均衡模糊C均值聚类算法

文传军[1] 汪庆淼[2] 詹永照[3]

(常州工学院理学院 常州 213002)[1] (苏州大学计算机学院 苏州 215021)[2]

(江苏大学计算机科学与通信工程学院 镇江 212013)[3]

摘 要 模糊C均值聚类算法没有考虑各类样本容量因素,当各类样本容量差异较大时,其聚类判决将向小样本类倾斜。提出一种新的聚类算法——均衡模糊C均值聚类,对模糊C均值聚类算法最小化目标函数进行修正,使得改进的目标函数包含了样本容量因素,利用粒子群算法并以样本模糊隶属度为编码对象求解参数优解。从理论上分析了该算法的性质,通过仿真实验验证了所提算法对平衡、不平衡数据集的有效性。

关键词 模糊C均值聚类,样本容量,均衡化,粒子群,全局优解

中图法分类号 TP391.4 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2014.08.053

Equalization Fuzzy C-means Clustering Algorithm

WEN Chuan-jun[1] WANG Qing-miao[2] ZHAN Yong-zhao[3]

(School of Science, Changzhou Institute of Technology, Changzhou 213002, China)[1]

(School of Computer Science and Technology, Soochow University, Suzhou 215021, China)[2]

(School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China)[3]

Abstract Fuzzy C-means clustering(FCM) is a fast and effective clustering algorithm, but it doesn't consider the difference of the samples size, while the capacities of each class are of large difference, and the decision of FCM will be benificial to the class with less samples. A new clustering algorithm was proposed in the paper and named as equalization fuzzy C-means clustering(EFCM). The minimum objective function of FCM was modified and the factor of samples size was added in EFCM objective function. The parameter optimal solutions of EFCM were calculated through PSO algorithm in which sample fuzzy memberships are seted as coding object. The properties of EFCM were obtained by theoretical analysis. The effectiveness of EFCM for balansed and unbalanced datasets was proved by simulation experiments.

Keywords Fuzzy C-means clustering, Samples size, Equalization, Particle swarm, Global optimal solution

1 引言

聚类分析是机器学习和数据挖掘中的一个重要研究课题,是非监督模式识别中的一种重要方法。模糊C均值聚类算法(Fuzzy c-means clustering,FCM)是基于目标函数的一种模糊聚类算法,因计算简便、易于计算机实现和良好的聚类表现而成为应用最为广泛的聚类算法[1-3]。FCM算法也存在一些缺陷:如容易陷入局部最优解[4];对类别 K 值及模糊加权指数 m 的选择没有准则可依循;对异常数据较为敏感;聚类结果可能不平衡等。为克服FCM算法存在的不足,研究者从各自不同的角度提出了一系列扩展算法,文献[5]为了克服FCM初始聚类中心和局部优解的问题,将改进的PSO算法应用于FCM;文献[6]也使用了文献[5]类似的方法,将PCM、FCM与PSO算法相结合,以提高算法的抗噪和分类性能;Huang等人[7]给出了确定 m 取值的理论途径,揭示了 m 值与数据结构的局部关系;Wu等人[8]研究发现大的 m 取值使得FCM对噪声和野值更鲁棒。

不平衡数据集是指某类样本数量明显少于其它类样本的数据集,有监督分类在处理不平衡数据集时,分类判决总会倾向于多数类(负类),而导致少数类(正类)的分类精度很低[9]。关于不平衡数据集分类的研究主要集中于有监督分类,而较少涉及聚类分析领域,实际上聚类分析也存在不平衡数据分类问题。模糊C均值聚类系列算法是基于类内加权平方误差最小化准则的,样本通常归类于最近聚类中心所代表的类,当模糊聚类算法基于不平衡数据集进行分类时,聚类因样本容量差异而倾向于正类,导致聚类结果不平衡。文献[10]提出广义均衡模糊聚类算法(GEFCM),将样本容量因素引入FCM目标函数中,使得GEFCM算法在 $m>1$ 时对平衡或不平衡数据集均有效。本文在文献[10]的基础上做了进一步研究,提出了均衡模糊C均值聚类算法(equalization fuzzy C-means clustering,EFCM),将均衡模糊聚类算法扩展到 $m=1$ 时的情况,EFCM在FCM最优化准则中引入类别容量因素,利用粒子群优化算法求解目标函数,对样本模糊隶属度编码计算参数优解,理论分析和讨论了算法的性质,通过仿真实验

