数据库上测试基于特征表示学习的AE-MDPC算法的聚类性能。CAS-PEAL-R1[1]是中科院计算所采集的人脸数据库的子库，从中随机选40人，每人5幅480×360像素的图片。IMM[2]包括40人，每人6张像素为640×480的图片。BioID[3]包含在各种光照和复杂背景下的灰度面部图像，实验随机选26人，每人6张像素为384×286图片。

首先，将每幅图像缩成32×24像素；其次，设置AE的编码器网络结构为768-1024-256-z，这3个数据集上的z分别取25，18和58；然后，用AE提取特征和降维，并与PCA降维进行比较。算法的聚类性能如表3所列，而3种算法在CAS-PEAL-R1上的聚类可视化结果如图7所示。

表3 3种聚类算法在3个人脸数据集上的性能对比

Table 3 Clustering performance comparison of three algorithms on three face datasets

Algorithm	BioID			CAS-PEAL-R1			IMM		
	ACC	*AMI*	*ARI*	*ACC*	*AMI*	*ARI*	*ACC*	*AMI*	*ARI*
DPC	0.686	0.732	0.597	0.375	0.484	0.244	0.375	0.290	0.123
PCA-DPC	0.916	0.924	0.879	0.720	0.793	0.693	0.442	0.343	0.161
AE-MDPC	0.917	0.935	0.892	0.800	0.836	0.758	0.450	0.422	0.273

(a)DPC 聚类结果 (b)PCA-DPC 聚类结果 (c)AE-MDPC 聚类结果

图7 3种算法在CAS-PEAL-R1数据集上的聚类结果

Fig. 7 Clustering results of three algorithms on CAS-PEAL-R1 database

由图7(a)可以看出，DPC算法的聚类效果较差，不同人的图片有着相同的类标，如将7个人的35张图片和另外2个人的3张图片都标识为39；同一人的5张图片却被分成2类或者3类，如第1个人的5张图片被标识为39，4和37。图7(b)是经过PCA降维后再进行DPC聚类的结果，聚类效果得到了很好的改善，共发生5次2个人的10张图片被分配同一类标的情况，也有同一人的5张图片被分成2类或者3类，但是这两种错分现象有了大幅减少。图7(c)是经过自动编码器提取特征和降维后再进行MDPC聚类的结果，进一步改善了聚类结果，只发生了2次2个人的10张图片被分配相同类标的情况，并且同一人的5张图片只被分成了2类。AE-MDPC算法准确地发现了类中心，聚类效果远优于DPC算法。

对比图7(a)和图7(b)发现，PCA借助特征向量得到人

1) http://www.jdl.ac.cn/peal/index.html

2) http://www.imm.dtu.dk/~aam/aamexplorer/

3) http://www.humanscan.de/support/downloads/facedb.php

脸图片的整体特征表示，这种特征表示提高了区分性，表现在多个人的图片被分配相同的类标现象大大减少，从而提高了聚类精度。对比图 7(b)和图 7(c)可以发现，非线性映射的 AE 学习到人脸图片上更具有辨识力的特征，进一步提高了聚类性能。仔细观察不难发现，在组内图片中人脸大小不同的情况下，PCA-DPC 和 AE-MDPC 的归类会发生错误。另外，表情和发型也会影响聚类结果。最后，本文使用的 DPC 没有判别噪音和取消其类标，因此对图像数据聚类的精度不高。但是，基于图像维度约简的 PCA-DPC 和 AE-MDPC 却受此影响较小。

由表 3 可知，在这 3 组人脸数据集上，AE-MDPC 和 PCA-DPC 算法的聚类性能相比 DPC 算法都有不同程度的提升。在 BioID 和 CAS-PEAL-R1 上，AE-MDPC 算法相比 DPC 算法提升了 23%和 41.5%的聚类精度，而在 IMM 上提升了 7.5%。AE-MDPC 的聚类精度在 IMM 上提升不大的原因是该数据集上的人脸有向右较大的姿态，并且一些人之间的差异很难辨识，从而导致自动编码器没有提取到很好的特征表示。

