(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106056167 A
(43)申请公布日 2016.10.26

(21)申请号 201610512937.6

(22)申请日 2016.07.01

(71)申请人 山东大学
地址 250199 山东省济南市历城区山大南路27号

(72)发明人 江铭炎 郭宝峰 孙舒琬 陈蓓蓓

(74)专利代理机构 济南金迪知识产权代理有限公司 37219
代理人 杨树云

(51)Int.Cl.
G06K 9/62(2006.01)
G06N 3/00(2006.01)

权利要求书4页 说明书7页 附图1页

(54)发明名称

一种基于高斯核混合人工蜂群算法的归一化可能性模糊熵聚类方法

(57)摘要

本发明涉及一种基于高斯核混合人工蜂群算法的归一化可能性模糊熵聚类方法，包括：(1)归一化预处理，得到新样本集X_New；(2)参数初始化：(3)计算到初始聚类中心的距离，计算隶属矩阵U和可能性矩阵T，得到初始适应度值fitness(i)；(4)进入采蜜蜂阶段：(5)进入跟随蜂阶段：(6)进入侦察蜂阶段：(7)得到最终最优聚类中心Vbest，并由Vbest得到对应的隶属矩阵U，并按照$c_i=\mathrm{argmax}(u_{ij})$得到最终聚类。本发明提出的方法具有较好的噪声鲁棒性，在一定程度上较少参数的人为依赖性，引入人工蜂群算法后，算法的全局特性得到提高，避免了参数初始值敏感问题。本发明的可行性和有效性都得提高。

CN 106056167 A

1.一种基于高斯核混合人工蜂群算法的归一化可能性模糊熵聚类方法,其特征在于,包括以下步骤:

(1)对样本集X进行归一化预处理,得到新样本集X_New;

(2)参数初始化:人工蜂群算法种群数NP,采蜜蜂的数量SN,局部最优限制次数limit,最大迭代次数maxcycle;初始聚类中心V,模糊加权指数m,聚类数c,阈值ε,协方差矩阵σ^2,熵系数λ,高斯核函数的宽度参数δ;

(3)计算新样本集X_New中的样本到初始聚类中心的距离,并计算对应的隶属矩阵U和可能性矩阵T,得到每只采蜜蜂的初始适应度值fitness(i);

(4)进入采蜜蜂阶段:采蜜蜂进行邻域搜索,产生每只采蜜蜂的适应度值新解fitness(sol),并更新隶属矩阵U和可能性矩阵T;

(5)比较fitness(i)和fitness(sol),如果fitness(i)<fitness(sol),则fitness(i)=fitness(sol),否则,fitness(i)不变;

(6)进入跟随蜂阶段:跟随蜂按概率p_i选择跟踪采蜜蜂,并对采蜜蜂进行邻域搜索,产生每个采蜜蜂的适应度值新解fitness(sol1),并更新隶属矩阵U和可能性矩阵T,比较fitness(i)和fitness(sol1),如果fitness(i)<fitness(sol1),则fitness(i)=fitness(sol1),否则,fitness(i)不变;

(7)进入侦察蜂阶段:判断采蜜蜂转侦察蜂的条件是否满足,如果采蜜蜂的适应度值fitness(i)在limit次迭代中均未发生变化,则认为该采蜜蜂的适应度值fitness(i)为局部最优解,放弃局部最优解,同时该采蜜蜂转变为侦查蜂,按照

V_i=rand(c,s).*(max(X_New)-min(X_New))+min(X_New)在解空间进行新的搜索;否则,该采蜜蜂不转变;s为新样本集X_New中每个样本元素的维数,max(X_New)为行向量,由新样本集X_New中每列的最大值组成,min(X_New)也为行向量,由新样本集X_New中每列最小值组成,rand(c,s)为由[0,1]构成的c*s矩阵,*表示矩阵之间对应元素相乘;

(8)重复步骤(3)至(7),直到最大迭代次数maxcycle或者满足||Vbest(iter+1)-Vbest(iter)||<ε;||Vbest(iter+1)-Vbest(iter)||表示第iter+1次迭代得到最优聚类中心和第iter次迭代得到最优聚类中心的欧式距离;

