



一种改进的基于局部密度的聚类算法

关晓惠¹, 钱亚冠², 孙欣欣¹

(1. 浙江水利水电学院, 浙江 杭州 310018; 2. 浙江科技学院理学院, 浙江 杭州 310023)

摘要: 聚类分析一直是机器学习和数据挖掘领域一个比较活跃而且极具挑战性的研究方向。Alex 提出的基于局部密度的聚类算法是一种快速、有效的聚类方法, 但该方法通过手工选取确定聚类个数和聚类中心。为此, 对原算法进行改进, 在初步选取候选聚类中心的基础上, 使用基于密度连通的算法优化选取聚类中心, 然后使用大密度最近邻方法确定样本类别。实验证明, 该方法能有效解决聚类个数和聚类中心无法确定的问题, 同时在聚类评价指标上显示出较好的聚类效果和性能。

关键词: 局部密度; 类簇中心; 评价指标

中图分类号: TN929.5

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2016008

An improved clustering algorithm based on local density

GUAN Xiaohui¹, QIAN Yaguan², SUN Xinxin²

1. College of Information Engineering and Art Design, Zhejiang University of
Water Resources and Electric Power, Hangzhou 310018, China

2. College of Science, Zhejiang University of Science and Technology, Hangzhou 310023, China

Abstract: Clustering analysis is an important and challenging research field in machine learning and data mining. A fast and effective clustering algorithm based on the idea of local density was proposed by Alex. But the number of clusters and cluster centers in the algorithm were determined by hand. Therefore, the candidates of cluster centers based on local density were firstly selected and then density connectivity method was used to optimize the candidates. The classes of samples are the same as the nearest center with bigger local density. Experiments show that the proposed method has a better cluster efficiency and can handle the problems of uncertain cluster number and cluster centers.

Key words: local density, cluster center, evaluation criterion

1 引言

聚类是指在没有任何先验知识的情况下, 根据数据特征的相似性将同类数据聚集在一起的过程, 属于无监督分类的范畴。聚类的目标是使得同一类簇内对象的相似性尽可能大, 不同类簇之间对象的相似性尽可能小。聚类作为

一种重要的数据分析和挖掘手段, 已被广泛应用于语音识别、字符识别、图像处理、信息安全、金融等领域。

迄今为止, 国内外研究人员相继提出很多聚类算法, 主要分为基于层次的聚类、基于划分的聚类、基于密度的聚类、基于网格的聚类、基于模型的聚类等^[1]。基于层次的聚类是指对样本集合进行合并或者分裂, 直到满足某一个终

止条件,代表算法有 BIRCH 算法、CURE 算法^[23]。优点是能得到不同粒度的聚类结构,缺点是很难确定合并和分裂的准则。基于划分的聚类是指首先将所有数据粗略地划分为 K 个类,然后通过迭代算法使某个准则达到最优来对划分进行修正。代表算法有 k -means 算法、 k 中心点方法及其改进^[4-7]。优点是算法简单、速度快,缺点是 K 值需要事先指定,而且只能发现圆形类簇。基于密度的聚类算法是指根据数据对象的分布密度,将密度足够大的数据对象聚类在一起,样本空间被低密度区间划分开,代表算法有 DBSCAN 算法、OPTICS 算法、DENCLUE 算法^[7-9]。优点是发现任意形状类簇,缺点是参数的设置对聚类结果影响较大。基于网格的聚类是指将数据空间量化为有限单元,构成一个可以聚类的网格结构,代表算法有 STING 算法、CLIQUE 算法^[10,11]。优点是运算速度快,缺点是存在量化尺度问题。基于模型的聚类是指寻找给定数据与某种数据模型的最佳拟合,代表方法有 COBWEB 算法、AutoClass 算法、SOM 算法^[12-14]。

近年来随着人工智能、机器学习、模式识别、数据挖掘等领域的不断发展,又提出了许多新的聚类算法。为了解决样本点不仅仅只属于某一个类的问题,提出了模糊聚类^[15-17],用模糊理论的方法对数据进行软划分。谱聚类是一种基于图论的聚类方法^[18],通过计算数据之间相似矩阵的特征值和特征向量进行聚类。子空间聚类是针对高维数据空间出现的一种有效聚类方法^[19],通过特征选择在不同的子空间上进行聚类。然而,在很多聚类方法中都需要提供聚类个数作为参数,目前还没有一个很好的办法可以保证获得准确的聚类数目,这一直是聚类分析中的一个难点^[20]。Frey 提出一种利用亲密度传播进行聚类的方法^[21]。该方法无需事先指定聚类数目,能够快速、有效地处理大规模数据集,但对于比较松散的聚类结构就会得到较多的聚类数目。