到稿日期:2013-06-25 返修日期:2013-10-13 本文受国家自然科学基金(61170126)资助。

文传军(1976-),男,博士,主要研究方向为模式识别、模糊聚类,E-mail:wcjyhy@qq.com;詹永照(1962-),男,教授,博士生导师,主要研究方向为分布式计算、计算机图形学。

验证了算法的有效性。

2 均衡模糊 C 均值聚类(EFCM)

2.1 算法构想

K 均值聚类算法(硬聚类,K-means clustering,HCM)及其系列改进算法都具有类似的构造思路,即首先确定非相似性指标最优化准则,亦即最小化代价函数,然后根据该准则获取参数迭代公式,最后建立聚类算法迭代流程。

HCM 算法的目标为最小化代价函数:

$$J=\sum_{i=1}^{c}Ji=\sum_{i=1}^{c}(\sum_{k,x_k\in Gi}\| x_k-c_i\|^2) \tag{1}$$

其中,c 表示聚簇划分数,G_i 表示第 i 组样本划分,c_i 为其聚类中心,HCM 取各组向量的均值为聚类中心,即 $c_i=\frac{1}{n_i}\sum_{x_k\in G_i}x_k$。

由于 HCM 算法仅仅考虑样本与聚类中心的距离,没有考虑各类样本的容量差异,使得 HCM 算法对正本类倾斜。为了克服类样本容量差异对聚类判决的影响,HCM 算法的一个改进优化准则为:

$$J=\sum_{i=1}^{c}\frac{Ji}{n_i}=\sum_{i=1}^{c}(\frac{1}{n_i}\sum_{k,x_k\in Gi}\| x_k-c_i\|^2) \tag{2}$$

其中,n_i 表示第 i 组样本的容量,用以消除各聚簇容量差异对聚类结果的影响。

FCM 与 HCM 的主要区别在于 FCM 用模糊划分,使得每个给定数据点用 0,1 间的隶属度来确定样本属于各个组的程度,这更加符合客观实际,FCM 最优化准则为:

$$J(U,P)=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^m\| x_j-c_i\|^2 \tag{3}$$

其中,u_{ij} 为样本 x_j 属于第 i 组聚簇的程度,且有 $\sum_{i=1}^{c}u_{ij}=1$;m 表示模糊度控制参数,当 $m=1$ 时,FCM 则退化为 HCM,式(3)退化为式(1)[4]。

FCM 最优化准则与 HCM 类似,仅仅考虑了样本与聚类中心的距离,距离越小则隶属程度越高。由于 FCM 没有考虑各类样本容量的差异,样本容量大的类因样本空间分布广泛,其边缘样本与该类中心的距离有可能大于与其它类中心的距离,由此导致 FCM 的误判,即原始 FCM 仅适用于各类样本容量均衡的情况。文献[9]提出 GEFCM,在 FCM 目标函数中引入样本容量信息,如式(4)所示,使得目标函数对样本容量差异较大的情况也能够适应,但 GFCM 要求模糊加权指数 $m>1$。

$$J(U,P)=\sum_{j=1}^{n}\sum_{i=1}^{c}\frac{u_{ij}^m}{\sum_{s=1}^{n}u_{is}^m}\| x_j-c_i\|^2 \tag{4}$$

2.2 EFCM 最优化准则及公式推导

均衡 FCM 聚类算法(EFCM)在 GFCM 的基础上,取 $m=1$,则 EFCM 最小化目标函数为:

$$J(U,P)=\sum_{i=1}^{c}\sum_{j=1}^{n}\frac{u_{ij}}{\sum_{j=1}^{n}u_{ij}}\| x_j-c_i\|^2 \tag{5}$$

其中,$\sum_{i=1}^{c}u_{ij}=1,\forall j=1,\cdots,n$。

由 EFCM 目标函数及约束条件 $\sum_{i=1}^{c}u_{ij}=1,\forall j=1,\cdots,n$,利用拉格朗日乘子法可得:

$$L(U,P,\lambda)=\sum_{j=1}^{n}\sum_{i=1}^{c}\frac{u_{ij}}{\sum_{j=1}^{n}u_{ij}}\| x_j-c_i\|^2-\sum_{j=1}^{n}\lambda_j(\sum_{i=1}^{c}u_{ij}-1) \tag{6}$$