(9)得到最终最优聚类中心Vbest,并由Vbest得到对应的隶属矩阵U,并按照c_i=argmax(u_{ij})得到最终聚类。

2.根据权利要求1所述的一种基于高斯核混合人工蜂群算法的归一化可能性模糊熵聚类方法,其特征在于,所述步骤(1)中,样本集X包含n个样本的数据,X={x_1,x_2,...,x_n},x_j=(x_{j1},x_{j2},...,x_{js})^T∈R^s,x_j的样本元素为实数域R的S维空间中的样本,1≤j≤n,具体步骤包括:对样本集X进行归一化预处理,归一化预处理公式如式(I)所示:

$$x_{jk}'=\frac{x_{jk}-\overline{x_k}}{(x_k)_{\max}-(x_k)_{\min}}\quad(\mathrm{I})$$

式(I)中,x_{jk}'为新样本集X_New中的元素,j=1,2,...,n,k=1,2,...,s,$(x_k)_{max}$、$(x_k)_{min}$分别为样本集X第k维属性上的最大值、最小值,$\overline{x_k}$为样本集X第k维属性上的均值,

$\overline{x_k}=\frac{1}{n}\sum_{j=1}^{n}x_{jk}$，归一化处理后，得到新样本集X_New。

3.根据权利要求2所述的一种基于高斯核混合人工蜂群算法的归一化可能性模糊熵聚类方法，其特征在于，所述步骤(2)中，具体步骤包括：

A、初始化SN个初始聚类中心：V_1=rand(c,s).*(max(X_New)-min(X_New))+min(X_New)；

B、计算目标函数中的协方差矩阵σ^2，计算公式如式(Ⅱ)所示：

$$\sigma^2=\frac{1}{n}\sum_{j=1}^{n}\|\varphi(x_j)-\varphi(\bar{x})\|^2 \quad (Ⅱ)$$

式(Ⅱ)中，$\varphi(\bar{x})$为映射到高维特征空间后的样本均值：$\varphi(\bar{x})=\frac{1}{n}\sum_{j=1}^{n}\varphi(x_j)$；将$\varphi(\bar{x})=\frac{1}{n}\sum_{j=1}^{n}\varphi(x_j)$代入式(Ⅱ)中，消去$\varphi(\bar{x})$，得到式(Ⅲ)：

$$\sigma^2=\frac{1}{n}\sum_{j=1}^{n}k(x_j,x_j)-\frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}k(x_j,x_k) \quad (Ⅲ)$$

式(Ⅲ)中，核函数Φ(x)采用高斯核函数$k(x,y)=exp(-\frac{\|x-y\|^2}{2\delta^2})$，$\delta=0.4$。

4.根据权利要求3所述的一种基于高斯核混合人工蜂群算法的归一化可能性模糊熵聚类方法，其特征在于，所述步骤(3)中，具体步骤包括：

C、映射到高维特征空间后，计算新样本集X_New的中样本到初始聚类中心V的欧式距离；

D、依据式(Ⅳ)计算隶属矩阵U：

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}(\frac{D_{ij}}{D_{kj}})^{m-1}},\forall i,j \quad (Ⅳ)$$

式(Ⅳ)中，$D_{ij}=\|\varphi(x_j)-\varphi(v_i)\|^2=k(x_j,x_j)-2k(x_j,v_i)+k(v_i,v_i)$，为高维特征空间中新样本集X_New中样本j与初始聚类中心V的v_i之间的欧式距离的平方；D_{kj}为高维特征空间中新样本集X_New中样本j与初始聚类中心V的v_k之间的欧式距离的平方；

E、依据式(V)计算可能性矩阵T：

$$t_{ij}=exp(-\frac{m^2c(D_{ij}+\lambda)}{\sigma^2+m^2c\lambda}),\forall i,j \quad (Ⅴ)$$