2014 年 Alex Rodriguez 和 Alessandro Laio 在 Science 上提出一种简洁的聚类算法^[22]。与以往的聚类算法相比,该方法能够处理任意形状类簇,而且对数据变换有很好的顽健性。但该方法中聚类个数和聚类中心无法自动确定,需要手工选取,这无疑限制了算法的应用范围和领域。本文提出的基于局部密度的聚类算法,是对该算法的一种改进。在初步选取候选聚类中心的基础上,增加一个优化选取聚类中心的过程,使用基于密度连通的算法合并或剔除不正确的聚类中心,使用大密度最近邻方法确定样本类

别。实验证明,该方法具有较好的聚类效果和性能,有效解决了聚类个数不确定的问题。

2 聚类过程

2.1 算法思想

本文算法的核心思想是基于局部密度的概念,它表示与该点的距离在一定范围的点的个数,也就是说一个点附近点的个数越多,其局部密度越大。该算法认为聚类中心是由一些局部密度比较低的点围绕,并且这些点距离其他高局部密度的点的距离都比较大。为此定义两个量。

(1) 局部密度 ρ_i

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \chi(x) = \begin{cases} 1, & x < 0 \\ 0, & x \geq 0 \end{cases} \quad (1)$$

其中, $d_c > 0$ 为截断距离,需要用户确定。推荐做法是选择 d_c ,使得每个点的平均邻居数为总点数的 1%~2% (假设为 t)。为了将聚类算法扩展到异形类簇,本文使用高斯核函数来定义局部密度,既避免了不同的点具有相同局部密度的问题,又能识别异形类簇。

$$\rho_i = \sum_j e^{-(d_{ij}/d_c)^2} \quad (2)$$

(2) 到较高局部密度点的最近距离 δ_i

$$\delta_i = \min_{j: \rho_j > \rho_i} (d_{ij}) \quad (3)$$

表示所有局部密度大于 ρ_i 的点中,与 x_i 距离最近的点 x_j 与 x_i 之间的距离。对于密度最大的点, $\delta_i = \max_j d_{ij}$,表示与 x_i 距离最大的数据点与 x_i 之间的距离。

2.2 确定类簇中心

类簇中心是指局部密度比较大,且距离其他较大局部密度的点的距离比较远的点。首先计算所有点的 ρ_i 和 δ_i ,以 ρ 为横坐标,以 δ 为纵坐标形成决策图,选择 ρ_i 和 δ_i 都比较大的点作为类簇的中心。为了定量确定类簇的中心点,定义 $\gamma_i = \rho \delta_i$,然后对 $\{\gamma_i | i=1, \dots, N\}$ 进行降序排序,选择 γ_i 大于某个阈值 λ 的点为中心点。此时可能会存在两种特殊情况:第一种情况是一些 ρ 很大但 δ 值很小的点会被选为中心点,这样可能会造成同一个类簇中有两个中心点存在,将本来属于同一个类簇的数据点分成两个不同的类簇;第二种情况是 ρ 很小,但 δ 很大,这样会把部分异常点视为聚类的中心,本文的做法是对 ρ 和 δ 都设置各自的阈值,将大于阈值的点视为候选中心点。然后使用基于密度的连通性算法将候选中心点合并或剔除,具体算法如下。

**算法 1** DCC(determining-clustering-center)

输入: $X=\{x_1, x_2, \dots, x_N\}$ 是需要聚类的数据点; N 是数据点个数; $\{\rho_1, \rho_2, \dots, \rho_N\}$ 为每个数据的局部密度; $\{\delta_1, \delta_2, \dots, \delta_N\}$ 为每个样本点到高局部密度的最小距离。

输出: 类簇中心点 $\{x_{m_1}, x_{m_2}, \dots, x_{m_k}\}$ 。

for $i \leftarrow 1$ to N

 计算 $\gamma_i = \rho \delta_i$;

end for

对 $\{\gamma_1, \gamma_2, \dots, \gamma_N\}$ 从大到小进行排序, 得到降序下标序 $\{q_1, q_2, \dots, q_N\}$;

