基于样本空间分布密度的初始聚类中心优化 K-均值算法*

谢娟英[1,2], 郭文娟[1], 谢维信[2,3], 高新波[2]

(1. 陕西师范大学 计算机科学学院, 西安 710062; 2. 西安电子科技大学 电子工程学院, 西安 710071; 3. 深圳大学 信息工程学院 ATR 国家重点实验室, 深圳 518060)

摘 要: 针对传统 K-均值聚类算法对初始聚类中心敏感、现有初始聚类中心优化算法缺乏客观性,提出一种基于样本空间分布密度的初始聚类中心优化 K-均值算法。该算法利用数据集样本的空间分布信息定义数据对象的密度,并根据整个数据集的空间信息定义了数据对象的邻域;在此基础上选择位于数据集样本密集区且相距较远的数据对象作为初始聚类中心,实现 K-均值聚类。UCI 机器学习数据库数据集以及随机生成的带有噪声点的人工模拟数据集的实验测试证明,本算法不仅具有很好的聚类效果,而且运行时间短,对噪声数据有很强的抗干扰性能。基于样本空间分布密度的初始聚类中心优化 K-均值算法优于传统 K-均值聚类算法和已有的相关 K-均值初始中心优化算法。

关键词: 聚类; K-均值聚类; 初始中心; 邻域; 样本分布密度

中图分类号: TP301.6 **文献标志码**: A **文章编号**: 1001-3695(2012)03-0888-05

doi:10.3969/j.issn.1001-3695.2012.03.024

K-means clustering algorithm based on optimal initial centers related to pattern distribution of samples in space

XIE Juan-ying[1,2], GUO Wen-juan[1], XIE Wei-xin[2,3], GAO Xin-bo[2]

(1. *School of Computer Science, Shaanxi Normal University, Xi'an* 710062, *China*; 2. *School of Electronic Engineering, Xidian University, Xi'an* 710071, *China*; 3. *National Laboratory of Automatic Target Recognition (ATR), School of Information Engineering, Shenzhen University, Shenzhen Guangdong* 518060, *China*)

Abstract: To overcome the sensible of traditional K-means clustering algorithm to initial centers, and avoid the arbitrary of available improved K-means algorithms for discovering good initial centers, this paper proposed a new algorithm to find the optimal initial centers for K-means clustering algorithm. It defined the density and the neighborhood for each sample according to the natural pattern distribution of exemplars in data space, so that the samples chose as initial seeds not only lie in the higher density area, but also far away from each other. It tested the new algorithm on some well-known datasets from UCI machine learning repository and on some synthetic datasets with different proportion noises using many different measures. The experimental results demonstrate that our new algorithm achieves excellent clustering result in short run time and is insensible to noisy data. It outperforms the traditional K-means clustering algorithm and those available algorithms for improving the initial seeds of K-means clustering algorithm.

Key words: clustering; K-means clustering; initial centers; neighborhood; density of pattern distribution

0 引言

聚类是模式识别、数据挖掘、机器学习等领域的重要研究内容之一,在识别数据的内在结构方面具有极其重要的作用[1]。基于物以类聚原理,聚类将一组个体按照相似性归成若干类别,使得同一类的个体间差别尽可能小,而不同类的个体间差别尽可能大[2]。根据对象间相似性度量和聚类评价准则的不同,聚类算法可分为划分方法、层次方法、基于密度的方法、基于网格的方法以及基于模型的方法[1,2]。

K-均值聚类算法是最经典且应用最广泛的划分方法之一,其理论可靠、算法简单、收敛速度快并且能有效地处理大数据集。但传统 K-均值算法过度依赖初始条件,如聚类数目 k 值需要预先给定、聚类结果依赖于初始聚类中心、不同的样本输入次序会使聚类结果发生改变等。在 k 值确定的情况下,不同的初始聚类中心可能得到不同的聚类结果;初始聚类中心与聚类结果之间具有明显的函数对应关系;算法运算效率对初始聚