式(V)中，m为模糊加权指数，λ为熵系数，0.01≤λ≤1；

F、每个采蜜蜂的初始适应度值fitness(i)的求取公式如式(Ⅵ)所示：

$$fitness(i)=\begin{cases}\frac{1}{1+fobj(i)}, & fobj(i)\geq 0\\ 1+|fobj(i)|, & fobj(i)<0\end{cases} \quad (Ⅵ)$$

式(Ⅵ)中，j=1,2,...,s,fobj(i)是指目标函数J(U,V,T)的极小值。

5.根据权利要求4所述的一种基于高斯核混合人工蜂群算法的归一化可能性模糊熵聚类方法，其特征在于，所述步骤(4)中，具体步骤包括：

G、采蜜蜂邻域搜索产生适应度值新解fitness(sol)，邻域搜索公式如式(Ⅶ)所示：

$v_{ij}=x_{ij}+rand(x_{ij}-x_{kj})+\beta(xbest-x_{ij})$ (Ⅶ)

式(Ⅶ)中，xbest代表已搜索到最优的聚类中心，i＝1,2,...,SN，x_{ij}表示第i个采蜜蜂的第j维分量，v_{ij}为搜索到的邻域值，rand、β为搜索系数，rand、β的值均为(0,1)；

H、依据式(Ⅳ)、式(Ⅴ)计算步骤G求取的适应度值新解fitness(sol)对应的隶属矩阵U1和可能性矩阵T1。

6.根据权利要求5所述的一种基于高斯核混合人工蜂群算法的归一化可能性模糊熵聚类方法，其特征在于，所述步骤(5)中，具体步骤包括：

比较fitness(i)及fitness(sol)，按照贪婪准则更新当前解：如果fitness(sol)≥fitness(i)，fitness(i)＝fitness(sol)；否则，舍弃fitness(sol)，继续保留fitness(i)。

7.根据权利要求6所述的一种基于高斯核混合人工蜂群算法的归一化可能性模糊熵聚类方法，其特征在于，所述步骤(6)中，具体方法为：

I、跟随蜂以概率p_i选择跟随适应度值较优的采蜜蜂并在其周围进行进一步邻域搜索：概率p_i的求取公式如式(Ⅷ)所示：

$$p_i=\frac{fitness(i)}{\sum fitness(i)} \quad (Ⅷ)$$

如果p_i>rand，就选择跟随该采蜜蜂，rand为(0,1)之间的实数，进入步骤J；否则，就不跟随该采蜜蜂；

J、跟随蜂邻域搜索产生适应度值新解fitness(sol1)，邻域搜索公式如式(Ⅶ)所示；

K、依据式(Ⅳ)、式(Ⅴ)计算步骤J求取的适应度值新解fitness(sol1)对应的隶属矩阵U2和可能性矩阵T2；

L、按照贪婪准则比较fitness(i)及fitness(sol1)，更新当前解：如果fitness(sol1)≥fitness(i)，接受fitness(sol1)，fitness(i)＝fitness(sol1)；否则，舍弃fitness(sol1)，继续保留fitness(i)。

8.根据权利要求7所述的一种基于高斯核混合人工蜂群算法的归一化可能性模糊熵聚类方法，其特征在于，所述步骤(9)，具体步骤包括：

①计算每只采蜜蜂的适应度值fitness(i)和对应的目标函数值fobj(i)，i＝1,2,...,SN，选取fobj(i)中值最小的采蜜蜂对应的聚类中心作为最优的聚类中心Vbest；

②循环迭代得到最终的最优聚类中心Vbest，根据式(Ⅸ)、式(Ⅹ)得到最终的隶属矩阵U；

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}(\frac{D_{ij}}{D_{kj}})^{m-1}},\forall i,j \quad (Ⅸ)$$

$$t_{ij}=exp(-\frac{m^2c(D_{ij}+\lambda)}{\sigma^2+m^2c\lambda}),\forall i,j \quad (Ⅹ)$$

式(Ⅸ)中，$D_{ij}=\|\varphi(x_j)-\varphi(v_i)\|^2=k(x_j,x_j)-2k(x_j,v_i)+k(v_i,v_i)$，为高维特征空间中新样