选取 $\gamma_i > \varepsilon$ 并且 $\rho_i > \sigma$ 的点为候选类簇中心点 $\{x_{q_1}, x_{q_2}, \dots, x_{q_w}\}$;

任意候选类簇中心 x_{q_i}, x_{q_j}

 if (在 d_c 邻域内如果存在一条直接密度可达数据点链 $\{p_1, p_2, \dots, p_i\}$, 满足 $x_{q_i} = p_1, x_{q_j} = p_i$)

x_{q_i}, x_{q_j} 属于同一个类簇, 并选择 ρ 值比较大的点为合并后的类簇中心;

 end if

return

2.3 聚类

类簇中心确定以后, 需要确定每个点划分给某个类簇的可靠性。本文使用大密度最近邻方法将每个点归类到局部密度比自己大的最近邻的簇。聚类算法如下。

算法 2 LDC(local-density-clustering)

输入: $X=\{x_1, x_2, \dots, x_N\}$ 是需要聚类的数据点; N 是数据点个数。

输出: 每个数据点的类别 $C=\{c_1, c_2, \dots, c_N\}$ 。

计算数据点间的相互距离 d_{ij} ;

将 d_{ij} 按从小到大排序 $\{d_1, d_2, \dots, d_N\}$; 其中 $M = \frac{1}{2}N(N-1)$;

$d_c = d_{[M]}$; 其中 $t=2\%$;

对每个数据点计算 ρ_i 和 δ_i ;

调用函数 MC , 确定类簇中心点 $\{x_{m_1}, x_{m_2}, \dots, x_{m_k}\}$;

for $i \leftarrow 1$ to N

 if (x_i 不属于类簇中心) then

$d = d_M$ (将 d 设置为最大的数据点间距);

 for 每个类簇中心 $\{x_{m_k}\}$

 if ($\rho_i \leq \rho_{m_k} \&\& d > d_{m_k}$)

$d = d_{m_k}$;

 end if

 end for

$c_i = c_{m_k}$;

end if

end for

return C ;

3 评价指标

评价一个聚类算法的好坏一般基于这样的原则: 簇中的成员尽可能地互相靠近, 簇与簇之间的距离尽可能远。假设 $P=\{P_1, P_2, \dots, P_s\}$ 为人工标注的分类结果, $C=\{C_1, C_2, \dots, C_m\}$ 为聚类算法的划分。本文采用以下评价指标。

(1) *purity*: 正确聚类的样本数占总样本数的比例

$$purity(P, C) = \frac{1}{N} \sum_k \max_j |P_k \cap C_j| \quad (4)$$

(2) *R* 指数: 表示 C 和 P 之间的相似程度

假设 a 表示两个点在 C 和 P 中均属于同一个簇的个数; b 表示两个点在 C 中属于相同的簇, 在 P 中属于不同簇的个数; c 表示两个点在 C 中属于不同的簇, 在 P 中属于相同簇的个数; d 表示两个点在 C, P 中均属于不同簇的个数。 R 值越大说明 C 和 P 的吻合度越高, 说明 C 的聚类效果越好。

$$R \text{ 指数: } R = \frac{a+b}{a+b+c+d} \quad (5)$$

(3) *F-measure*: 由准确率和召回率两个指标组合而成。

$$\text{准确率: } P(P_j, C_i) = \frac{|P_j \cap C_i|}{|C_i|} \quad (6)$$

$$\text{召回率: } R(P_j, C_i) = \frac{|P_j \cap C_i|}{|P_j|} \quad (7)$$

$$F \text{ 值: } F(P_j, C_i) = \frac{2 \cdot P(P_j, C_i) \cdot R(P_j, C_i)}{P(P_j, C_i) + R(P_j, C_i)} \quad (8)$$

对每个类簇 P_j , 计算 F 值 $F(P_j) = \max_{1 \leq i \leq m} F(P_j, C_i)$, 对所有类簇的 F 值做加权平均, 得到整个聚类结果的 F 值:

$$F = \sum_{j=1}^s \frac{|P_j| \cdot F(P_j)}{n}$$

4 实验与结果分析**4.1 实验数据**

UCI 数据库是一个专门用于测试分类、聚类算法的国际通用标准测试数据库, 包含 Wine、Iris、Glass 等数据集。其中 Iris 数据集包含 3 类, 每一类代表一种类型的鸢尾