收稿日期: 2011-08-23; **修回日期**: 2011-09-30 **基金项目**: 中央高校基本科研业务费专项资金重点资助项目(GK200901006);陕西省自然科学基础研究计划资助项目(2010JM3004);中央高校基本科研业务费专项资金资助项目(GK201001003)

作者简介:谢娟英(1971-),女,陕西西安人,副教授,硕导,CCF 会员,主要研究方向为智能信息处理、模式识别、机器学习、数据挖掘(xiejuany@snnu.edu.cn;);郭文娟(1986-),女,甘肃武威人,硕士研究生,主要研究方向为智能信息处理、模式识别;谢维信(1941-),男,广东广州人,教授,博导,主要研究方向为智能信息处理与目标识别、智能人机交互、图像处理和模式识别;高新波(1972-),男,山东莱芜人,教授,博导,主要研究方向为机器学习,计算智能和视觉信息处理、分析和理解及无线通信.

类中心敏感;算法缺乏对数据集样本实际分布情况的考虑,随机选取聚类中心增加了聚类结果的随机性和不可靠性。由此可见,聚类中心初始化问题在 K-均值聚类算法中有着重要的地位。国内外诸多学者致力于 K-均值算法的最佳初始中心选择研究,以改善 K-均值算法的聚类效果、聚类时间等性能[3~12],这些方法在一定程度上改进了 K-均值算法,但是没有充分利用数据集样本的空间分布信息,有一定的主观性。

为此,本文提出一种基于样本空间分布密度的初始聚类中心优化 K-均值算法,充分利用数据集样本的空间分布信息,根据数据集的样本分布特征划分数据集,启发式地确定初始聚类中心,避免了传统 K-均值算法的聚类中心随机选取,并利用样本的空间分布信息尽可能地避免现有相关算法的主观性。UCI 机器学习数据库数据集以及随机生成的带有噪声点的人工模拟数据集的实验测试证明,本文算法具有更好的聚类效果,聚类准确率优于传统 K-均值算法以及文献[8,10,11]的基于密度的优化初始聚类中心 K-均值算法,且对噪声数据有很强的抗干扰性能。

1　K-均值算法及其研究现状

传统 K-均值聚类算法以聚类误差平方和作为度量聚类质量的准则函数。首先随机选取 k 个数据对象作为初始聚类中心,根据其余数据对象到各聚类中心的距离将其分配到各个类簇中,重新计算各类簇的新中心,再次分配样本到新中心,计算各类簇的新中心,再分配样本。如此反复迭代直到准则函数收敛,即聚类中心不再发生变化。

K-均值算法的初始聚类中心影响算法的聚类效果,算法的迭代次数依赖于初始聚类中心与实际聚类中心的偏差,因此,恰当的聚类中心不仅可以提高 K-均值聚类结果的准确性,还可以加快聚类算法的收敛速度。国内外诸多学者致力于 K-均值算法初始聚类中心选择的研究。Kaufman 等人[3]提出一种启发式方法,估计数据点的局部密度,以该密度为启发选择 K-均值算法的初始聚类中心;Dhillon 等人[4]在迭代过程中重新计算类中心以提高聚类性能;Khan 和 Deelers 等人[5,6]也分别在初始化问题上作了有意义的尝试;钱线等人[7]运用谱方法估计特征中心来初始化聚类中心;袁方等人[8]通过密度参数计算数据对象所在区域的密度,选择相互距离最远的 k 个处于高密度区域的点作为初始聚类中心;赖玉霞等人[9]采用聚类对象分布密度方法,选择相互距离最远的 k 个处于高密度区域的点作为初始聚类中心;汪中等人[10]采用密度敏感的相似性度量计算数据对象的密度,启发式地生成初始聚类中心;王赛芳等人[11]通过计算点密度来选取初始聚类中心;韩凌波等人[12]通过计算每个数据对象的密度参数选取 k 个处于高密度分布的点作为初始聚类中心。上述方法一定程度上改进了 K-均值算法的性能,但算法中需要主观选择一些参数,这使得算法在一定程度上缺少了客观性;同时在聚类准确率、初始化聚类中心所用时间和聚类迭代次数方面有待进一步提高。