本集X_New中样本j与初始聚类中心V的v_i之间的欧式距离的平方；D_{kj}为高维特征空间中新样本集X_New中样本j与初始聚类中心V的v_k之间的欧式距离的平方；式(Ⅹ)中，m为模糊加权指数，λ为熵系数，$0.01 \leqslant \lambda \leqslant 1$；

③按照式(XI)求取样本u_{ij}所属的类别c_i：

$$c_i = \mathrm{argmax}(u_{ij}) \quad (XI)$$。

一种基于高斯核混合人工蜂群算法的归一化可能性模糊熵聚类方法

技术领域

[0001] 本发明涉及一种基于高斯核混合人工蜂群算法的归一化可能性模糊熵聚类方法，属于大数据挖掘、机器学习技术领域。

背景技术

[0002] 模糊聚类分析作为无监督分析数据、理解数据、认知事物的重要手段，由于引入模糊集合和模糊数学的思想，通过隶属度函数建立了样本数据与类别之间的不确定性描述，有效地解决了现实中不精确、没有明显边界“亦此亦彼”的聚类问题。模糊聚类拥有较好的数据表达能力与聚类效果，现已成功应用于海量数据实时聚类分析、模式分类、风险趋势预测、决策分析中，为人们深入理解数据、深层利用数据、挖掘数据中潜在价值信息做出重要共献。

[0003] 现阶段理论研究和实际应用中比较广泛的是基于目标函数的模糊聚类，包括模糊C均值聚类(Fuzzy C-means Clustering,FCM)、可能性C均值聚类(Possibilistic C-means Clustering,PCM)、可能性模糊C均值聚类(Possibilistic Fuzzy C-means CLustering,PFCM)。FCM算法对初始聚比较敏感，并且容易陷入局部最优解而得不到最佳的聚类划分；PCM克服了FCM对噪声敏感的问题，对噪声鲁棒性有所提高，但容易引起一致性聚类问题；PFCM兼具FCM与PCM的优点，具有较好的噪声鲁棒性，又不会产生重合的聚类，但PFCM涉及的参数较多，通常这些参数都需要人为指定而缺乏理论依据，这无形中增加了聚类的计算复杂度，同时算法的稳定性也受到影响。此外，这些基于目标函数的模糊聚类算法适合处理线性可分的、低维、凸型结构数据，然而在聚类算法在处理高维、非线性可分、非凸结构数据时聚类算法的性能很不稳定。

发明内容

[0004] 针对现有技术的不足，本发明提供了一种基于高斯核混合人工蜂群算法的归一化可能性模糊熵聚类方法(Hybrid Methods for Possibilistic Fuzzy Entropy Clustering Based on Artificial Bee Colony Algorithm and kernel function,ABC_KPFECM)；

[0005] 本发明通过对原始样本数据归一化处理，解决了量纲不统一对聚类结果产生的影响；此外，本发明引入高斯核函数，将原样本空间的数据映射到高维特征空间，解决了高维、非凸、非线性可分结构数据聚类不稳定的问题；最后，本发明还引入具有独特全局寻优能力的人工蜂群算法，优化提高了算法的全局寻优特性。该方法具有较好的噪声鲁棒性，不会产生一致性聚类问题，也避免了参数的人为依赖性，同时具有较好的全局特性，算法的整体性能得到提高。

[0006] 本发明的数学模型为：

[0007]

$$J(U,T,V)=\sum_{i=1}^{c}\sum_{j=1}^{n}(u_{ij}^{m}+t_{ij})\|\varphi(x_j)-\varphi(v_i)\|^2+\frac{\sigma^2}{m^2c}\sum_{i=1}^{c}\sum_{j=1}^{n}(t_{ij}\log t_{ij}-t_{ij})+\lambda\sum_{i=1}^{c}\sum_{j=1}^{n}t_{ij}\log t_{ij}$$

[0008] 其中$\sigma^2=\frac{1}{n}\sum_{j=1}^{n}\|\varphi(x_j)-\varphi(\bar{x})\|^2$，$\varphi(\bar{x})=\frac{1}{n}\sum_{j=1}^{n}\varphi(x_j)$

