一种基于共享近邻的密度峰值聚类算法

刘奕志 程汝峰 梁永全
(山东科技大学计算机科学与工程学院 山东 青岛 266590)
(山东省智慧矿山信息技术重点实验室 山东 青岛 266590)

摘 要 基于加权K近邻的密度峰值发现算法(FKNN-DPC)是一种简单、高效的聚类算法,能够自动发现簇中心,并采用加权K近邻的思想快速、准确地完成对非簇中心样本的分配,在各种规模、任意维度、任意形状的数据集上都能得到高质量的聚类结果,但其样本分配策略中的权重仅考虑了样本间的欧氏距离。文中提出了一种基于共享近邻的相似度度量方式,并以此相似度改进样本分配策略,使得样本的分配更符合真实的簇归属情况,从而提高聚类质量。在UCI真实数据集上进行实验,并将所提算法与K-means,DBSCAN,AP,DPC,FKNN-DPC等算法进行对比,验证了其有效性。

关键词 聚类,共享近邻,相似性度量,密度峰值
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.02.022

Clustering Algorithm Based on Shared Nearest Neighbors and Density Peaks

LIU Yi-zhi CHENG Ru-feng LIANG Yong-quan
(College of Information Science and Engineering, Shandong University of Science and Technology, Qingdao, Shandong 266590, China)
(Provincial Key Laboratory for Information Technology of Wisdom Mining of Shandong Province, Qingdao, Shandong 266590, China)

Abstract Robust clustering by detecting density peaks and assigning points based on fuzzy weighted K-nearest neighbors(FKNN-DPC) is a simple and efficient clustering algorithm, which can automatically detect the cluster center and assign the non-cluster center sample based on weighted K-nearest neighbors quickly and accurately. It is powerful in recognizing high quality cluster in any scale , any dimension, any size and any shape of the data set, but the weight calculation in assigning strategies only considers the Euclidean distance between samples. In this paper, a similarity measure based on shared neighborhood was proposed, and the sample assigning strategy was improved by this similarity, so that the cluster is more consistent with the real attribution, thus improving the clustering quality. The effectiveness of the algorithm is verified by comparing the experiments on the UCI real data set with the K-means, DBSCAN, AP, DPC, and FKNN-DPC algorithm.

Keywords Clustering, Shared nearest neighbors, Similarity measure, Density peak

1 引言

随着信息技术和计算机技术的飞速发展,数据量急剧增加。为了帮助用户从这些海量数据中提取出有价值的知识,数据挖掘技术应运而生。聚类分析技术作为数据挖掘领域的一个重要分支,逐渐成为研究的热点并被广泛应用于模式识别、图像处理、生物信息分析等领域。聚类是一个根据样本间的相似性将数据集划分成簇的过程,使得簇内的对象最大程度地相似,同时不同簇间的对象最大程度地相异[1-4]。根据聚类原理,聚类算法大致可分为5类[5]:基于划分的聚类算法、基于层次的聚类算法、基于网格的聚类算法、基于密度的聚类算法和基于模型的聚类算法。

基于划分的聚类算法以K-means[6]为代表,其思想是根据距离将样本划分到最近的簇中心所代表的簇,并迭代更新簇中心。该算法具有简单高效的优点,但只能发现类球状簇,且其易受初始中心点选择和噪声数据的影响。

依据层次分解的方式,基于层次的聚类算法可分为凝聚的和分裂的两类,其中比较有代表性的算法有CURE[7]和CHAMELEON。CURE算法不用单个中心或对象来代表一个簇,而是选择数据空间中数目固定且具有代表性的点来共

到稿日期:2017-04-08 返修日期:2017-06-11 本文受国家自然科学基金(61203305,61433012),山东省自然科学基金(ZR2015FM013),山东省重点研发计划(攻关)(2016GSF120012),山东省"泰山学者"攀登计划资助。
刘奕志(1993-),男,硕士,CCF学生会员,主要研究方向为数据挖掘与机器学习;**程汝峰**(1992-),男,硕士,CCF学生会员,主要研究方向为数据挖掘与机器学习;**梁永全**(1968-),男,教授,博士生导师,CCF会员,主要研究方向为分布式人工智能、数据挖掘与机器学习、电子商务等,E-mail:lyq@sdust.edu.cn(通信作者)。

同代表相应的簇,从而识别任意形状的簇,且对噪声数据不敏感;但该算法的效果受数据采样质量的影响较大。CHAMELEON 算法[8]将数据建模为图,通过相对互连性和相对近邻性两个指标来控制簇的分裂与合并,以发现高质量的任意形状的簇;但其易受噪声数据的影响且参数难以确定。

