一种改进的密度加权的模糊 C 聚类算法①

王行甫, 程用远, 覃启贤

(中国科学技术大学 计算机学院, 合肥 230027)

摘 要: 模糊 C 均值聚类算法(FCM)是一种流行的聚类算法,在许多工程领域有着广泛的应用. 密度加权的模糊 C 均值算法(Density Weighted FCM)是对传统 FCM 的一种改进,它可以很好的解决 FCM 对噪声敏感的问题. 但是 DWFCM 与 FCM 都没有解决聚类结果很大程度上依赖初始聚类中心的选择好坏的问题. 提出一种基于最近邻居节点对密度的 FCM 改进算法 Improved-DWFCM,通过最近邻居节点估计节点密度的方法解决聚类结果对初始簇中心依赖的问题. 仿真结果表明这种算法选择出来的初始聚类中心与最终结果的簇中心非常接近, 大大提高了算法收敛的速度以及聚类的效果.

关键词: 模糊聚类; 基于密度加权的模糊 C 聚类; 初始聚类中心; 最近邻居节点对; 密度

Improved Density Weighted Fuzzy C Means Algorithm

WANG Xing-Fu, CHENG Yong-Yuan, QIN Qi-Xian

(School of Computer Science, University of Science and Technology of China, Hefei 230027, China)

Abstract: Fuzzy C Means algorithm is popular soft clustering algorithm. It has been applied in many engineering fields. Density weighted FCM is its variant, which can solve FCM's problem: sensitive to outlier and noise data. However, performances of both algorithms are heavily depend on proper initial cluster centers. This paper proposes a novice algorithm: Improved density weighted FCM based on nearest neighbor pair and its density, simulation results show initial center produced by the algorithm are very close to final cluster center. Thus IDWFCM can convergent very quickly and improve the performance.

Key words: fuzzy C means; improved density weighted fuzzy C means; initial cluster center; nearest neighbor data pair; density

聚类是一种重要的非监督模式学习算法, 在许多工程领域如: 模式识别、系统建模、图像处理、通信、数据挖掘方面有广泛的应用. 聚类是将一组没有标签的样本数据分成几簇, 使得同一簇内的数据其相似度最大, 不同簇中数据之间相似度最小.

有许多聚类算法被提出, 这些算法可以粗略的分为两类: 硬聚类算法、软聚类算法. 对硬聚类算法来说, 一个样本数据必须只能属于唯一的一个簇. 然而软聚类没有这个限制, 一个样本对象可能同时属于几个簇, 这种属于不同簇的程度用模糊隶属度函数来描述. 这种模糊的性质在有些应用中十分重要. 聚类算法中模糊 C 均值算法(FCM)是一种十分流行的聚类算法, 许多学者对 FCM 进行广泛的研究[1-5]. 然而, FCM 也有一些缺点: 初始簇中心选择对聚类结果好坏有很大的影响; 容易受到噪声数据影响; 由于隶属度函数矩阵通常很大, 因此 FCM 的计算复杂度通常较高.

许多针对 FCM 的改进算法先后被设计出来解决 FCM 算法的不足, 其中比较有名的如: 模糊概率 FCM, 基于可信度的 CFCM, 基于密度加权的 FCM. 其中基于密度加权的模糊 C 聚类(DWFCM)[6-7]可以很好的解决对噪声数据敏感的问题. 但是这些算法都没有考虑初始簇中心的选择问题, 随机产生的簇中心的

① 基金项目:国家科技重大专项(2012ZX10004-301-609);国家自然科学基金(60970128);安徽省教学研究计划 2010
收稿时间:2012-01-16;收到修改稿时间:2012-03-06

聚类效果通常比较差.

本文提出的 IDWFCM 可以选择一个很好的初始簇中心从而大大降低计算法复杂度, 改善聚类效果. 本文结构如下: 第一部分是对传统FCM以及密度加权FCM 理论的一个快速回顾. 第二部分详细的介绍了我们的算法 IDWFCM. 第三部分对对仿真结果进行分析和总结. 最后一部分是对全文的一个总结.