2　基于样本空间分布密度的初始聚类中心优化 K-均值算法

在用欧氏距离作为相似性度量的 K-均值聚类算法中,一般认为选择相互距离较远的 k 个数据对象作为初始聚类中心比随机选取 k 个数据对象作为聚类中心更具有代表性[8]。然而如果只是单纯地选取相互距离较远的 k 个数据对象来代表 k 个不同的类簇,有时会选取到孤立点或类边缘点,或者一个类中选取了两个以上的数据对象,这将导致聚类结果偏差,聚类过程收敛缓慢,甚至出现空类现象。比较合适的聚类中心应位于数据分布比较密集的区域,且相互之间距离较远。

文献[8,10,11]是近年国内学者提出的基于密度的 K-均值初始聚类中心优化算法,但这些算法在定义密度时带有很大的主观性,未能很好地反映数据集的分布特点。

本文基于样本空间分布密度的初始聚类中心优化 K-均值算法,充分利用样本的空间分布信息,根据数据集中数据对象的自然分布情况,定义样本的密度和邻域,密度越小表明该数据对象周围样本分布越密集。首先选出密度最小的数据对象,即数据集中处于数据最密集区域的样本,以该数据对象为中心,将整个数据集划分成不同的环形区域,即该样本的不同邻域;分析各个邻域中包含的数据对象数目,选出包含样本最多的前 k 个邻域,分别从这些邻域中选择周围数据最密集的数据对象作为聚类初始中心。这样所选的初始聚类中心不仅处于数据密集区域,并且各个中心点间距离较远。由于各个数据集规模不同,通过邻域半径调节系数可以调节邻域大小,从而选出最合适的初始中心点。

本文算法相关概念、详细步骤以及时间复杂度分析分别如下。

2.1　相关基本概念

设给定含有 n 个数据对象的数据集合 $X=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,每个数据对象含有 p 维特征,现欲将该数据集合划分为 k 个类簇 $C_j(j=1,2,\cdots,k,k<n)$。第 i 个数据对象的第 j 个特征值为 x_{ij}。

定义 1　任意两个数据对象间的欧氏距离定义为

$$d(\boldsymbol{x}_i,\boldsymbol{x}_j)=\sqrt{\sum_{a=1}^{p}(x_{ia}-x_{ja})^2}\quad i=1,2,\cdots,n;j=1,2,\cdots,n$$

定义 2　数据集中每个数据对象 $\boldsymbol{x}_i$ 对应的密度 density($\boldsymbol{x}_i$)定义为

$$\text{density}(\boldsymbol{x}_i)=\sum_{j=1}^{n}\frac{d(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\sum_{l=1}^{n}d(\boldsymbol{x}_l,\boldsymbol{x}_j)}\quad i=1,2,\cdots,n$$

定义 3　聚类误差平方和 E 的定义为:$E=\sum_{j=1}^{k}\sum_{t=1}^{n_j}\|\boldsymbol{x}_{jt}-\boldsymbol{m}_j\|^2$,$\boldsymbol{x}_{jt}$是第 j 类的第 t 个数据对象,$\boldsymbol{m}_j$ 是第 j 类的聚类中心,n_j 是第 j 类样本(数据对象)个数。

定义 4　数据对象的邻域半径 R 定义为:$R=n^{cR}\times\frac{1}{n}\sum_{i=1}^{n}e^{-\text{density}(\boldsymbol{x}_i)}$,$cR$ 是邻域半径调节系数。