基于密度的聚类算法从数据对象的分布密度出发,连接密度足够大的相邻区域,可以发现任意形状的簇,并能有效处理噪声数据。

顾名思义,基于密度的聚类算法是根据数据对象的分布密度将数据集划分为簇的过程。DBSCAN[9]是一个典型的基于密度的聚类方法,它将核心对象定义为邻域半径 eps 内包含最少数据对象个数 $MinPts$ 的数据,并通过不断扩展与核心对象可达的数据对象来生成簇。该算法具有能够发现任意形状簇、对噪声数据不敏感的优点,但参数的设置缺乏理论依据。

基于模型的聚类算法为每个簇建立一个数学模型,并通过调整使得该模型能够较好地拟合数据的分布。OPEHCA[10]是一种最新的基于模型的聚类算法,它结合了树状层次结构和遗传算法的思想,通过结合层次的概念在不同层次上为每个簇找到不同的分布模型,并通过遗传算法寻找最优的模型。

基于网格的聚类算法从对数据空间进行划分的角度出发,将空间划分为有限数目的单元以构成一个可以进行聚类分析的网络结构。CLIQUE[11]是一个经典的基于网格的聚类算法,结合了网格与密度的思想,在处理大规模高维数据时具有较好的效果。

近邻传播算法[12]是一种基于数据对象间信息传递的聚类算法,通过反复迭代交换近邻样本间的信息,寻找最优的代表点与簇结果集合,并寻找使所有数据对象与最近簇代表点样本的相似度之和最大且最优的簇代表点集合。该算法具有简单、高效的特点,且不需要有关簇数量的先验知识,但不能发现任意形状的簇。

近年来,基于密度的聚类算法因具有能够发现任意形状的簇以及所需有关数据集的先验知识少的特点,得到了许多研究者的关注。Alex 和 Alessandro 提出了一种基于密度与距离的聚类算法:密度峰值聚类算法(DPC)[13]。该算法假设理想簇的中心是局部密度相对较大且距离比其密度大的样本相对较远的样本点,能够自动确定类簇数量与各类簇中心,并且可以得到任意形状的簇,在各种规模的数据集上都取得了很好的聚类结果。此外,还出现了许多基于 DPC 的改进算法[14-18]。

本文结合共享近邻(SNN)的思想,提出了一种新的相似度度量方式。该方式更能体现样本的真实簇归属情况,并利用此相似度改进了检测密度峰值和基于模糊加权 K-近邻分配的鲁棒聚类算法(FKNN-DPC)[17]的分配策略,使样本对于其真正所属的簇有更大的归属度,从而提高了聚类质量。本文在 9 个 UCI 真实数据集上进行实验,将所提算法与 FKNN-DPC,DPC,K-means,DBSCAN,AP 算法的聚类结果进行了对比,证实了其有效性。本文第 2 节回顾了密度峰值算法的原理以及 FKNN-DPC 算法对于 DPC 算法样本分配策略的改进;第 3 节给出了结合共享近邻的相似度的定义并详细描述了改进后的 SNN-FKNN-DPC 算法;第 4 节展示了改进后的算法在 9 个 UCI 数据集上的聚类结果以及与几种经典聚类算法的对比;最后总结全文。

2 FKNN-DPC 的原理

本节详细介绍了 DPC 算法的核心思想和 FKNN-DPC 算法中的分配策略。

DPC 算法假设理想的簇中心是局部密度相对较大(密度峰值)且距离密度比它大的样本相对较远的样本点。为了找到同时满足这两个条件的样本,DPC 算法引入了样本 i 的局部密度 ρ_i 和样本 i 到密度比它大的样本的最小距离 δ_i,其中 ρ_i 的定义如式(1)和式(2)所示,δ_i 的定义如式(5)所示。

$$\rho_i=\sum_{j\neq i}\chi(d_{ij}-d_c) \tag{1}$$

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

在式(1)和式(2)中,d_c 为截断距离,d_{ij} 为样本 i 与样本 j 之间的欧氏距离,其定义如式(3)所示。

$$d_{ij}=\sqrt{\sum_{k=1}^{n}(x_{ik}-x_{jk})^2} \tag{3}$$

$$\chi(x)=\begin{cases}1, & x>0\\ 0, & x<0\end{cases} \tag{4}$$

式(1)为数据集的规模较大时样本 i 的局部密度计算方式;当数据集的规模较小时,为减小截断距离的选择对算法的影响,DPC 算法采用高斯核函数来估计样本 i 的局部密度,如式(2)所示。

$$\delta_i=\begin{cases}\min\limits_{j:\rho_j>\rho_i}(d_{ij}), & \exists j,\rho_j>\rho_i\\ \max\limits_{j}(d_{ij}), & \text{otherwise}\end{cases} \tag{5}$$