[0009] 公式中φ为数据空间X到高维特征空间H的映射，即$\varphi:x\to\varphi(x)\in H$，与之对应的核函数为$k(x,y)=\varphi(x)*\varphi(y)$，$\varphi(x_j)$为映射到高维特征空间的样本，$\varphi(v_i)$为映射到高维特征空间的聚类中心；参数满足m>1，λ>0，在满足约束条件$\sum_{i=1}^{c}u_{ij}=1,\forall i,j$，以及0≤$u_{ij}$，$t_{ij}$≤1，令$\|\varphi(x_j)-\varphi(v_i)\|^2=k(x_j,x_j)-2k(x_j,v_i)+k(v_i,v_i)=D_{ij}$，目标函数J(U,V,T)取极小值时必须满足：

[0010]
$$u_{ij}=\frac{1}{\sum_{k=1}^{c}(\frac{D_{ij}}{D_{kj}})^{m-1}},\forall i,j$$

[0011]
$$t_{ij}=exp(-\frac{m^2c(D_{ij}+\lambda)}{\sigma^2+m^2c\lambda}),\forall i,j$$

[0012]
$$v_i=\frac{\sum_{j=1}^{N}(u_{ij}^{m}+t_{ij})k(x_j,v_i)x_j}{\sum_{j=1}^{N}(u_{ij}^{m}+t_{ij})k(x_j,v_i)}$$

[0013] 本发明的技术方案为：

[0014] 一种基于高斯核混合人工蜂群算法的归一化可能性模糊熵聚类方法，包括以下步骤：

[0015] (1)对样本集X进行归一化预处理，得到新样本集X_New；

[0016] (2)参数初始化：人工蜂群算法种群数NP，采蜜蜂的数量SN，局部最优限制次数limit，最大迭代次数maxcycle；初始聚类中心V，模糊加权指数m，聚类数c，阈值ε，协方差矩阵σ^2，熵系数λ，高斯核函数的宽度参数δ；

[0017] (3)计算新样本集X_New中的样本到初始聚类中心的距离，并计算对应的隶属矩阵U和可能性矩阵T，得到每只采蜜蜂的初始适应度值fitness(i)；

[0018] (4)进入采蜜蜂阶段：采蜜蜂进行邻域搜索，产生每只采蜜蜂的适应度值新解fitness(sol)，并更新隶属矩阵U和可能性矩阵T；

[0019] (5)比较fitness(i)和fitness(sol)，如果fitness(i)<fitness(sol)，则fitness(i)=fitness(sol)，否则，fitness(i)不变；

[0020] (6)进入跟随蜂阶段：跟随蜂按概率p_i选择跟踪采蜜蜂，并对采蜜蜂进行邻域搜索，产生每个采蜜蜂的适应度值新解fitness(sol1)，并更新隶属矩阵U和可能性矩阵T，比较fitness(i)和fitness(sol1)，如果fitness(i)<fitness(sol1)，则fitness(i)=fitness(sol1)，否则，fitness(i)不变；

[0021] (7)进入侦察蜂阶段：判断采蜜蜂转侦察蜂的条件是否满足，如果采蜜蜂的适应度

值fitness(i)在limit次迭代中均未发生变化，则认为该采蜜蜂的适应度值fitness(i)为局部最优解，放弃局部最优解，同时该采蜜蜂转变为侦查蜂，按照V_i=rand(c,s).*(max(X_New)-min(X_New))+min(X_New)在解空间进行新的搜索；否则，该采蜜蜂不转变；s为新样本集X_New中每个样本元素的维数，max(X_New)为行向量，由新样本集X_New中每列的最大值组成，min(X_New)也为行向量，由新样本集X_New中每列最小值组成，rand(c,s)为由[0,1]构成的c*s矩阵，*表示矩阵之间对应元素相乘；

[0022] (8)重复步骤(3)至(7)，直到最大迭代次数maxcycle或者满足||Vbest(iter+1)-Vbest(iter)||<ε；||Vbest(iter+1)-Vbest(iter)||表示第iter+1次迭代得到最优聚类中心和第iter次迭代得到最优聚类中心的欧式距离；