定义 5　数据对象的 M 邻域。对于任意数据对象 $\boldsymbol{x}_i$,以 $\boldsymbol{x}_i$ 为中心,以 $M-1$ 倍的 R 和 M 倍的 R 为半径形成的环形区域所包含的数据对象集合,称为数据对象 $\boldsymbol{x}_i$ 的 M 邻域,用 δ_M 表示,形式化定义如下:

$$\delta_M=\{\boldsymbol{x}_j\mid(M-1)\times R<d(\boldsymbol{x}_i,\boldsymbol{x}_j)<M\times R\}$$

$$M = 1,2,\cdots,M_{\max};j = 1,2,\cdots,n$$

$$M_{\max} = \text{int}[(\max(d(x_i,x_j)) - \min(d(x_i,x_j)))/R]$$

2.2　算法描述

1)初始化中心点

a)根据定义 2 计算数据集中每个数据对象 x_i 的密度值 density(x_i),并依据 density(x_i)值将数据集升序排列;初始化聚类中心点集 C 为空,即 $C = \varnothing$。

b)从当前 x 中选择 density(x_i)值最小的样本 $x_{\min}$,计算该数据对象到数据集中其他样本的距离。

c)根据定义 5 求得数据对象 $x_{\min}$ 的 M 邻域 $\delta_M(M = 1,2,\cdots,M_{\max})$。

d)统计数据对象 $x_{\min}$ 的 M 邻域 δ_M 中的样本数量,降序排列,从中选出包含样本个数最多的前 k 个邻域 $\delta_l(l = 1,2,\cdots,k)$。

e)分别在邻域 $\delta_l(l = 1,2,\cdots,k)$ 中选出密度值最小的数据对象 x_l,将该对象加入到中心点集 C 中,即 $C = C \cup \{x_l\}$。

f)输出包含有 k 个中心点的初始中心点集 C,即 $|C| = k$。

2)更新类簇中心点

a)将数据集中数据对象分配到与其距离最近的聚类中心所在类簇中,根据定义 3 计算聚类误差平方和 E。

b)计算每一类簇样本的平均值,以该值作为新的聚类中心点,更新中心。

3)分配数据

a)将数据集中的数据对象分配到与其距离最近的聚类中心点所在类簇中。

b)根据定义 3 计算聚类误差平方和,若聚类误差平方和没有变化,则结束迭代;否则转到 2)继续执行。

2.3　算法复杂度分析

传统 K-means 的时间复杂度为 $O(nkT)$,其中 T 为算法的迭代次数,k 为类簇数,n 为数据集样本个数。本文算法与传统 K-means 算法相比,增加了初始类簇中心的选择,时间复杂度变为 $O(n^2 + nkt)$,t 为聚类过程迭代次数,即算法步骤 2)和 3)的重复执行次数。但是本文算法选择的初始聚类中心是理想的最佳初始聚类中心,因此 $t < T$。在小规模数据集上,本文算法的时间性能依然能保持传统 K-means 的时间性能。然而,本文算法在数据集规模很大的情况下,此时 $k << n$,时间性能将不如传统 K-means,因为选择最佳初始聚类中心的时间消耗 $O(n^2)$将占主导地位。实验测试也证明了这里的复杂度分析。

3　实验结果与分析

本文实验分别在 UCI 数据集和随机生成的人工模拟数据集两大类数据集上进行。实验环境为:Intel 酷睿 2 6320@ 1.86 GHz,1 GB 内存,160 GB 硬盘,WindowsXP 操作系统,MATLAB 应用软件。

3.1　UCI 机器学习数据库数据集实验

选用 UCI 机器学习数据库[13]中的 iris 等九个聚类算法测试常用的数据集对本文基于样本空间分布密度的初始聚类中心优化 K-均值算法进行测试,并与传统 K-means 聚类算法和文献[8,10,11]中的算法进行比较。同时为了测试本文算法对于大数据集的处理能力,对 segmentation 数据集不仅选用了常用的包含 210 个样本的数据集,还选用了包括 2 310 个样本的数据全集进行模拟实验。为了区分不同大小的 segmentation 数据集,将 segmentation 数据集全集后标有-test。实验所用 UCI 数据集描述如表 1。

表 1　数据集描述

数据集	样本数	属性数	类数
soybean-small	47	35	4
iris	150	4	3
wine	178	13	3
segmentation	210	19	7
ionoshpere	351	34	2
WDBC	569	30	2
pima indians diabetes	768	8	2
yeast	1 484	8	10
segmentation-test	2 310	19	7