利用式(6)对 c_i 求偏导,并令其结果等于 0,从而得到 EFCM 聚类中心迭代公式:

$$c_i=\frac{\sum_{j=1}^{n}u_{ij}x_j}{\sum_{j=1}^{n}u_{ij}},\forall i=1,\cdots,c \tag{7}$$

2.3 EFCM 算法参数估计存在的问题及解决方法

在 FCM 及 GFCM 算法中,u_{ij} 和 c_i 是通过 AO 交替迭代算法得到估计值的。但在 EFCM 中,由于 $m=1$,若仿效 FCM 模糊隶属度 u_{ij} 的推导方法,利用 Lagrange 辅助函数式(6)对 u_{ij} 求导,只会得到式(8):

$$u_{ij}=\sum_{j=1}^{n}u_{ij}-\frac{n-1}{\sum_{i=1}^{c}\frac{(\sum_{j=1}^{n}u_{ij})^2}{\| x_j-c_i\|^2}}\frac{(\sum_{j=1}^{n}u_{ij})^2}{\| x_j-c_i\|^2} \tag{8}$$

$\forall i=1,\cdots,c\quad\forall j=1,\cdots,n$

式(8)无法判断方程对 u_{ij} 是否可解,也不能保证通过 AO 交替迭代收敛到稳定点,所以 EFCM 无法仿效 FCM 算法通过交替迭代实现对参数的估计。

引入粒子群算法寻求目标函数的最优化和参数的优解估计,粒子群算法模拟鸟群群体智能进行全局最优解搜索,现有基于粒子群的聚类算法[5,6]将聚类中心作为编码对象,模糊隶属度 u_{ij} 依然依靠梯度法迭代公式计算,无法解决 EFCM 算法 u_{ij} 估计所存在的问题,因此改进为以各样本隶属度为编码对象,聚类中心利用式(7)计算。粒子群算法简洁高效,为全局优化算法,对粒子维数没有约束,适合 EFCM 算法目标函数寻优和参数估计。

2.4 EFCM 算法迭代求解流程

粒子群适应度函数定义为:

$$f(x)=\frac{1}{J(U,P)+1} \tag{9}$$

EFCM 用下列步骤确定聚类中心 c_i 和隶属矩阵 $U=(u_{ij})_{c\times n}$:

步骤 1 令 $t=0$,初始化多个 $c\times n$ 维粒子的位置 $X_i(t)$ 和速度 $V_i(t)$,其中粒子位置的每一维是介于 0,1 间的随机数。

步骤 2 将粒子位置 $X_i(t)$ 的每 c 维分量构成一组进行单位化,单位化后的位置分量对应模糊隶属度 u_{ij},即满足约束条件 $\sum_{i=1}^{c}u_{ij}=1$。

步骤 3 由式(7)计算 c 个聚类中心 c_i,$i=1,\cdots,c$。

步骤 4 根据式(5)、式(9)计算多个粒子的适应度函数值。如果迭代次数超过某个限值,或群体最优解适应度相对上次群体最优解适应度的改变量小于某个阈值,则算法停止。

步骤 5 记录及更新个体最优解 $P_i(t)$ 和群体最优解 $P_g(t)$,进而更新粒子速度 $V_i(t+1)$ 及位置 $X_i(t+1)$,令 $t=t+1$,返回步骤 2。

最后,EFCM 将样本 x_j 的聚类类别判定为:$k=\arg\max_{i=1,\cdots,c}u_{ij}$。

3 EFCM 性质分析

3.1 EFCM 均衡性

若 EFCM 中模糊隶属度为硬性的,则有:

$$u_{ij}=\begin{cases}1, & x_j\in G_i\\ 0, & x_j\notin G_i\end{cases} \tag{10}$$

则 EFCM 最小化代价函数式(5)转换为改进 HCM 代价函数式(2),由于改进 HCM 代价函数考虑了各类样本容量对聚类的影响,因此 EFCM 也是包含了样本容量因素的。