1 模糊C聚类与密度加权模糊C聚类算法

1.1 模糊 C 聚类(Fuzzy C Means)

模糊 C 聚类算法(Fuzzy C Means)是一种目前被广泛使用的软聚类算法. 它假设簇的数目 C 是固定的. 然后 FCM 可以转换为下列目标的极小值问题.

$$J_{fcm}=\sum_{i=1}^{c}\sum_{k=1}^{N}u_{ik}^{m}d_{ik}^{2} \tag{1}$$

其中, $d_{ik}=\|x_k-v_i\|$

约束条件为:

$$\sum_{i=1}^{N}u_{ik}=1;k=1,...,N \tag{2}$$

其中, u_{ij} 在[0,1]之间, 表示样本集合中数据 Xj 属于 i 簇的程度; m 为影响隶属度矩阵模糊化的因子, 通常 m=2. N 代表样本的数目. V 矩阵是聚类中心矩阵. 利用拉格朗日数乘法, 问题可以等价为在约束条件下求下面方程的极小值.

$$L(U,\lambda)=\sum_{i=1}^{c}\sum_{k=1}^{N}u_{ik}^{m}d_{ik}^{2}+\sum_{k=1}^{N}\lambda_k(1-\sum_{i=1}^{c}u_{ik}) \tag{3}$$

由公式(3)可以得到下面的迭代方程:

$$u_{ik}=\left(\sum_{j=1}^{C}\left(d_{ik}/d_{jk}\right)^{2/(m-1)}\right)^{-1} \tag{4}$$

然后我们假定 u_{ik} 已知, 带入(1)式中, 可以得到 V 的迭代公式:

$$v_i=\sum_{k=1}^{N}(u_{ik}^{m}x_k)\Big/\sum_{k=1}^{N}(u_{ik}^{m}) \tag{5}$$

FCM 算法中我们通常随机的选择簇的初始中心. 簇的初始中心对聚类结果有很大的影响, 如果初始簇中心能选择的接近最终的簇中心, 那么FCM的可以很快的收敛. 从上面的公式我们还可以看出, 噪声数据或离群数据和普通数据的影响力是一样的. 基于密度的加权 FCM 能很好的解决对噪声数据敏感的问题.

1.2 基于密度加权的模糊 C 聚类算法(Density Weighted Fuzzy C Means)

密度加权的模糊 C 聚类算法是基于以下假设: 离群数据以及噪声数据通常远离普通数据. 我们给离群数据、噪声数据一种势能描述. DFCM 的公式如下:

$$L(U,\lambda)=\sum_{k=1}^{n}\sum_{i=1}^{c}D_k(u_{ik})^{m}\|x_k-v_i\|^{2}+\sum_{k=1}^{n}\lambda_k\left(1-\sum_{i=1}^{c}u_{ik}\right) \tag{6}$$

$$D_k=\sum_{y=1}^{n}\exp\left(-\frac{h\times\|x_k-x_y\|}{STD}\right) \tag{7}$$

其中, D_k 是势能函数. h 是解析度因子, STD 是输入数据的标准方差. 从 D_k 可以看出, 如果数据离得比较近, 那么 D_k 的值就比较大. 因此 D_k 可以用来表述某样本数据周围的数据密度. DWFCM 的迭代公式如下:

$$v_i=\frac{\sum_{k=1}^{n}D_k(u_{ik})^{m}x_k}{\sum_{k=1}^{n}D_k(u_{ik})^{m}} \tag{8}$$

DWFCM 可以很好克服 FCM 对噪声数据敏感的缺点.

2 改进的密度加权FCM(Improved Density Weighted Fuzzy C Means)

2.1 概述

传统的模糊 C 聚类以及密度加权 FCM 的聚类结果都对初始簇的选择有很大的依赖. 随机产生的簇中心通常会导致比较坏的聚类结果. 本文提出算法能产生与最终簇中心非常近的初始簇中心, 因此能大大降低迭代收敛的步数, 从而降低计算复杂性改进聚类结果. 本算法是对初始簇中心的选择是基于最近邻居数据对以及它们周围的数据密度. 它是基于以下两点对输入数据的观察:

1) 簇中心周围的数据密度通常很大.

2) 由于簇中心的密度通常比较大,簇中心周围通常有一些距离非常近的邻居节点对.

通过计算这些邻居节点对之间的距离以及它们周围的数据密度, 选出密度最大的邻居节点对, 同时对一些比较近的簇中心我们采用归并的操作进行融合, 这样通常产生的初始簇中心与结果簇中心很接近.

2.2 算法步骤

基于上面两点观察, 我们的算法包括两个步骤:

初始簇中心生成、传统 DWFCM 步骤. 为了更好的描述算法, 我们先引入一些辅助变量. 变量 D 描述的是距离矩阵, 其中 d(i,j)表示数据 i 和数据 j 之间的欧式距离. 很明显 D 是一个对称矩阵,我们可以只存储上三角来节约存储空间. U 是隶属度函数矩阵, V 是簇中心矩阵.