花,每类有50个数据,共150个样本,在3个类簇中分布均匀,其中一类与另外两类线性可分,另外两类有部分重叠。Wine数据集包含178个样本,13个数值型属性,共分成3类,每类中样本数量不同。Glass数据集共有69个样本,包含3类,每类占总数据量的1/3。另外,Leuk72_3k也是比较常用的聚类测试数据集。

4.2 类簇中心选择

算法首先根据局部密度和到高密度样本的距离来确定类簇中心,然后计算其他非中心样本与类簇中心的距离,从而决定样本归属。因此,算法中类簇中心点的选择不但决定着聚类的个数,还影响其他样本的类别归属。图1(a)为Iris数据样本经过多维尺度变换后样本的分布情况,图1(b)为 $\{\gamma_i | i=1, \dots, N\}$ 从大到小排序后的结果。如果选择 γ_i 最大的2个样本作为类簇中心,则整个数据被分成2个类簇,如果选择 γ_i 值最大的前5个样本作为类簇中心,则样本被分成5个类簇。为了更合理地确定类簇中心,首先给 γ_i 设置一个相对较小的阈值(本实验的阈值为6),使较多的样本点成为候选类簇中心,然后使用算法1对候选类簇中心进行合并,得到最优的类簇中心,图1(c)中菱形的点为候选类簇中心。图1(d)中菱形的点为合并后的类簇中心,样本的不同形状标示根据最优类簇中心聚类后的结果。

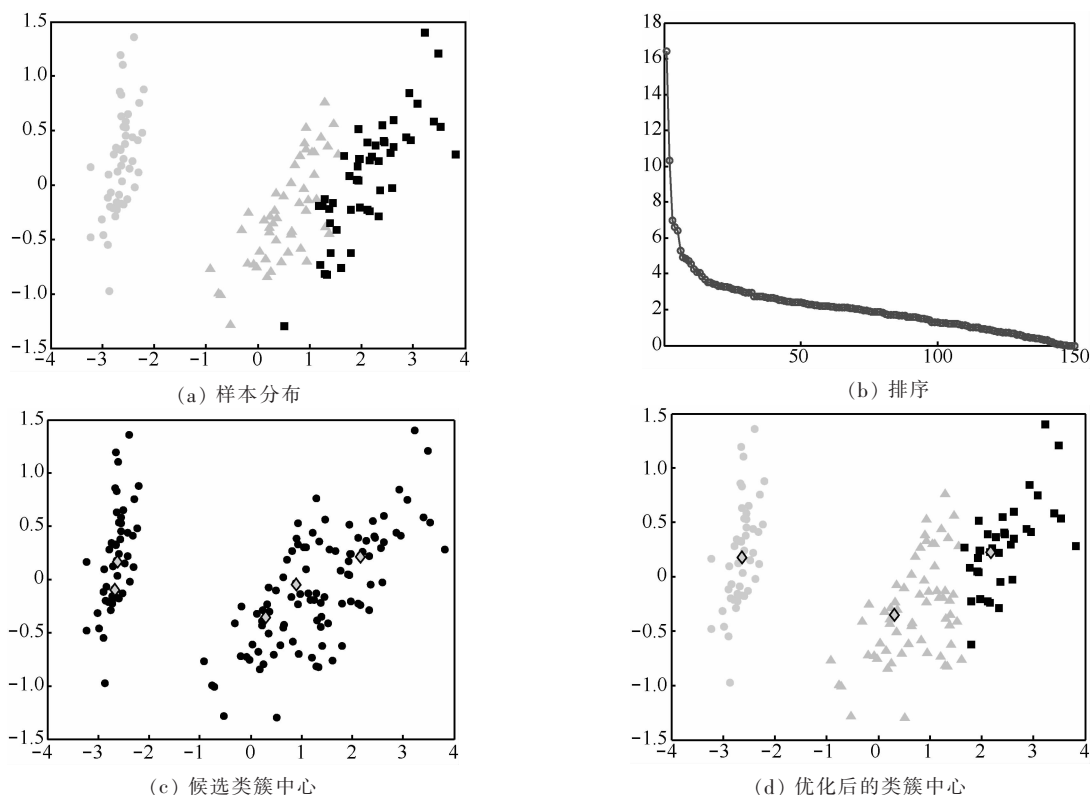


图1 聚类中心选择

4.3 d_c 对算法结果的影响

d_c 的选择决定局部密度的大小,如果取得太大, ρ_i 的区分度不大,类簇中心不准确,如果取得太小,类簇中心的个数过多,会导致同一类簇的数据被划分为不同的类簇。为了证明 d_c 的大小对实验结果的影响,本文针对不同的数据集,分别采用不同大小的 d_c 做实验,得出的实验结果如图2所示(t 为 d_c 的值使得每个点的平均邻居数占有所有点的比例)。