当数据集为平衡数据集时,样本容量因素 $\sum_{j=1}^{n} u_{ij}$ 可近似为常量,由式(5)可知,EFCM 算法将退化为 FCM 算法,说明 EFCM 算法对平衡数据集适用;当数据集为不平衡数据集时,负类样本容量信息大于正类样本容量信息,则式(5)降低了负类对目标函数的贡献。为了使目标函数最小化,算法将向负类样本倾斜,从而校正了算法因距离因素造成的对负类样本的误判。

3.2 EFCM 与 FCM、GEFCM 的比较

通过 EFCM 和 FCM 的最小化代价函数式(5)和式(3)的比较可知,EFCM 包含了类样本容量因素,所以更符合实际情况;而 FCM 只考虑了距离因素对类别判决的贡献,适合于处理类样本容量平衡的情况,当各类样本容量差异较大时,容易导致误判。

GEFCM、EFCM 目标函数均衡化思想是一致的,都是在目标函数中引入样本的容量信息,使得算法对均衡或不均衡数据集都有效。GEFCM 利用梯度法获取模糊隶属度迭代公式,为了算法的收敛性,要求 $m>1$;EFCM 的目标函数雷同 GEFCM 目标函数当 $m=1$ 时的情况,EFCM 引入粒子群算法对模糊隶属度进行编码,克服了 $m>1$ 的限制。

3.3 EFCM 的收敛性

EFCM 的目标函数式(5)是恒大于等于 0 的,即是有下界的,EFCM 引入粒子群算法在全局空间中寻找最优解,粒子群算法及式(9)保证了模糊聚类目标函数值的单调递减性,从而说明迭代算法所求得的一系列目标函数值收敛于某个极小值。

4 仿真实验

为了验证 EFCM 算法的有效性,首先进行了平衡、不平衡数据集仿真实验,类别数 $K=2$ 已知。

第一个为不平衡数据集仿真实验:样本集由两个二维高斯随机分布样本的类别组成,两类的类中心分别为(5,5),(10,10),第一类的样本数为 300,协方差矩阵取为[6 0;0 6],第二类的样本数为 50,协方差矩阵取为[1 0;0 1]。两类样本方差的差异在于两类样本容量的不同,样本数多的类,则样本分布较广泛,方差体现了样本的扩散情况,按照样本比例给出方差取值,样本的平面分布如图 1 所示。

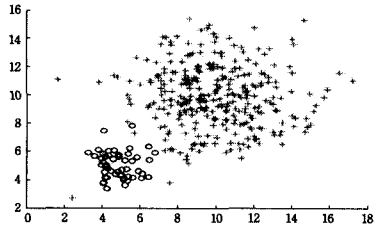

图 1 仿真实验 1 两类样本空间分布

采用 FCM 和 EFCM 对两类样本集进行聚类。FCM 模糊度控制参数取值为 2;当利用 PSO 求解 EFCM 时,粒子数取为 20,迭代 100 次,粒子位置的每维参数取值范围为[0.01,0.99],粒子位置的每维参数对应一个模糊隶属度,所以在计算适应度函数值时,需要先将每 c 维分量归一化。需要说明的是,直接利用 PSO 生成的随机初始粒子计算 EFCM 效果非常差,极易陷入局部优解,其结果毫无价值,所以先利用 FCM 算法对数据集进行聚类,选取 FCM 所训练出来的模糊隶属度,作为 PSO 的一个初始粒子。

由于两类样本容量差距较大,单纯考察聚类正确度并不能够反映各算法的有效性,因此选用各类正确度的几何平均 $g=\sqrt[n]{\prod_i acc_i}$ 作为比较标准,其中 acc_i 表示第 i 类正确度($i=1,2,n=2$),每个算法进行 10 次测试,取测试结果平均值作比较,测试结果如表 1 所列。

表 1 仿真实验 1 结果比对

聚类模型	各类正确度几何平均平均值
FCM(m=2)(%)	86.6
EFCM(%)	94.8

第二个为平衡数据集仿真实验,第一个仿真实验中两类的类中心分别为(5,5),(10,10),两类样本数都为 100,协方差矩阵取为[2 0;0 2],样本的平面分布如图 2 所示,实验运算结果如表 2 所列。