关于算法聚类结果的评价,除采用常用的聚类误差平方和、聚类时间、聚类准确率评价方法之外,还采用 Rand 指数、Jaccard 系数[14~16]和 Adjusted Rand index 参数[17]对聚类结果进行比较分析。后三个聚类评价指标都是在已知正确分类信息的前提下对聚类算法的聚类结果进行评价的有效指标。其中,Adjusted Rand index 参数是最好的聚类有效性评价准则[18]。

三个评价指标的定义如下:设 U 和 V 分别是关于数据集的两种划分,其中 U 是已知的正确划分,而 V 是通过某种聚类算法得到的划分结果,定义 a、b、c、d 四个参数。设 a 为在 U 和 V 都在同一类的样本对数目;b 表示在 U 中为同一类,而在 V 中却不在同一类的样本对数目;c 表示在 V 中为同一类,而在 U 中却不在同一类的样本对数目;d 为 U 和 V 都不在同一类簇的样本对数目。$a + b + c + d = n(n-1)/2$,其中,n 为数据集中所有样本数,也即数据集的规模。

设 $M = a + b + c + d$,则 M 表示所有可能的样本对。Rand 指数、Jaccard 系数和 Adjusted Rand index 参数的定义如下:

Rand 指数:$R = (a + d)/M$

Jaccard 系数:$J = a/(a + b + c)$

Adjusted Rand index 参数:

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

其中:R 表示 Rand 指数;J 表示 Jaccard 系数;RI 表示 Adjusted Rand index 参数。

从定义可知,Rand 指数表示聚类结果与原始数据集样本分布的一致性;Jaccard 系数表示实现正确聚类样本对占聚类前或后在同一类簇样本对的比率;RI 值越大表示实现正确聚类的样本对越多,聚类效果越好(其上界为 1,表示聚类结果与原始数据集的样本分布完全一致;下界为 -1,表示聚类结果与原始数据集的样本分布完全不一致)。

UCI 数据集的实验结果如表 2、3 和图 1~4 所示,其中 K-

means 算法的实验结果为算法运行 20 次所得结果的平均值。表 2 和 3 分别是 K-means、文献[8,10,11]和本文算法聚类误差平方和、聚类时间的比较。图 1 ~ 4 分别是 K-means、文献[8,10,11]和本文算法 Rand 指数、Jaccard 系数、Adjusted Rand index 参数以及聚类准确率结果的比较。

表 2　UCI 数据集上聚类误差平方和比较

数据集	K-means 算法	文献[8]算法	文献[10]算法	文献[11]算法	本文算法
soybean-small	249.418	228.516	228.516	347.406	222.114
iris	98.57	78.9451	78.9451	78.9451	78.9408
wine	2.42E +06	2.37E +06	2.37E +06	2.63E +06	2.37E +06
segmentation	3.24E +06	3.24E +06	2.95E +06	2.91E +06	2.90E +06
ionoshpere	2.72E +03	2.42E +03	2.42E +03	2.42E +03	2.42E +03
WDBC	7.99E +07	7.79E +07	7.79E +07	7.79E +07	7.79E +07
pima indians diabetes	5.64E +06	5.14E +06	5.14E +06	5.14E +06	5.14E +06
yeast	58.607	57.455	46.600	47.533	46.547
segmentation-test	1.62E +07	1.58E +07	2.09E +07	1.58E +07	1.40E +07

由表 2 可见,在 wine 数据集上,本文算法的聚类误差平方和等于文献[8,10]算法的聚类结果,小于传统 K-means 和文献[11]的算法;在 ionoshpere、WDBC 和 pima indians diabetes 数据集上,本文算法的聚类误差平方和等于文献[8,10,11]算法的聚类误差平方和,明显小于传统 K-means 算法;在其他数据集上,本文算法的聚类误差平方和小于其他四种算法的聚类结果。因此,本文算法具有更好的聚类效果。