对于密度最高的样本 i,$\delta_i=\max_j(d_{ij})$。

DPC 算法构造密度 ρ 相对于距离 δ 的决策图,选择使 ρ 和 δ 同时取较大值的样本作为簇中心。在选定簇中心以后,DPC 算法按照密度从大到小的顺序将剩余样本依次分配到比它密度大且与其距离最近的样本所在的簇,一步完成所有样本的分配,十分高效。

然而,此分配策略存在不可忽视的弊端:一旦出现错分的样本,便会影响密度比它低的样本的分配,使错误不断扩大,类似“多米诺骨牌”效应。针对此问题,FKNN-DPC 基于 K 近邻的思想提出了一种新的样本分配策略。首先根据式(6)—式(8)剔除噪声点,然后执行样本分配策略。策略分为两步:策略 1 基于图的连通性,采用广度优先搜索从寻找各类簇中心的 K 近邻样本开始对样本进行类簇分配;策略 2 根据加权 K 近邻对未扩展到的样本点和噪声点进行分配。

$$\delta_i^K=\max_{j\in KNN_i}d_{ij} \tag{6}$$

$$\tau=\frac{1}{N}\sum_{i=1}^{N}\delta_i^K \tag{7}$$

$$Outliers=\{o\mid\delta_o^K>\tau\} \tag{8}$$

对于每一个未分配的样本，根据式(9)－式(11)计算其关于每个簇的归属度。其中，式(9)为样本 i 与样本 j 之间的相似度，表示样本 i 与样本 j 之间的距离越远时其相似度越低；式(10)为样本 i 与样本 j 之间的相似度关于样本 j 的 k 近邻相似度的归一化值；式(1)为样本 i 关于簇 c 的归属度。

$$w_{ij}=\frac{1}{1+d_{ij}} \tag{9}$$

$$\gamma_{ij}=\frac{w_{ij}}{\sum\limits_{l\in kNN_j} w_{jl}} \tag{10}$$

$$p_i^c=\sum_{j\in KNN_i,y_j=c}\gamma_{ij}*w_{ij} \tag{11}$$

3　核心工作

本节针对 FKNN-DPC 分配策略中存在的缺陷，提出了基于共享近邻的相似度，并详细说明了基于此相似度改进的新样本分配策略。

3.1　基于共享近邻的相似度

FKNN-DPC 算法提出的新分配策略虽然能有效解决 DPC 分配策略中样本错误分配的影响不断放大的问题，但其相似度度量方式仅考虑了样本间的距离，无法反映真实簇的归属对样本间相似度的影响。

如图 1 所示，图中圆形点与方形点分属于不同的簇，并且样本 a 和 b 之间的距离与样本 a 和 c 之间的距离相等，即 $d_{ab}=d_{ac}$，如果按照 FKNN-DPC 中的相似度计算方式，则 $w_{ab}=w_{ac}$；然而根据样本 a,b,c 的真实归属情况，样本 a 和 b 属于同一个簇，样本 a 和 c 属于不同的簇，在对样本 a 进行分配时，样本 b 对样本 a 的影响应该大于样本 c 对样本 a 的影响，即样本 a 和 b 之间的相似度应大于样本 a 和 c 之间的相似度。因此，衡量两个样本之间的相似度时，仅考虑距离是不全面的。本文基于共享近邻的思想提出了一种新的相似度计算方式，如式(12)所示：

$$s_{ij}=\exp\left(-\frac{d_{ij}}{SNN_{ij}+1}\right) \tag{12}$$

其中，SNN_{ij} 的定义如下：

$$SNN_{ij}=\mathrm{card}(\varepsilon NN(i)\cap\varepsilon NN(j)) \tag{13}$$

其中，$\varepsilon NN(i)$ 表示位于样本 i 的以 ε 为半径的邻域内的样本集合，$\mathrm{card}(X)$ 表示集合 X 中包含的元素数量。式(12)中的分母加 1 是为了避免出现分母为 0 的情况。