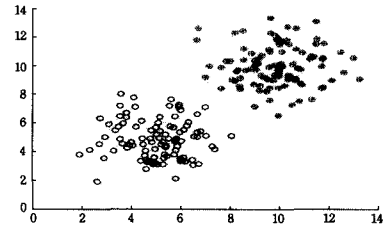

图 2 仿真实验 2 两类样本空间分布

表 2 仿真实验 2 结果比对

聚类模型	正确度几何平均平均值
FCM(m=2)(%)	99.0
EFCM(%)	98.5

根据平衡、不平衡数据集仿真实验结果对比可知,EFCM 相较于 FCM 而言,对样本容量的差异性具有更好的包容性,当样本容量差异较大时,EFCM 表现更为有效;同时,EFCM 在平衡及不平衡数据集中都有较好表现,这也说明了 EFCM 算法性能的稳定性。

EFCM 算法的求解是基于粒子群算法的,所以 EFCM 算法具有通用性。为了测试 EFCM 算法的通用性能,选择 UCI 机器学习数据库中的公共数据集进行算法比对测试,所选数据集包括 Pima、Wine 两个数据集[9],两个数据集的信息如表 3 所列,实验设置及测试标准与仿真实验 1、2 相同。

表 3 实验数据集

数据集	样本数	维数	类别比
Pima	768	8	500:268
Wine	178	13	59:71:48

实验结果如表 4 所列,实验比对标准与仿真实验 1、2 相同。

表 4 基于 Pima 及 Wine 数据集的对比测试

数据集	FCM(m=2)(%)	EFCM(%)
Pima	55.45	54.88
Wine	72.46	73.25

由表 4 可知,EFCM 相较于 FCM 而言,也有良好的表现,同时也说明了 EFCM 算法具有良好的数据集通用性能。

结束语 模糊聚类算法有着广泛的应用前景和重要的社

会价值[11]。FCM算法是聚类分析中应用最为广泛的一种聚类算法，但由于对类样本容量缺乏考虑，因此不适用于不平衡数据集。EFCM算法在目标函数中引入样本容量因素，使得算法对样本容量差异性保持鲁棒，利用PSO算法对模糊隶属度进行估计，避免了EFCM算法梯度法参数估计不能保证迭代过程收敛的弊病，理论分析及实验测试验证了所提算法的有效性。

参考文献

[1] Zhao F, Jiao L C, L H Q. Kernel generalized fuzzy c-means clustering with spatial information for image segmentation[J]. Digital Signal Processing: A Review Journal, 2013, 23(1): 184-199

[2] Kannan S R, Ramathilagam S, Chung P C. Effective fuzzy c-means clustering algorithms for data clustering problems[J]. Expert Systems with Applications, 2012, 39(7): 6292-6300

[3] 孙晓鹏，纪燕杰，李翠芳，等. 三维网格模型增量式聚类检索[J]. 计算机科学，2011，38(11)：248-251

[4] Bezdek J C, Hathaway R J, Sobin M, et al. Convergence and theory for fuzzy c-means clustering: counterexamples and repairs [J]. IEEE Transactions on Systems, Man and Cybernetics, 1987, 17(5): 873-877

[5] Zhang Z H, Liu S, Ji C P. An improved fuzzy C-means based on IPSO[J]. International Review on Computers and Software, 2012, 7(1): 241-245

[6] Chen D H, Liu Z J, Wang Z H. A novel fuzzy clustering algorithm based on kernel method and particle swarm optimization [J]. Journal of Convergence Information Technology, 2012, 7(3): 299-307

[7] Huang M, Xia Z X, Wang H B. The range of the value for the fuzzifier of the fuzzy c-means algorithm[J]. Pattern Recognition Letters, 2012, 33(16): 2280-2284

[8] Wu K L. Analysis of parameter selections for fuzzy c-means [J]. Pattern Recognition, 2012, 45(1): 407-415

[9] 王熙照，崔芳芳，鲁淑霞. 密度加权近似支持向量机[J]. 计算机科学，2012，39(1)：182-184

[10] 文传军，詹永照，柯佳. 广义均衡模糊C均值聚类算法[J]. 系统工程理论与实践，2012，32(12)：2751-2755

[11] Yu Y, Zhang B B, Chen L. An improved fuzzy C-means cluster algorithm for radar data association[J]. International Journal of Advancements in Computing Technology, 2012, 4(20): 181-189