初始簇中心生成步骤如下:

Step1. 计算距离矩阵 D, 存储 D 方便后面步骤使用.

Step2. 对 D 矩阵的每一行按升序排序,排序后 D[i][1]表示到第 i 个样本数据的最小距离. 利用 DD 矩阵存储每一行排序数据的下标索引.

Step3. 利用前面得到的矩阵 D,DD 选择距离最小的 k 对数据,进行排序.

Step4. 计算这些数据对的数据密度, 按照降序排序. (对每一行 i,我们统计有多少数据 j 的距离 d(i,j)<r, 利用这个值作为密度,r 一个预先设定的常量)

Step5. 按照数据密度从大到小选择数据对, 如果有两组数据对之间的距离小于某一个阈值我们就选择他们的中点进行归并形成一个簇中心.

Step6. 如果得到 k 组数据对我们就停止算法.

利用第一个步骤中产生的簇中心做为密度加权算法模糊 C 聚类的初始簇中心, 聚类步骤如下:

Step1. 利用第一个步骤得到的簇中心 Vk 作为初始簇中心.

Step2. 设置 k=1 作为迭代次数.

Step3. 根据公式(4)迭代更新 Uk.

Step4. 根据公式(5)迭代更新 Vk+1

Step5. k+1,然后转到 step3 直到

$$e=\sum_{i=1}^{c}\left\|V_i(k+1)+V_i(k)\right\|^2<\zeta$$

3 仿真结果分析

我们采用图 1 所示的数据作为例子对 IDWFCM 与传统 FCM 进行比较. 图 1 中的数据分布是由平面二维高斯分布产生的. 在算法的第一阶段我们先按照数据密度求出前 k 对最近邻居点对. 图 2 中所示的 6 个大的点(距离非常近的两个点)就是我们最后产生的 6 对最大邻近点对, 并且这邻接点对的密度最大. 从图 2 可以看出它们与簇中心已经比较接近了.

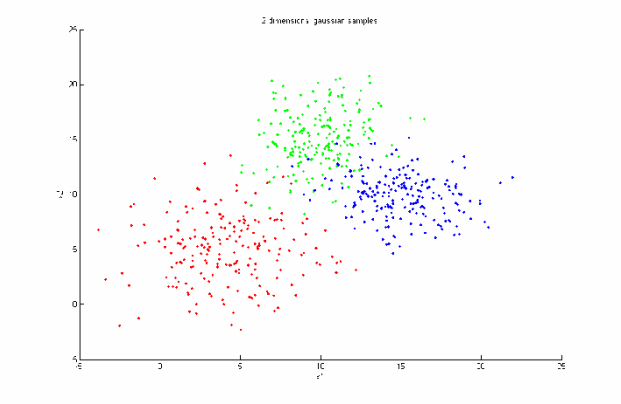

图 1　实验数据分布

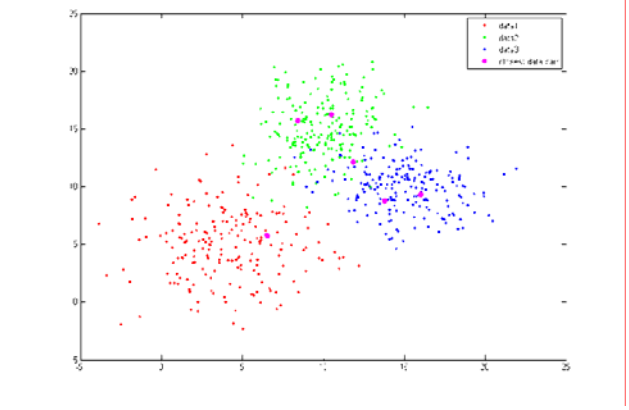

图 2　没有归并后的初始簇中心

但是有一些最近邻近点对距离非常近, 它们应该属于同一簇的. 如果我们不对这些点对进行归并操作, 最后会得到的簇数目就会不正确. 图 3 所示的就是归并的过程. 图中被圈起来的节点对进行了融合, 取它们的中心作为新的簇中心.