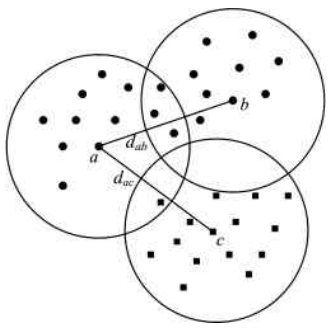

图 1　共享近邻说明图

Fig. 1　Definition graph of shared nearest neighbor

上述定义的相似度计算方式能够有效地区分不同簇间的样本，例如图 1 中，虽然样本 a 和 b 与样本 a 和 c 的距离相同，但 $SNN_{ab}>SNN_{ac}$，因此样本 a 和 b 之间的相似度大于样本 a 和 c 之间的相似度，更符合数据样本的真实簇归属情况。SNN_{ij} 考虑到了样本 i 和 j 的邻域状况，属于同一个簇的两点应该经过一系列高密度点可达，也就是说被一系列点粘合起来。

这种基于共享近邻的相似度计算方式能够扩大有相同簇归属的样本间的相似性，因此在进行加权 K 近邻分配时能够使得与待分配样本实际属于同一簇的样本对当前要分配点的影响，大于实际属于不同簇的样本和噪声点对要分配点的影响，从而使分配结果更准确。

3.2　基于 SNN 相似度的改进算法

根据本文提出的共享近邻相似度，对 FKNN-DPC 中的分配策略进行改进：在分配策略 1 中，以相似 K 近邻作为簇向外扩展的依据，代替原始的基于距离的 K 近邻集合；在分配策略 2 中，以共享近邻相似度代替原始的权重计算公式。其中，样本 i 的相似 K 近邻集合 $simi\_KNN(i)$ 为与 i 相似度最大的 K 个样本组成的集合。相应的归属度计算公式如式(14)所示。

$$p_i^c=\sum_{j\in simi\_KNN_i,y_j=c}\gamma_{ij}*s_{ij} \tag{14}$$

其中，γ_{ij} 为样本 i 与样本 j 之间的相似度关于样本 j 与其相似 K 近邻的相似度之和的标准化，表示已分配样本 j 对于分配样本 i 的贡献度，如式(15)所示。

$$\gamma_{ij}=\frac{w_{ij}}{\sum\limits_{l\in simi\_kNN_j} w_{jl}} \tag{15}$$

改进后的样本分配策略 1、样本分配策略 2 和算法的主要流程如下。

(1)样本分配策略 1

1)从簇中心集合 CC 中选出一个未分配的样本点 i 作为一个新的类簇中心，并将其标记为已分配。

2)将样本 i 的相似 K 近邻分配到样本 i 所在的簇，并初始化扩展队列 T，将样本 i 的相似 K 近邻加入到扩展队列的队尾。

3)在扩展队列中取队首样本 o，若样本 o 的相似 K 近邻集合 $simi\_KNN(i)$ 中的每个样本 p 满足以下条件：①还未被分配；②非噪声点；③$s_{op}\leqslant\mathrm{mean}(\{s_{pq}\mid q\in simi_kNN(p)\})$，则判定 p 为可扩展样本，将 p 分配到 o 所属的簇，并将样本 p 加入到扩展队列 T 的尾部。

4)若扩展队列 T 不为空，则转步骤 5)。

5)若簇中心集合 CC 中还有未分配的样本，则转步骤 1)，否则结束分配策略 1。

(2)样本分配策略 2

假设经过策略 1 的处理，还有 na 个未分配的样本。

1)根据式(13)计算每一个未分配样本归属于每个簇的归属度 p_i^c，这样就构成了一个 $na*|CC|$ 的归属度矩阵，其中 $|CC|$ 代表簇中心的个数。

2)构造两个长度为 na 的向量 VA 和 VP，分别用来存储

每个样本的归属度最大值以及使其归属度达到最大的相应簇的标号。

3)在 VA 中找到归属度值取得最大的点 p,当其最大归属度 $VP[p]>0$ 时,将其分配到最可能在的簇,即 $VP[p]$,并转步骤 4);否则,退出分配策略 2。

4)更新归属度矩阵和 VA,VP 向量,更新 p 的相似 K 近邻中的每一个样本 q 的归属度值为:$p_q^c=p_q^c+\gamma_{qp}*w_{qp}$,并令 $VA[q]=\max\{p_q^c|c=1,\cdots,|CC|\}$,$VP[q]=\arg\max_c\{p_q^c|c=1,\cdots,|CC|\}$。