(上接第 232 页)

根据优先值对混淆集进行排序，可提高文本校对的效率[14]。汉字的字频信息在一定程度上反映了被写错的概率，一般情况下，高频字被错写成低频字的概率比低频字错写成高频字的概率要大；形相似度越高，写错的概率越大。设$\{v_1, v_2 \cdots v_m\}$为得到的错别字混淆集合，用 N_i 表示混淆集合$\{v_1, v_2, \cdots, v_m\}$中 $v_i(1 \leqslant i \leqslant m)$与正确汉字的形相似度，按以下步骤进行排序：

如果混淆集中某一候选串 v_i 中 $N_i > \lambda$(λ 为设定的阈值，经验值为 0.68)，根据 N_i 的大小进行排序。对剩余的混淆集集合，根据字频的大小进行排序。

5.3 实验结果分析

本实验系统中，实验结果的好坏与一些因素有关，如：

(1)分词的准确度

由于本实验通过大规模语料获取混淆集是在分词的基础上进行的，因此分词的准确度对实验结果有影响，但是到目前为止还没有一种分词方法可以达到100%的分词准确度。另外，由于中文用法的繁杂，语料选取的领域广泛，故在分词过程中会碰到词典的未登录词，这对分词及最终结果有影响。

(2)召回率和正确率的权衡

在本实验系统中所选取的阈值是一个经验值。它取值的高低直接影响到召回率和正确率。阈值选取过高会导致召回率下降，过低则会导致正确率下降，因此这存在一个权衡的过程。

(3)内部规则的不完备性

完整的规则能对错别字混淆集进行很大程度的补充和完善，由于目前对错别字混淆集补充还没有一个统一的规则，本课题所罗列的规则难免有一定的局限性，还应该继续深入考虑挖掘，使其更加完备。

结束语 本文对种子汉字的错别字混淆集进行了规则自补充、数据开源外部补充，得到一个种子汉字的错别字字典，同时，大规模语料获取的错别词词典不受领域的限制，应用的领域广泛，对文本校对有很大的帮助。今后将从内部规则和语法分析入手，进行更深入的研究，以提高系统的完善性。

参考文献

[1] 刘亮亮，王石，王东升，等. 领域问答系统中的文本错误自动发现方法[J]. 中文信息学报，2013，3：77

[2] 张磊，周明，黄昌宁，等. 中文文本自动校对[J]. 语言文字应用，2001(1)：19

[3] 陈笑蓉，秦进，汪维家，等. 中文文本校对技术的研究与实现[J]. 计算机科学，2003，11(16)：53

[4] Zhang Zhao-huang. A Pilot Study on Automatic Chinese Spelling Error Correction[J]. Communication of COLIPS, 1994, 4(2): 143

[5] 于勐，姚天顺. 一种混合的中文文本校对方法[J]. 中文信息学报，1998，12(2)：31

[6] 丰强泽，曹存根. 语音查询中的辨音方法：中国，CN1514387[P]. 2004-07-21

[7] 戴耿毅，余静涛. 基于双数组 Trie 树算法的字典改进和实现[J]. 软件导刊，2012，11(7)：17

[8] 李慧，杨炳儒，潘丽芳，等. 一种基于双数组 Trie 的 B2B 规则串提取方法[J]. 计算机科学，2013，40(5)：206

[9] 王静帆，邬晓钧，夏云庆，等. 中文信息检索系统的模糊匹配算法的研究和实现[J]. 中文信息学报，2007，21(006)：59

[10] 张仰森，曹元大，俞士汶. 基于规则与统计相结合的中文文本自动查错模型与算法[J]. 中文信息学报，2006，20(4)：1

[11] 张仰森，丁冰青. 中文文本自动校对技术现状及展望[J]. 中文信息学报，1998，12(3)：50

[12] 王贤明，胡智文，谷琼. 一种基于随机 n-Grams 的文本相似度计算方法[J]. 情报学报，2013，32(7)：716

[13] 吴春颖，王士同. 基于二元语法的 N-最大概率中文粗分模型[J]. 计算机应用，2007(12)：2902

[14] 张仰森. 中文校对系统中纠错知识库的构造及纠错建议的产生算法[J]. 中文信息学报，2000，15(5)：33