[0023] (9)得到最终最优聚类中心Vbest，并由Vbest得到对应的隶属矩阵U，并按照c_i=argmax(u_{ij})得到最终聚类。

[0024] 根据本发明优选的，所述步骤(1)中，样本集X包含n个样本的数据，X={x_1,x_2,...,x_n}，x_j=(x_{j1},x_{j2},...,x_{js})^T∈R^s，x_j的样本元素为实数域R的S维空间中的样本，1≤j≤n，具体步骤包括：

[0025] 为避免样本数据x_j中每个维度由于量纲不同对聚类结果造成的影响，先对样本集X进行归一化预处理。

[0026] 对样本集X进行归一化预处理，归一化预处理公式如式(Ⅰ)所示：

[0027] $$x_{jk}'=\frac{x_{jk}-\overline{x_k}}{(x_k)_{\max}-(x_k)_{\min}}\quad(\text{Ⅰ})$$

[0028] 式(Ⅰ)中，x_{jk}'为新样本集X_New中的元素，j=1,2,...,n,k=1,2,...,s，$(x_k)_{max}$、$(x_k)_{min}$分别为样本集X第k维属性上的最大值、最小值，$\overline{x_k}$为样本集X第k维属性上的均值，$\overline{x_k}=\frac{1}{n}\sum_{j=1}^{n}x_{jk}$，归一化处理后，得到新样本集X_New。

[0029] 根据本发明优选的，所述步骤(2)中，具体步骤包括：

[0030] A、初始化SN个初始聚类中心：V_1=rand(c,s).*(max(X_New)-min(X_New))+min(X_New)，l=1,2,...,SN，s为新样本集X_New中每个样本元素的维数，新样本集X_New是一个n*s维矩阵，max(X_New)为行向量，由新样本集X_New中每列的最大值组成，min(X_New)也为行向量，由新样本集X_New中每列最小值组成，rand(c,s)为由[0,1]构成的c*s矩阵，*表示矩阵之间对应元素相乘；

[0031] B、计算目标函数中的协方差矩阵σ^2，计算公式如式(Ⅱ)所示：

[0032] $$\sigma^2=\frac{1}{n}\sum_{j=1}^{n}\|\varphi(x_j)-\varphi(\overline{x})\|^2\quad(\text{Ⅱ})$$

[0033] 式(Ⅱ)中，$\varphi(\overline{x})$为映射到高维特征空间后的样本均值：$\varphi(\overline{x})=\frac{1}{n}\sum_{j=1}^{n}\varphi(x_j)$；将$\varphi(\overline{x})=\frac{1}{n}\sum_{j=1}^{n}\varphi(x_j)$代入式(Ⅱ)中，消去$\varphi(\overline{x})$，得到式(Ⅲ)：

[0034] $$\sigma^2=\frac{1}{n}\sum_{j=1}^{n}k(x_j,x_j)-\frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}k(x_j,x_k)\quad(\text{Ⅲ})$$

[0035] 式(Ⅲ)中,核函数Φ(x)采用高斯核函数$k(x,y)=exp(-\frac{\|x-y\|^2}{2\delta^2})$,δ=0.4。

[0036] 根据本发明优选的,所述步骤(3)中,具体步骤包括:

[0037] C、映射到高维特征空间后,计算新样本集X_New的中样本到初始聚类中心V的欧式距离;

[0038] D、依据式(Ⅳ)计算隶属矩阵U:

[0039] $$u_{ij}=\frac{1}{\sum_{k=1}^{c}(\frac{D_{ij}}{D_{kj}})^{m-1}},\forall i,j\quad(\text{Ⅳ})$$

[0040] 式(Ⅳ)中,$D_{ij}=\|\varphi(x_j)-\varphi(v_i)\|^2=k(x_j,x_j)-2k(x_j,v_i)+k(v_i,v_i)$,为高维特征空间中新样本集X_New中样本j与初始聚类中心V的v_i之间的欧式距离的平方;D_{kj}为高维特征空间中新样本集X_New中样本j与初始聚类中心V的v_k之间的欧式距离的平方;