5)如果没有剩余的未分配点,则结束分配策略 2,否则转步骤 3)。

算法 1 基于共享近邻的加权 K 近邻密度峰值聚类算法

输入:数据集 Data,近邻参数 K,邻域半径 ε

输出:聚类结果 CL

1. 数据预处理:缺失值填补与数据归一化。
2. 计算样本间的欧氏距离,根据式(1)、式(2)和式(5)计算每个样本的 ρ 和 δ。
3. 基于 ρ 和 δ 构造决策图,并选出簇中心点集合 CC。
4. 根据样本分配策略 1 分配非簇中心点。
5. 按照分配策略 2 分配分配策略 1 未分配的样本。
6. 将执行分配策略 1 和分配策略 2 后仍未分配的样本分配到与其最相似的已分配点所在的簇。

3.3 算法的复杂度分析

SNN-FKNN-DPC 算法的空间复杂度主要由 2 部分构成。1)存储相似度矩阵,其空间复杂度为 $O(n^2)$;2)存储每个样本的相似 K 近邻,空间复杂度为 $O(nK)$,其中 $K\leqslant n$。因此,算法的整体空间复杂度为 $O(n^2)$。

SNN-FKNN-DPC 算法的时间复杂度主要由 3 部分构成。1)计算相似度矩阵,包括计算距离矩阵与样本间的共享近邻个数,前者的时间复杂度为 $O(n^2)$。计算样本间的共享近邻数量时,需要先计算每个样本的近邻集合与集合的交集,其中计算一个样本的 ε-近邻集合的时间复杂度为 $O(n)$,若使用 KD-树则可有效检索特定距离内的所有点,从而将时间复杂度降低到 $O(\log n)$,因此计算 n 个样本的 ε-近邻集合的时间复杂度为 $O(n\log n)$;其次需要求集合的交,其时间复杂度为 $O(n_1n_2)$,n_1 和 n_2 分别代表两个样本的 ε-近邻集合中的元素的数量,运用哈希表可快速求解,将时间复杂度降低至 $O(\max\{n_1,n_2\})$。因为 $n_1,n_2\leqslant n$,所以计算相似度矩阵的整体时间复杂度为 $O(n^2)$。2)求每个样本的相似 K 近邻,计算一个样本的相似 K 近邻的时间复杂度为 $O(n)$,因此整体时间复杂度为 $O(n^2)$。3)分配样本的时间复杂度与 FKNN-DPC 中相应操作的时间复杂度相同,为 $O(n^2)$。因此,SNN-FKNN-DPC 的整体时间复杂度为 $O(n^2)$。

4 实验与结果

选取 9 个 UCI 真实数据集[19]对所提算法进行测试和评价,并将实验结果与 K-means,DBSCAN,AP,DPC,FKNN-DPC 进行比较,其中因为 K-means 算法本身具有随机性,所以本文采用 20 次重复实验的平均值作为其最终的实验结果。本文选用的 UCI 真实数据集如表 1 所列。

表 1 本文选用的 UCI 数据集

Table 1 UCI datasets

Dataset	No. records	No. attributes	No. clusters
Iris	150	4	3
Wine	178	13	3
Seeds	210	7	3
Ionosphere	351	34	2
Dermatology	366	34	6
Libras movement	360	91	15
Pima-indians-diabetes	768	8	2
Waveform	5000	19	3
Waveform(noise)	5000	21	3

采用聚类准确率(Acc)、AMI(adjusted mutual information)和 ARI(adjusted rand index)作为度量指标来对实验结果进行评价。这 3 种指标是最常用的评判聚类质量的指标[20],值越大说明聚类质量越高。其中,AMI 基于信息论,ARI 基于样本对计数,分别定义如下。

设 $U=\{U_1,\cdots,U_R\}$ 和 $V=\{V_1,\cdots,V_C\}$ 分别表示数据集 $S=\{s_1,s_2,\cdots,s_N\}$ 的真实划分和聚类结果,U 和 V 之间的 AMI 值的计算公式如式(16)所示。

$$AMI(U,V)=\frac{MI(U,V)-E\{MI(U,V)\}}{\max\{H(U),H(V)\}-E\{MI(U,V)\}} \tag{16}$$