表 3　UCI 数据集上聚类时间比较　　s

数据集	K-means 算法	文献[8]算法	文献[10]算法	文献[11]算法	本文算法
soybean-small	0.073	0.016	0.016	0.078	0
iris	0.003	0.046	0.031	0.188	0.031
wine	0.006	0.032	0.047	0.051	0.031
segmentation	0.118	0.142	0.094	0.063	0.053
ionoshpere	0.089	0.047	0.078	0.072	0.042
WDBC	0.013	0.125	0.156	0.094	0.072
pima indians diabetes	0.153	0.213	0.234	0.265	0.144
yeast	0.499	1.656	1.297	0.468	0.343
segmentation-test	1.109	4.031	4.297	3.838	3.609

表 3 聚类时间比较显示,本文算法时间性能明显优于文献[8,10,11]算法,具有更快的收敛速度。但在有些数据集上本文算法不及 K-means 算法,原因在于本文算法在初始化聚类中心点时,需要花费时间计算样本密度和划分邻域。

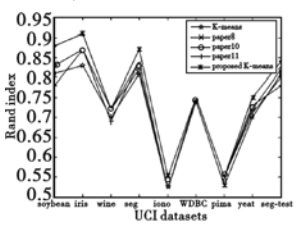

图1　UCI数据集上Rand指数比较

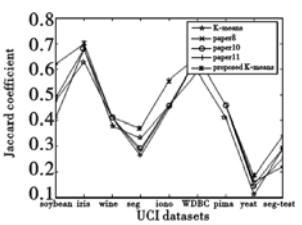

图2　UCI数据集上Jaccard系数比较

从图 1 ~ 3 所示的 Rand 指数、Jaccard 系数和 Adjusted Rand index 参数三个聚类有效性指标测试结果显示,本文算法优于 K-means 和文献[8,10,11]算法。图 4 显示,本文算法的聚类正确率明显高于 K-means 和文献[8,10,11]算法。

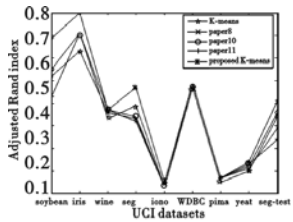

图3　UCI数据集上Adjusted Rand index参数比较

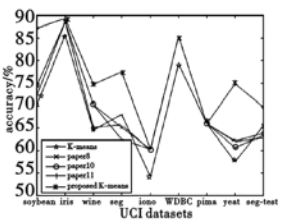

图4　UCI数据集上聚类准确率比较

从以上 UCI 数据集的实验结果分析可见,本文基于样本空间分布密度的初始聚类中心优化 K-均值算法具有良好的聚类效果,聚类时间短,聚类准确率更高。

3.2　人工模拟数据集实验

为进一步测试本文基于样本空间分布密度的初始聚类中心优化 K-均值算法对噪声数据的抗干扰性能,随机生成了分别含有 0、5%、10%、15%、20%、25%、30%、35%、40% 不同比例噪声的人工模拟数据集来对算法进行测试。模拟数据集包含三个类簇,每一类簇中含有 120 个二维样本,这些样本符合正态分布。其中第 i 类的横坐标 x 的均值为 μ_x^i,纵坐标 y 的均值为 μ_y^i,第 i 类的标准差为 σ^i。在第二类加入噪声点,噪声点的标准差为 σ^l。随机生成的人工模拟数据集生成参数如表 4 所示。

表 4　随机生成的人工模拟数据集的各项参数

参数项	第一类	第二类	第三类
均值	$\mu_x^1=0$ $\mu_y^1=0$	$\mu_x^2=6$ $\mu_y^2=2$	$\mu_x^3=6$ $\mu_y^3=-1$
标准差	$\sigma^1=1.5$	$\sigma^2=0.5$ $\sigma^l=2$	$\sigma^3=0.5$