从图2中可以看出,不同数据集下, d_c 对聚类结果的影响是不一样的。Iris和Wine数据集都有最优的 d_c 。对于Iris数据集,当 $t > 2\%$ 时,只能聚出2类,当 $t < 1\%$ 时,虽然能聚出3类,但聚类的准确率在降低。Leuk72_3k和Glass数据集的聚类结果基本不受 d_c 的影响。通过分析发现,Leuk72_3k数据集的类内样本点的距离远小于类间的距离。因此在不同的应用背景下,应该根据具体的问题选择合适的 d_c 参数。

4.4 聚类结果对比

为了验证算法的有效性,将本文中算法与经典的K-means算法和DBSCAN算法进行实验对比,并用purity、R指数、F-measure来衡量算法的优劣性。表1为几种聚类算法在不同数据集上的实验结果比较。

从表1可以看出,本文算法相对于K-means、DBSCAN算法在各指标上均有较大的提升,说明该算法有较好的聚

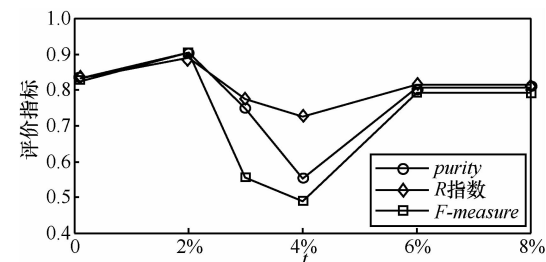
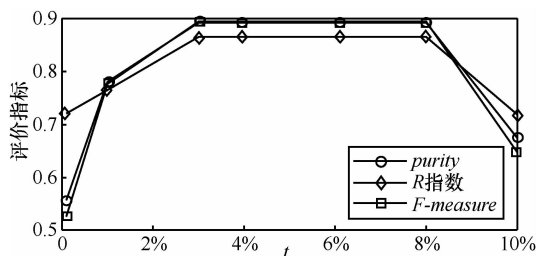
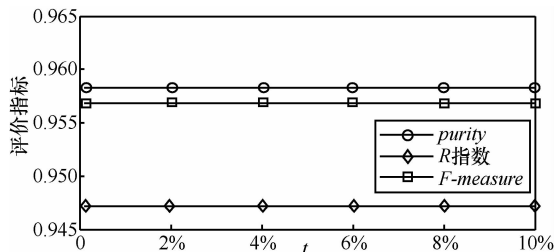
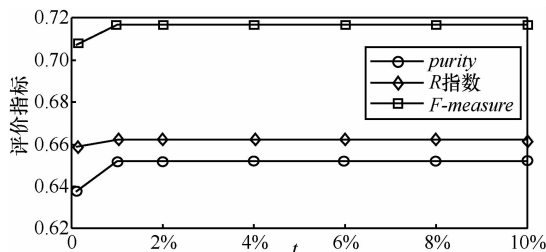
(a) Iris 数据集下不同 d_c 值的聚类结果(b) Wine 数据集下不同 d_c 值的聚类结果(c) Leuk72_3k 数据集下不同 d_c 值的聚类结果(d) Glass 数据集下不同 d_c 值的聚类结果图2 不同数据集下 d_c 对聚类结果的影响