其中,$MI(U,V)$表示 U 和 V 之间的互信息,如式(17)所示。

$$MI(U,V)=\sum_{u\in U}\sum_{v\in V}p(u,v)\frac{P(u,v)}{p_1(u)p_2(v)} \tag{17}$$

设 a 为在 U 和 V 中均属于同一类簇的样本对数目,b 为原划分 U 中属于同一类簇而在 V 中不属于同一类簇的样本对数目,c 为 U 中不属于同一类簇而在 V 中属于同一类簇的样本对数目,d 为 U 和 V 中均不属于同一类簇的样本对数目,则 ARI 的定义如式(18)所示。

$$ARI=\frac{2(ad-bc)}{(a+b)(b+d)+(a+c)(c+d)} \tag{18}$$

首先采用式(19)对数据集进行归一化,使得每个属性的取值都在[0,1]之内。

$$x_{ij}=\frac{x_{ij}-\min(x_j)}{\max(x_j)-\min(x_j)} \tag{19}$$

其中,x_{ij} 表示第 i 个样本的第 j 个属性的取值,$\max(x_j)$ 和 $\min(x_j)$ 分别表示第 j 个特征的最大值和最小值。

上述 6 个聚类算法在 9 个 UCI 数据集上的聚类结果比较如表 2 所列,其中 Par 表示该算法的参数取值,加粗标注表示该值最优,"—"表示没有对应值。在 Waveform 和 Waveform(noise)数据集上,因 DBSCAN 和 AP 算法的复杂度较高,参数难以调整至最优取值,故没有统计其聚类结果。

本文提出的算法在大部分数据集上的聚类结果(即 Acc,AMI,ARI 评价指标)都优于 FKNN-DPC,DPC,AP,DBSCAN,K-means 算法的相应指标值,当选取的簇中心相同时,聚类结果优于 FKNN-DPC,表明本文提出的算法在分配非簇中心样本时更准确;在 Iris 和 Pima-indians-diabetes 数据集上的结果与 FKNN-DPC 一致,原因在于这两个数据集中的簇严重重叠,其特殊的邻域状况使得本文提出的相似度计算方式退化为距离度量,而大多数情况下本文提出的算法更符合数据的真实分布情况。

表 2　UCI 数据集上各算法的 Acc，AMI，ARI 指标对比

Table 2　Comparison of Acc，AMI and ARI of each algorithm

algorithm	Iris				Wine				Ionosphere			
	Acc	AMI	ARI	Par	Acc	AMI	ARI	Par	Acc	AMI	ARI	Par
proposed algorithm	**0.973**	**0.912**	**0.922**	7/0.8	**0.978**	**0.908**	**0.933**	7/0.7	**0.858**	**0.385**	**0.501**	6/1.6
FKNN-DPC	**0.973**	**0.912**	**0.922**	7	0.949	0.831	0.852	7	0.752	0.284	0.355	8
DPC	0.887	0.767	0.720	2	0.882	0.706	0.672	2	0.681	0.238	0.276	0.65
AP	0.907	0.756	0.757	6	0.854	0.686	0.616	6	0.709	0.127	0.173	15
DBSCAN	0.893	0.775	0.732	0.14/9	0.876	0.678	0.660	0.42/10	0.607	0.086	0.036	0.2/7
K-means	0.825	0.692	0.660	2	0.932	0.815	0.830	3	0.712	0.129	0.178	2

algorithm	Seeds				Libras movement				Dermatology			
	Acc	AMI	ARI	Par	Acc	AMI	ARI	Par	Acc	AMI	ARI	Par
proposed algorithm	**0.924**	**0.767**	**0.791**	8/1.2	**0.603**	0.507	**0.407**	7/0.6	**0.867**	**0.875**	**0.871**	7/0.8
FKNN-DPC	**0.924**	0.759	0.790	8	0.436	0.508	0.308	9	0.768	0.847	0.718	7
DPC	0.900	0.717	0.734	2	0.361	0.390	0.214	0.5	0.697	0.588	0.490	2
AP	0.895	0.685	0.715	10	0.450	0.497	0.277	2.5	0.814	0.771	0.717	5
DBSCAN	0.881	0.644	0.686	0.17/8	0.350	0.408	0.154	0.96/5	0.787	0.709	0.727	0.7/3
K-means	0.890	0.671	0.705	3	0.443	**0.519**	0.304	15	0.691	0.786	0.654	6

algorithm	Waveform				Waveform(noise)				Pima-indians-diabetes			
	Acc	AMI	ARI	Par	Acc	AMI	ARI	Par	Acc	AMI	ARI	Par
proposed algorithm	**0.727**	**0.382**	**0.372**	5/0.9	**0.659**	**0.296**	**0.257**	5/1.4	0.648	0.001	0.013	7/0.8
FKNN-DPC	0.703	0.324	0.350	5	0.648	0.247	0.253	5	0.648	0.001	0.013	6
DPC	0.586	0.318	0.268	0.5	0.535	0.184	0.164	0.3	0.650	0.034	0.078	4
AP	—	—	—	—	—	—	—	—	0.624	0.045	0.089	35
DBSCAN	—	—	—	—	—	—	—	—	0.540	0.017	0.035	0.15/6
K-means	0.501	0.364	0.254	3	0.512	0.364	0.252	3	**0.668**	**0.050**	**0.102**	2