在随机生成含有 0、5%、10%、15%、20%、25%、30%、35%、40% 不同比例噪声的九个人工模拟数据集上分别运行 K-means、文献[8,10,11]以及本文算法,实验结果如表 5、6 和图 5 ~ 8 所示。其中 K-means 算法的实验结果为算法运行 20 次的平均值。表 5 和 6 是 K-means、文献[8,10,11]以及本文算法的聚类误差平方和以及聚类时间比较。图 5 ~ 8 分别是 K-means、文献[8,10,11]以及本文算法的 Rand 指数、Jaccard 系数、Adjusted Rand index 参数和聚类正确率的结果比较。

表 5　人工数据集上聚类误差平方和比较

噪声数据比例/%	K-means 算法	文献[8]算法	文献[10]算法	文献[11]算法	本文算法
0	686.393	641.567	641.567	641.567	614.011
5	779.487	704.984	704.984	704.984	689.742
10	677.311	609.696	583.410	583.410	559.696
15	901.892	804.424	784.424	784.424	775.633
20	855.461	824.131	809.964	804.875	800.964
25	907.977	756.806	733.913	733.913	727.126
30	952.142	776.566	776.566	776.566	755.171
35	911.357	860.585	860.585	860.585	832.421
40	989.306	907.221	907.221	907.221	882.850

表 5 结果显示,在含有不同比例噪声的数据集上,本文算法的聚类误差平方和均小于其他四种算法,有很好的聚类效

果。表6聚类时间的比较显示,本文算法时间性能优于文献[8,10,11]算法,具有更快的收敛速度。但在个别数据集上,本文算法不及 K-means 算法,原因在于本文算法在初始化聚类中心点时,需要花费时间计算样本密度和邻域。

表6　人工数据集上聚类时间比较　　s

噪声数据比例/%	K-means 算法	文献[8]算法	文献[10]算法	文献[11]算法	本文算法
0	0.005	0.047	0.063	0.063	0.023
5	0.003	0.031	0.234	0.047	0.234
10	0.003	0.047	0.047	0.047	0.036
15	0.118	0.142	0.063	0.094	0.094
20	0.025	0.047	0.063	0.047	0.047
25	0.004	0.047	0.063	0.063	0.047
30	0.153	0.213	0.163	0.132	0.125
35	0.003	0.077	0.063	0.063	0.063
40	0.004	0.047	0.109	0.132	0.109

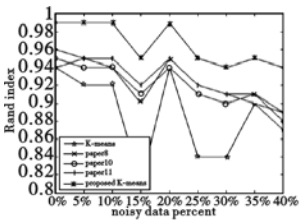

图5　人工数据集上Rand指数比较

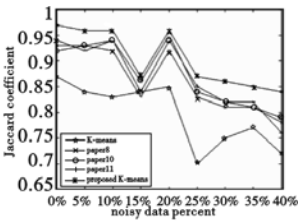

图6　人工数据集上Jaccard系数比较

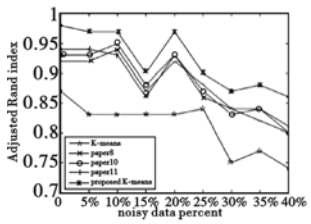

图7　人工数据集上Adjusted Rand index参数比较

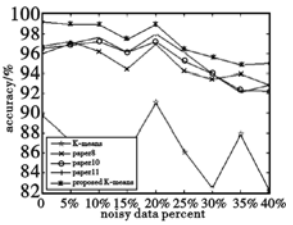

图8　人工数据集上聚类准确率比较

由图5~8的 Rand 指数、Jaccard 系数以及 Adjusted Rand index 参数比较显示,在含有不同比例噪声的数据集上,本文算法的这三个非常有效的聚类结果评价指标均高于 K-means、文献[8和10,11]算法。由图8的聚类准确率比较显示,本文算法聚类结果的准确率最高,明显优于其他四种算法。