实验结果表明，AE-MDPC 算法较为有效地解决了 DPC 算法处理高维数据时聚类性能不佳的问题。针对缩小人脸图片尺寸可能影响算法聚类性能的问题，在 3 个人脸数据集上对 AE-MDPC 算法做了一系列实验。AE-MDPC 算法在不同尺寸图片上的聚类性能如表 4－表 6 所列。可以看到，对于 BioID 图片，当其像素为 48×36 时，AE-MDPC 算法的聚类精度最高，为 0.961；对于 CAS-PEAL-R1，AE-MDPC 算法最好的聚类精度对应的图片像素是 60×45；IMM 的图片缩小到 64×48 像素时，AE-MDPC 的聚类精度是 0.546。在 3 种算法的聚类性能对比实验(见表 3)中，采用 32×24 小片是为了平衡 DPC 的计算复杂度，因此牺牲了 AE-MDPC 算法的一部分聚类精度。

表 4 不同尺寸的 BioID 人脸上的聚类性能

Table 4 Clustering performances of different size of BioID face

指标	96×72	48×36	32×24
ACC	0.904	**0.961**	0.917
AMI	0.908	0.949	0.935
ARI	0.857	0.922	0.892

表 5 不同尺寸的 CAS-PEAL-R1 人脸上的聚类性能

Table 5 Clustering performances of different size of CAS-PEAL-R1 face

指标	120×90	60×45	32×24
ACC	0.760	**0.880**	0.800
AMI	0.790	0.883	0.836
ARI	0.662	0.834	0.758

表 6 不同尺寸的 IMM 人脸上的聚类性能

Table 6 Clustering performances of different size IMM face

指标	320×240	64×48	32×24
ACC	0.375	**0.546**	0.450
AMI	0.293	0.560	0.422
ARI	0.167	0.435	0.273

4.3 MNIST 数据集

本节在手写数字数据集 MNIST 上进行实验，以进一步检验 AE-MDPC 算法对大数据集聚类的有效性。MNIST 数据集包含 10 个手写阿拉伯数字，共有 70000 张 28×28 的图片[27]，是机器学习领域常用的数据集之一。为了确保实验在 CPU 机器上顺利运行，随机选取了 5000 张图片进行实验。AE 编码器的结构是 784-500-500-2000-10，隐含层节点个数远大于输入的维度能使网络稀疏化提取更加有用的特征。

如表 7 所列，本文提出的 AE-MDPC 优于 DPC 和 PCA-DPC 算法，且在这 3 个聚类指标上均有很大幅度的提升。可见，AE 的非线性特征表示比 PCA 的线性特征嵌入更能挖掘出适合图像聚类的特征，这再次验证了本文提出的 AE-MDPC 算法对图像数据的聚类更具有活力。

表 7 3 种聚类算法在 MNIST 数据集上的性能对比

Table 7 Comparison three algorithms on MNIST dataset

Algorithms	*ACC*	*AMI*	*ARI*
DPC	0.356	0.273	0.163
PCA-DPC	0.249	0.150	0.040
AE-MDPC	0.725	0.713	0.621

结束语 本文针对 DPC 算法在处理复杂分布的数据和高维图像数据时聚类性能不佳的问题进行了改进，提出了一种基于最优特征表示并结合流形距离的 AE-MDPC 算法。将基于流行距离的高斯核函数值的累加和作为密度值描述了数据点的全局一致性，使潜在类中心的密度值变为局部最大，从而被选为类中心点。自动编码器学到最优特征表示，同时也剔除了属性噪声。本文提出的 AE-MDPC 算法提高了密度值的分离度，避免了 DPC 算法的“多米诺骨牌效应”，在经典人工数据集和真实图像数据集上的聚类性能均优于 DPC 算法。

为了进一步挖掘图像数据的有价值信息，提取有利于聚类的特征，下一步工作将利用具有不确定性的聚类结果对逐层训练好的 AE 进行监督式微调；另外，将探索卷积编码器以进一步提高所得特征的辨识能力。

参 考 文 献

[1] QUEEN J M. Some methods for classification and analysis of multivariate observations[C]// Proc of the fifth Berkeley symposium on mathematical statistics and probability. Oakland: Lucien Marie Le Cam, 1967: 281-297.

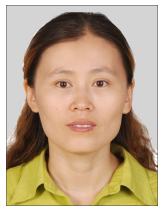
[2] ISMKHAN H. I-k-means-+: An Iterative Clustering Algorithm Based on an Enhanced Version of the k-means[J]. Pattern Recognition, 2018, 79: 402-413.

[3] JIA R Y, LI Y G. K-means algorithm for self-determination of cluster number and initial center[J]. Computer Engineering and Application, 2018, 54 (7): 152-158.