经过实验统计，ε 的取值范围通常为 0.5～1.5。一种选择 ε 的方法为：根据式(6)－式(8)识别出噪声点后，计算噪声点与其 K 近邻的平均距离，并将此距离作为参数 ε 的取值。

结束语　本文提出了一种基于共享近邻的相似度度量方式，并基于此相似度度量方式改进了 FKNN-DPC 算法中的样本分配策略，使得样本分配更符合真实的簇归属情况，并对参数的设定给出了指导性建议，使得该算法更易使用。本文算法在 9 个 UCI 真实数据集上的实验结果优于 K-means，DBSCAN，AP，DPC，FKNN-DPC 的结果，表明改进后的分配策略能够有效地提高聚类的质量。如何高效地计算相似度，并将此算法应用于大数据环境下，是需要进一步研究的问题。

参 考 文 献

[1] JAIN A K，DUBES R C. Algorithms for clustering data[M]. Upper Saddle Riever：Prentice Hall，1988.

[2] BERKHIN P. A Survey of Clustering Data Mining Techniques [J]. Grouping Multidimensional Data，2006，43(1)：25-71.

[3] XU R，ND W D. Survey of clustering algorithms[J]. IEEE Transactions on Neural Networks，2005，16(3)：645-678.

[4] GELBARD R，GOLDMAN O，SPIEGLER I. Investigating diversity of clustering methods：An empirical comparison[J]. Data & Knowledge Engineering，2007，63(1)：155-166.

[5] SUN J G，LIU J，ZHAO L Y. Clustering Algorithms Research [J]. Journal of Software，2008，19(1)：48-61.(in Chinese)
孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. 软件学报，2008，19(1)：48-61.

[6] MACQUEEN J. Some Methods for Classification and Analysis of MultiVariate Observations[C]// Proc. of Berkeley Symposium on Mathematical Statistics and Probability. 1967：281-297.

[7] GUHA S，RASTOGI R，SHIM K，et al. CURE：An Efficient Clustering Algorithm for Large Databases[J]. Information Systems，1998，26(1)：35-58.

[8] KARYPIS G，HAN E H，KUMAR V. CHAMELEON A hierarchical clustering algorithm using dynamic modeling[J]. Computer，1999，32(8)：68-75.

[9] ESTER M，KRIEGEL H P，XU X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise [C]// International Conference on Knowledge Discovery and Data Mining. 1996：226-231.

[10] FAN J. OPE-HCA：an optimal probabilistic estimation approach for hierarchical clustering algorithm[J/OL]. Neural Computing & Applications，2015：1-11. https：//link. springer. com/article/10.1007%2Fs00521-015-1998-5.

[11] AGRAWAL R，GEHRKE J，GUNOPULOS D，et al. Automatic subspace clustering of high dimensional data for data mining applications[M]// ACM SIGMOD Record. ACM，1998：94-105.

[12] FREY B J，DUECK D. Clustering by passing messages between data points[J]. Science，2007，315(5814)：972-976.

[13] RODRIGUEZ A，LAIO A. Clustering by fast search and find of density peaks[J]. Science，2014，344(6191)：1492-1496.

[14] DU M，DING S，JIA H. Study on density peaks clustering based on k-nearest neighbors and principal component analysis[J]. Knowledge-Based Systems，2016，99：135-145.

(下转第 146 页)

了新的要求，社群化制造将是适应这一要求的新的发展模式。针对社群化制造系统规模庞大、内部关系结构复杂、建模困难等问题，本文提出了基于 SLE 范式的社群化制造系统模型，包括个体模型、交互模型与社会模型 3 个部分；进一步引入了“计算实验”的思想，模拟社群化制造系统在不同演化机制下的演化过程，并通过对演化结果的分析验证了基于 SLE 范式的社群化制造系统模型具有巨大的优越性。在以后研究中需要对社群化制造系统计算模型进行进一步完善，使其能够增添更丰富的数据信息，模拟更多样的外部环境，以更贴近现实系统。

参 考 文 献

[1] DAHLGREN E, GÖÇMEN C, LACKNER K, et al. Small Modular Infrastructure[J]. The Engineering Economist, 2013, 58(4): 231-264.

[2] ANDREADIS G. A collaborative framework for social media aware manufacturing[J]. Manufacturing Letters, 2015, 3(1): 14-17.