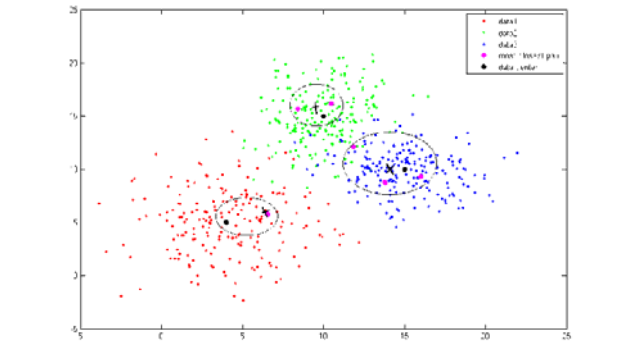

图 3　归并后的初始簇中心分布

从上面这些图可以看出, 经过归并以后的中心(图

3 中十字型的点)已经非常接近真实的簇中心了. 因为 IDWFCM 产生的初始簇中心与最终簇中心非常接近, 因此 IDWFCM 能非常快的收敛. 在我们的实验中, 传统的 FCM 平均要经过 21 次的迭代才能收敛, 而 IDWFCM 平均只需要 7 次迭代就可以收敛. 图 4 与图 5 显示了 2 中算法收敛的路径. 从图中我们看以看出 IDWFCM 收敛的速度要比 FCM 快很多.

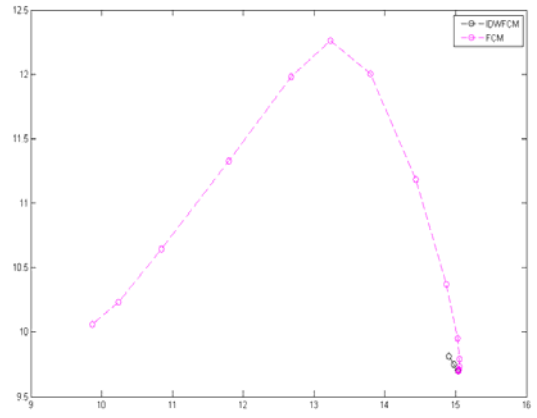

图 4 算法收敛路径 1

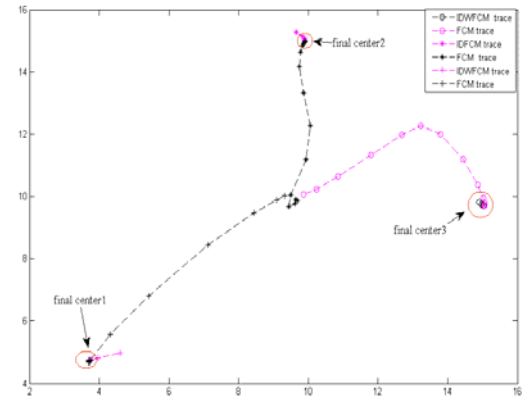

图 5 算法收敛路径 2

4 结语

模糊C聚类算法是一种广泛被使用的软聚类算法. 但是它的聚类性能与初始簇中心的选择有很大的关系, 随机的簇中心会大大降低聚类的效果. 较好的中心选择不但可以降低计算复杂度还可以改进聚类准确度. 本文我们提出一种改进的基于密度加权的模糊聚类算法(IDWFCM), 这种算法可以产生一个与真实簇中心比较接近的的初始中心. 仿真表明. IDWFCM 与传统 FCM 相比, 收敛速度明显提高, 聚类性能得到提高.

参考文献

1 Bezde JC. Pattern Recognition With Fuzzy Objective Function Algorithms. New York: Plenum, 1981.

2 Lesk J. Towards robust fuzzy clustering. Fuzzy Sets and Systems, 2003,137:215–233.

3 Chen JL, Wang JH. A new robust clustering algorithm-density-weighted fuzzy c-means. IEEE International Conference on Systems, Man, and Cybernetics, 1999,3: 90–94.

4 Zhang HZ, Chen H, Bao LX. An Improved Fuzzy C Means Clustering Algorithm and Its Application in Traffic Condition Recognition. 2010 7th Conference on Fuzzy System and Knowledge Discovery(FSDK 2010). 2010.

5 沈红斌,王士同,吴小俊.离群模糊核聚类算法.软件学报, 2004,15(7):1021−1029.

6 Dave RN, Krishnapm R. Robust clustering methods: A unified view. IEEE Trans. Fuzzy Syst.270–293,199.

7 Chen JL, Wang JH. A new robust clustering algorithm-density-weighted fuzzy c-means. Systems, Man, and Cybernetics, IEEE SMC'99 Conference Proceedings, 1999,3: 90–94.

(上接第 213 页)

Computer Vision and Image Understanding. 2011,115(6): 885–900.

9 Wang Z, Bovik AC, Sheikh HR, et al. From error visibility to structural similarity. IEEE Trans. on Image Processing, 2004,13(4):1−14.

10 Loza A, MihayLova L, Canagarajah N, et al. Structural similarity-based object tracking in video sequences. Proc. of the 9th International Conference on Information Fusion. Florence, USA: IEEE Press, 2006. 1−6.