[4] BIRANT D, KUT A. ST-DBSCAN: An algorithm for clustering spatial-temporal data [J]. Data & Knowledge Engineering, 2007, 60(1): 208-221.

[5] HOU J, GAO H J, LI X L. DSets-DBSCAN: A Parameter-Free Clustering Algorithm[J]. IEEE Transactions on Image Proces-

sing, 2016, 25(7): 3182-3193.
[6] TRAN T N, DRAB K, DASZYKOWSKI M. Revised DBSCAN algorithm to cluster data with dense adjacent clusters[J]. Chemometrics & Intelligent Laboratory Systems, 2013, 120(2): 92-96.
[7] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.
[8] XIE J Y, GAO H C, XIE W X. K Nearest Neighbor Optimized Density Peak Fast Search Clustering Algorithms[J]. Chinese Science: Information Science, 2016, 46 (2): 258-280.
[9] MEHMOOD R, BIE R, DAWOOD H, et al. Fuzzy Clustering by Fast Search and Find of Density Peaks[J]. Personal & Ubiquitous Computing, 2016, 20(5): 785-793.
[10] LI C Y, DING G Y, WANG D K. Clustering by Fast Search and Find of Density Peaks with Data Field[J]. Chinese Journal of Electronics, 2016, 25(3): 397-402.
[11] XU J, WANG G Y, DENG W H. DenPEHC: Density Peak based Efficient Hierarchical Clustering [J]. Information Sciences, 2016, 373(12): 200-218.
[12] LU Y H, XIA C. Optimal K-Nearest Neighbor and Local Density Clustering Algorithms for Uncertain Data[J]. Control and Decision-making, 2016, 31(3): 541-546.
[13] WANG P F, YANG Y W, KE Y Q. Research on Optimization of fast clustering algorithm for peak density[J]. Computer Engineering and Science, 2018, 40(8): 1503-1510.
[14] XIE J Y, QU Y N. K-medoids clustering algorithm for initial center of peak density optimization [J]. Computer Science and Exploration, 2016, 10(2): 230-247.
[15] WANG S L, WANG D K, LI C Y, et al. Clustering by fast search and find of density peaks with data field[J]. Chinese Journal of Electronics, 2016, 25(3): 397-402.
[16] XIE J Y, GAO H C, XIE W X, et al. Robust clustering by detecting density peaks and assigning points based on fuzzy weighted K-nearest neighbors[J]. Information Sciences, 2016, 354(C): 19-40.
[17] DU M J, DING S F, JIA H J. Study on density peaks clustering based on k-nearest neighbors and principal component analysis [J]. Knowledge-Based Systems, 2016, 99: 135-145.
[18] LIU R, WANG H, YU X M. Shared-nearest-neighbor-based Clustering by Fast Search and Find of Density Peaks[J]. Information Sciences, 2018, 450: 200-226.
[19] GOTTUMUKKAL R. An improved face recognition technique based on modular PCA approach [J]. Pattern Recognit Lett, 2004, 25(4): 429-436.
[20] KE Y, SUKTHANKAR R. PCA-SIFT: a more distinctive representation for local image descriptors [C] // Proc of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington: IEEE, 2004.
[21] ZHANG J Q, ZHANG H Y. Fast search clustering algorithm for density peak based on manifold distance[J]. Computer Knowledge and Technology, 2017, 13(2): 179-182.
[22] YANG H, FU Y, FAN D. The effect of noise characteristics on the internal validity of clustering[J]. Computer Science, 2018, 45(7): 22-30.
[23] TENENBAUM J B, SILVA V D E, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319-2323.
[24] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors [J]. Nature, 1986, 323(6088): 533-536.
[25] YOSHUA B, LAMBLIN P, POPOVICI D, et al. Greedy layer-wise training of deep networks[C] // Advances in Neural Information Processing Systems (NIPS'06). 2006: 153-160.
[26] VINH N X, EPPS J, BAILEY J. Bibliometrics: Information theoretic measures for clusterings comparison[C] // Proc of the International Conference on Machine Learning. New York: ACM, 2010, 2837-2854.
[27] YANN L C, L'EON B, YOSHUA B, et al. Gradient-based learning applied to document recognition [J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.



CHEN Jun-fen, born in 1976. Ph.D, associate professor, master supervisor, is member of China Computer Federation. Her main research interests include data mining, machine learning and image processing.