[3] IZVERCIAN M, POTRA S A. Prosumer-oriented Relationship Management Capability Development for Business Performance [J]. Procedia Technology, 2014, 16: 606-612.

[4] YANG C C, SUN J, ZHAO Z Y. Personalized recommendation based on collaborative filtering in social network[C]// IEEE International Conference on Progress in Informatics and Computing. IEEE, 2010: 670-673.

[5] XUE X, HAN H F, WANG S F, et al. Computational Experiment-based Evaluation on Context-aware O2O Service Recommendation[J/OL]. IEEE Transactions on Services Computing, 2016. http://ieee xplore. ieee. org/document/7779158.

[6] JIANG P Y, DING K, LENG J W. Towards a cyber-physical-social-connected and service-oriented manufacturing paradigm: Social Manufacturing[J]. Manufacturing Letters, 2016, 7(1): 15-21.

[7] JIANG P Y, DING K, LENG J W, et al. Service-driven social manufacturing paradigm[J]. Commputer Integrated Manufacturing Systems, 2015, 21(6): 1637-1649. (in Chinese)
江平宇，丁凯，冷杰武，等. 服务驱动的社群化制造模式研究[J]. 计算机集成制造系统，2015，21(6)：1637-1649.

[8] LENG J W, JIANG P Y, ZHANG F Q, et al. Framework and Key Enabling Technologies for Social Manufacturing[J]. Applied Mechanics & Materials, 2013, 312(2): 498-501.

[9] DING K, JIANG P Y, ZHANG X. A Framework for Implementing Social Manufacturing System Based on Customized Community Space Configuration and Organization[J]. Advanced Materials Research, 2013, 712-715(6): 3191-3194.

[10] XUE X, WANG S F, GUI B, et al. A Computational Experiment-based Evaluation Method for Context-aware Services in Complicated Environment[J]. Information Sciences, 2016, 373(9): 269-286.

[11] FENG X, MA J H. Building smart communities with cyber-physical systems[C]// Proceedings of 1st International Symposium on From Digital Footprints to Social and Community Intelligence. ACM, 2011: 1-6.

[12] LANCICHINETTI A, FORTUNATO S, KERTÉSZ J. Detecting the overlapping and hierarchical community structure of complex networks[J]. New Journal of Physics, 2008, 11(3): 19-44.

[13] BIAMINO G. A Semantic Model for Socially Aware Objects[J]. Advances in Internet of Things, 2012, 2(3): 47-55.

[14] DING K, JIANG P Y, LENG J W, et al. Modeling and analyzing of an enterprise relationship network in the context of social manufacturing[J]. Proceedings of the Institution of Mechanical Engineers Part B Journal of Engineering Manufacture, 2015, 230(4): 1207-1217.

[15] XIONG G, CHEN Y R, SHANG X Q, et al. AHP fuzzy comprehensive method of supplier evaluation in social manufacturing mode[C]// Intelligent Control and Automation. IEEE, 2015: 3594-3599.

[16] PANG B H. Multi-criteria Supplier Evaluation Using Fuzzy AHP[C]// International Conference on Mechatronics and Automation. IEEE, 2007: 2357-2362.

[17] ANDZULIS J, RAPP A, TRAINOR K J, et al. Social media technology usage and customer relationship performance: A capabilities-based examination of social CRM[J]. Journal of Business Research, 2014, 67(6): 1201-1208.

(上接第 129 页)

[15] MEHMOOD R, ZHANG G, BIE R, et al. Clustering by fast search and find of density peaks via heat diffusion[J]. Neurocomputing, 2016, 208(C): 210-217.

[16] XIE J Y, GAO H C, XIE W X. K-nearest neighbors optimized clustering algorithm by fast search and nding the density peaks of a dataset[J]. Scientia Sinica Informationis, 2016, 46(2): 258-280. (in Chinese)
谢娟英，高红超，谢维信. K 近邻优化的密度峰值快速搜索聚类算法[J]. 中国科学：信息科学，2016，46(2)：258-280.

[17] XIE J, GAO H, XIE W, 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.

[18] ZHANG W, LI J. Extended fast search clustering algorithm: widely density clusters, no density peaks[J]. arXiv preprint arXiv: 1505. 05610, 2015.

[19] LICHMAN M. UCI Machine Learning Repository[OL]. http://archive. ics. uci. edu/ml.

[20] VINH N X, EPPS J, BAILEY J. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance[C]// International Conference on Machine Learning (ICML 2009). Montreal, Quebec, Canada, 2009: 2837-2854.