[0041] E、依据式(Ⅴ)计算可能性矩阵T:

[0042] $$t_{ij}=exp(-\frac{m^2c(D_{ij}+\lambda)}{\sigma^2+m^2c\lambda}),\forall i,j\quad(\text{Ⅴ})$$

[0043] 式(Ⅴ)中,m为模糊加权指数,λ为熵系数,0.01≤λ≤1;

[0044] F、目标函数J(U,V,T)的极小值fobj(i)对应着最好的聚类划分,人工蜂群算的蜜源位置对应着可行解(聚类中心),可行解的优劣取决于适应度函数fitness(i),每个采蜜蜂的初始适应度值fitness(i)的求取公式如式(Ⅵ)所示:

[0045] $$fitness(i)=\begin{cases}\frac{1}{1+fobj(i)}, fobj(i)\geq 0\\ 1+|fobj(i)|, fobj(i)<0\end{cases}\quad(\text{Ⅵ})$$

[0046] 式(Ⅵ)中,j=1,2,...,s,fobj(i)是指目标函数J(U,V,T)的极小值;

[0047] 根据本发明优选的,所述步骤(4)中,具体步骤包括:

[0048] G、采蜜蜂邻域搜索产生适应度值新解fitness(sol),引入具有记忆全局最优的xbest来提高搜索的效率和全局最优趋势,邻域搜索公式如式(Ⅶ)所示:

[0049] $v_{ij}=x_{ij}+rand(x_{ij}-x_{kj})+\beta(xbest-x_{ij})$ (Ⅶ)

[0050] 式(Ⅶ)中,xbest代表已搜索到最优的聚类中心,i=1,2,...,SN,x_{ij}表示第i个采蜜蜂的第j维分量,v_{ij}为搜索到的邻域值,rand、β为搜索系数,rand、β的值均为(0,1);

[0051] H、计算步骤G求取的适应度值新解fitness(sol)(聚类中心)对应的隶属矩阵U1和可能性矩阵T1。

[0052] 根据本发明优选的,所述步骤(5)中,具体步骤包括:

[0053] 比较fitness(i)及fitness(sol),按照贪婪准则更新当前解:如果fitness(sol)≥fitness(i),接受fitness(sol),fitness(i)=fitness(sol);否则,舍弃fitness(sol),继续保留fitness(i)。

[0054] 根据本发明优选的,所述步骤(6)中,具体方法为:

[0055] Ⅰ、跟随蜂以概率p_i选择跟随适应度值较优的采蜜蜂并在其周围进行进一步邻域搜索：概率p_i的求取公式如式(Ⅷ)所示：

[0056] $$p_i=\frac{fitness(i)}{\sum fitness(i)} \quad (Ⅷ)$$

[0057] 如果p_i>rand，就选择跟随该采蜜蜂，rand为(0,1)之间的实数，进入步骤J；否则，就不跟随该采蜜蜂；

[0058] J、跟随蜂邻域搜索产生适应度值新解fitness(sol1)，邻域搜索公式如式(Ⅶ)所示；

[0059] K、计算步骤J求取的适应度值新解fitness(sol1)对应的隶属矩阵U2和可能性矩阵T2；

[0060] L、按照贪婪准则比较fitness(i)及fitness(sol1)，更新当前解：如果fitness(sol1)≥fitness(i)，接受fitness(sol1)，fitness(i)＝fitness(sol1)；否则，舍弃fitness fitness(sol1)，继续保留fitness(i)。

[0061] 根据本发明优选的，所述步骤(9)，具体步骤包括：

[0062] ①计算每只采蜜蜂的适应度值fitness(i)和对应的目标函数值fobj(i)，i＝1，2，...，SN，选取fobj(i)中值最小的采蜜蜂对应的聚类中心作为最优的聚类中心Vbest；

[0063] ②循环迭代得到最终的最优聚类中心Vbest，根据式(Ⅸ)、式(Ⅹ)得到最终的隶属矩阵U；

[0064] $$u_{ij}=\frac{1}{\sum_{k=1}^{c}(\frac{D_{ij}}{D_{kj}})^{m-1}},\forall i,j \quad (Ⅸ)$$