表1 不同聚类算法实验结果比较

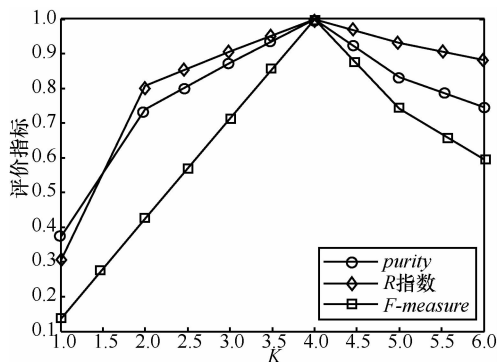
数据集	评价指标	本文算法	K-means	DBSCAN
Iris 数据集	<i>purity</i>	0.906 7	0.881 7	0.666 7
	<i>R</i> 指数	0.892 3	0.839 4	0.776 3
	<i>F-measure</i>	0.904 8	0.885 3	0.655 6
Wine 数据集	<i>purity</i>	0.893 3	0.724 9	0.696 6
	<i>R</i> 指数	0.865 4	0.691 9	0.721 2
	<i>F-measure</i>	0.895 8	0.704 3	0.706 0
leuk72_3k 数据集	<i>purity</i>	0.958 3	0.958 3	0.916 7
	<i>R</i> 指数	0.947 2	0.947 2	0.896 7
	<i>F-measure</i>	0.956 8	0.956 8	0.915 4
Glass 数据集	<i>purity</i>	0.752 2	0.710 1	0.579 7
	<i>R</i> 指数	0.762 4	0.683 3	0.638 5
	<i>F-measure</i>	0.717 1	0.726 8	0.516 2
4k2_far 数据集	<i>purity</i>	1	0.807 5	0.987 5
	<i>R</i> 指数	1	0.930 5	0.985 9
	<i>F-measure</i>	1	0.903	0.986 8

类效果 and 性能。

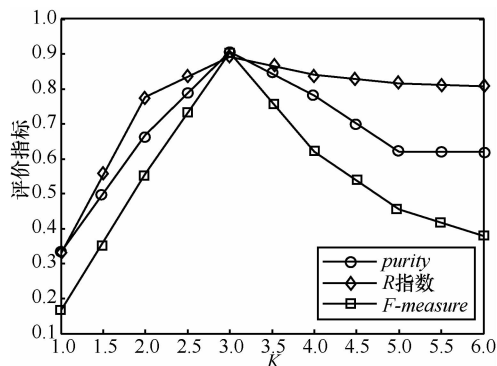
Alex 提出的算法中,聚类个数以及类簇中心都通过人工方式选定,为了确定最优的聚簇类数,本文采用最优评价指标方法来确定聚类个数。在给定的数据集上,通过选择不同的类簇中心个数,对数据集进行不同的划分,并计算不同划分的评价指标,如图3所示。选择评价指标最好的聚类个数为最佳聚类个数。从图3中可以看出,4k2_far 数据集的最优类簇个数为4, Iris 数据集的最优类簇个数为3。

5 结束语

针对基于局部密度的聚类算法无法自动选择类簇个数和类簇中心的问题,本文在该算法的基础上增加了一个优化



(a) 4k2_far 数据集不同划分的聚类指标



(b) Iris 数据集不同划分的聚类指标

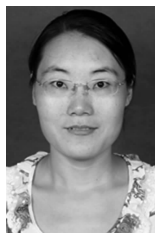
图3 不同数据集下类簇个数与各聚类指标的关系

选取聚类中心的过程,使用基于密度连通的算法合并或删除不正确的聚类中心。与其他聚类算法相比,该方法具有较好的聚类效果和性能,并有效地解决了聚类个数不确定的问题。本文还验证了不同的截断距离对聚类结果的影响,实验证明在实际应用中应该根据具体的聚类问题选择合适的参数。

参考文献:

- [1] 周涛, 陆惠玲. 数据挖掘中聚类算法研究进展 [J]. 计算机工程与应用, 2012, 48 (12): 100-109.
ZHOU T, LU H L. Clustering algorithm research advances on data mining [J]. Computer Engineering and Applications, 2012, 48 (12): 100-109.
- [2] ZHANG T, RAMAKRISHNAN R, LIVNY M. BIRCH: an efficient data clustering method for very large databases [C]// Proceedings of 1996 ACM-SIGMOD International Conference Management of Data, June 4-6, 1996, Montreal, Quebec, Canada. New York: ACM Press, 1996: 103-114.
- [3] GUHA S, RASTOGI R, SHIM K. CURE: an efficient clustering algorithm for large database [J]. Information Systems, 2001, 26(1): 35-58.
- [4] MACQUEEN J. Some methods for classification and analysis of multivariate observations [C]// Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, June 21-18, 1965, Berkeley, California, USA. California: University of California Press, 1967: 281-297.
- [5] KAUFMAN L, ROUSSEEUW P J. Finding Groups in Data: An Introduction to Cluster Analysis [M]. New York: John Wiley & Sons, 1990.
- [6] HUANG Z. Extensions to the k-means algorithm for clustering large data sets with categorical values [J]. Data Mining and Knowledge, Discovery II, 1998(2): 283-304.
- [7] 倪巍伟, 陈耿, 吴英杰. 一种基于局部密度的分布式聚类挖掘算法 [J]. 软件学报, 2008, 19(9): 2339-2348.
NI W W, CHEN G, WU Y J. Local density based distributed clustering algorithm [J]. Journal of Software, 2008, 19 (9): 2339-2348.
- [8] ANKERST M, BREUNIG M, KRIEGER H P, et al. OPTICS: ordering points to identify the clustering structure [C]// Proceedings of 1999 ACM-SIGMOD International Conference Management of Data (SIGMOD'99), June 1-3, 1999, Philadelphia, Pennsylvania, USA. New York: ACM Press, 1999: 49-60.
- [9] HINNEBURG A, KEIM D A. An efficient approach to clustering in large multimedia databases with noise [C]// Proceedings of 1998 International Conference Knowledge Discovery and Data Mining, August 27-31, 1998, New York, USA. New York: ACM Press, 1998: 58-65.
- [10] WANG W, YANG J, MUNTZ R. STING: a statistical information grid approach to spatial data mining [C]// Proceedings of 1997 International Conference Very Large Data Bases, August 2-29, 1997, Athens, Greece. New York: ACM Press, 1997: 186-195.
- [11] AGRAWAL R, GEHRKE J, GUNOPULOS D, et al. Automatic subspace clustering of high dimensional data for data mining applications [C]// Proceedings of 1998 ACM-SIGMOD International Conference Management of Data, June 2-4, 1998, Seattle, Washington, USA. New York: ACM Press, 1998: 94-105.
- [12] FISHER D. Improving inference through conceptual clustering [C]// Proceedings of 1987 AAAI Conference, July 13-17, 1987, Seattle, Washington, USA. [S.l.]: AAAI Press, 1987: 461-465.
- [13] FAYYAD V M, PIATETSKY S G, SMYTH P, et al. Bayesian Classification (AutoClass): Theory and Result. Advances in Knowledge Discovery and Data Mining [M]. Bridge City: The MIT Press, 1996: 153-180.
- [14] TEUVO K. The self-organizing map [J]. Neurocomputing, 1998, 21(13): 1-6.
- [15] CAI W L, CHEN S C, ZHANG D Q. Fast and robust fuzzy c-means clustering algorithms incorporating local information for image segmentation [J]. Pattern Recognition, 2007, 40(3): 825-833.
- [16] BASU B, SRINIVAS V V. Regional flood frequency analysis using kernel-based fuzzy clustering approach [J]. Water Resources Research, 2014, 50(4): 3295-3316.
- [17] LI X, WONG H S, WU S. A fuzzy minimax clustering model and its applications [J]. Information Sciences, 2012, 186: 114-125.
- [18] 周林, 平西建, 徐森, 等. 基于谱聚类的聚类集成算法 [J]. 自动化学报, 2012, 38 (8): 1335-1342.
ZHOU L, PING X J, XU S, et al. Cluster ensemble based on spectral clustering [J]. Acta Automatica Sinica, 2012, 38(8): 1335-1342.
- [19] 陈黎飞, 郭躬德, 姜青山. 自适应的软子空间聚类算法 [J]. 软件学报, 2010(10): 2513-2523.
CHEN L F, HUO G D, JIANG Q S. Adaptive algorithm for soft subspace clustering [J]. Journal of Software, 2010(10): 2513-2523.
- [20] SUN H, WANG S, JIANG Q. FCM-based model selection algorithms for determining the number of cluster [J]. Pattern Recognition, 2004, 37(10): 2027-2037.
- [21] FREY B J, DUECK D. Clustering by passing messages between data points [J]. Science, 2007, 315(5814): 972-976.
- [22] ALEX R, ALESSANDRO L. Clustering by fast search and find of density peaks [J]. Science, 2014, 344(6191): 1492-1496.

[作者简介]



关晓惠 (1977-), 女, 浙江水利水电学院副教授, 主要研究方向为机器学习与数据挖掘。

钱亚冠 (1976-), 男, 博士, 浙江科技学院理学院副教授, 主要研究方向为互联网流量分类、下一代互联网、机器学习与数据挖掘。

孙欣欣 (1973-), 女, 浙江水利水电学院副教授, 主要研究方向为计算机网络。