以上人工模拟数据集实验结果说明,本文基于样本空间分布密度的初始聚类中心优化 K-均值算法对含有噪声的人工模拟数据集有非常好的聚类效果,具有很强的抗噪声能力。

4　结束语

本文在分析传统 K-均值聚类算法和现有优化初始中心 K-均值聚类算法不足的基础上,提出一种基于样本空间分布密度的初始聚类中心优化 K 均值算法。该算法利用数据集样本的自然分布信息定义了数据对象的密度,并根据数据集的空间分布定义了数据对象的邻域,对数据集进行划分;选择数据集中位于样本分布密集区且相距较远的数据对象作为初始聚类中心点,克服了传统 K-均值算法对初始聚类中心敏感、聚类效果不理想的缺陷;同时在一定程度上克服了现有基于密度的初始聚类中心优化算法的主观性缺点。UCI 机器学习数据库数据集和随机生成的带有不同比例噪声的人工模拟数据集实验共同表明:本文基于样本空间分布密度的初始聚类中心优化 K-均值算法有很好的聚类效果,不仅聚类速度快、准确率高,而且对噪声数据有很强的抗干扰性能,聚类性能优于传统 K-menas 算法以及文献[8,10,11]的优化初始聚类中心 K-均值算法。

参考文献:

[1] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报, 2008,19(1):48-61.

[2] HAN Jia-wei,KAMBER M. 数据挖掘概念与技术[M]. 范明, 孟小峰,译. 北京:机械工业出版社,2000.

[3] KAUFMAN L, ROUSSEEUW P J. Finding groups in data: an introduction to cluster analysis[M]. New York:Wileys, 1990.

[4] DHILLON I S,GUAN Yu-qiang, KOGAN J. Refining clusters in high dimensional text data [C]//Proc of the 2nd SIAM Workshop on Clustering High Dimensional Data. 2002: 59-66.

[5] KHAN S S, AHMAD A. Cluster center initialization for K-means clustering [J]. Pattern Recognition Letters,2004, 25(11):1293-1302.

[6] DEELERS S, AUWATANAMONGKOL S. Enhancing K-means algorithm with initial cluster centers derived from data partitioning along the data axis with the highest variance[J]. Proceeding of World Academy of Science, Engineering and Technology,2007,26: 323-328.

[7] 钱线,黄萱菁,吴立德. 初始化 K-means 的谱方法[J]. 自动化学报,2007,33(4):342-346.

[8] 袁方,周志勇,宋鑫. 初始聚类中心优化的 K-means 算法[J]. 计算机工程,2007,33(3):65-66.

[9] 赖玉霞,刘建平. K-means 算法的初始聚类中心的优化[J]. 计算机工程与应用,2008,44(10):147-149.

[10] 汪中,刘贵全,陈恩红. 一种优化初始中心点的 K-means 算法[J]. 模式识别与人工智能, 2009,22(2):299-304.

[11] 王赛芳,戴芳,王万斌,等. 基于初始聚类中心优化的 K-均值算法[J]. 计算机工程与科学, 2010,32(10):105-107.

[12] 韩凌波,王强,蒋正峰,等. 一种改进的 K-means 初始聚类中心选取算法[J]. 计算机工程与应用,2010,46(17):150-152.

[13] FRANK A, ASUNCION A. UCI machine learning repository[R]. California:University of California, School of Information and Computer Science, 2010.

[14] 张惟皎,刘春煌,李芳玉. 聚类质量的评价方法[J]. 计算机工程, 2005,31(20):10-12.

[15] 于剑,程乾生. 模糊聚类方法中的最佳聚类数的搜索范围[J]. 中国科学(E辑), 2002,32(2):274-280.

[16] 杨燕, 靳蕃, KAMEL M. 聚类有效性评价综述[J]. 计算机应用研究, 2008, 25(6):1631-1632.

[17] HUBERT L, ARABIE P. Comparing partitions [J]. Journal of Classification,1985,2(1):193-218.

[18] VINH N X, EPPS J, NAILEY J. Information theoretic measures for clustering comparison: is a correction for chance necessary [C]// Proc of the 26th Annual International Conference on Machine Learning. New York:ACM Press,2009:1073-1080.