[0065] $$t_{ij}=exp(-\frac{m^2c(D_{ij}+\lambda)}{\sigma^2+m^2c\lambda}),\forall i,j \quad (Ⅹ)$$

[0066] 式(Ⅸ)中，$D_{ij}=\|\varphi(x_j)-\varphi(v_i)\|^2=k(x_j,x_j)-2k(x_j,v_i)+k(v_i,v_i)$，为高维特征空间中新样本集X_New中样本j与初始聚类中心V的v_i之间的欧式距离的平方；D_{kj}为高维特征空间中新样本集X_New中样本j与初始聚类中心V的v_k之间的欧式距离的平方；式(Ⅹ)中，m为模糊加权指数，λ为熵系数，0.01≤λ≤1；

[0067] ③按照式(Ⅺ)求取样本u_{ij}所属的类别c_i：

[0068] c_i＝argmax(u_{ij}) (Ⅺ)。

[0069] 本发明的有益效果为：

[0070] 1、本发明提出了一种基于高斯核混合人工蜂群算法的归一化可能性模糊熵聚类方法，通过对原始样本数据归一化处理，解决了多维大数据聚类分析时量纲不统一对聚类结果产生的影响。

[0071] 2、本发明提出的方法具有较好的噪声鲁棒性，不会产生一致性聚类问题，也在一定程度上较少参数的人为依赖性，引入人工蜂群算法后，算法的全局特性得到提高，避免了参数初始值敏感问题。

[0072] 3、本发明引入高斯核，进一步提高了算法在处理高维，非线性可分，非凸结构数据时的聚类性能，使得算法的可行性和有效性都得提高。

附图说明

[0073] 图1为本发明方法的流程图。

具体实施方式

[0074] 下面结合实施例和说明书附图对本发明作进一步限定，但不限于此。

[0075] 实施例

[0076] 本实施例结合机器学习标准测试集wine数据对本发明作进一步说明。

[0077] Wine数据是一个包含178个数据样本的13维数据集，包含3个类别。

[0078] 如图1所示，基于高斯核混合人工蜂群算法的归一化可能性模糊熵聚类方法，流程图如图1所示，包括以下步骤：

[0079] (1)输入待聚类样本wine数据，并对其进行归一化预处理，得到新样本X_New，使得新样本X_New落在区间[0,1]，避免由于量纲不同对聚类结果造成的影响。

[0080] (2)参数初始化，人工蜂群算法的种群数NP＝50，采蜜蜂数SN＝25，局部最优限制次数limit＝50，最大迭代次数maxcycle＝500；聚类数c＝3，初始聚类中心V＝rand(C,S).*(max(X)-min(X))+min(X)，模糊加权指数m＝2，阈值ε＝0.000001，协方差矩阵σ^2，熵系数λ＝0.4，高斯核函数的宽度参数δ＝0.4；

[0081] (3)计算新样本集X_New中样本到初始聚类中心V的欧式距离，并计算相应的隶属矩阵U和可能性矩阵T，得到每只采蜜蜂的初始适应度值。

[0082] (4)进入采蜜蜂阶段，采蜜蜂邻域搜索，产生每只采蜜蜂的适应度值新解fitness(sol)，并更新隶属矩阵U和可能性矩阵T；比较fitness(i)和fitness(sol)，如果fitness(i)<fitness(sol)，则fitness(i)＝fitness(sol)，否则，fitness(i)不变；

[0083] (5)进入跟随蜂阶段：跟随蜂按概率p_i选择跟踪采蜜蜂，并对采蜜蜂进行邻域搜索，产生每个采蜜蜂的适应度值新解fitness(sol1)，并更新隶属矩阵U和可能性矩阵T，比较fitness(i)和fitness(sol1)，如果fitness(i)<fitness(sol1)，则fitness(i)＝fitness(sol1)，否则，fitness(i)不变；

[0084] (6)进入侦察蜂阶段：判断采蜜蜂转侦察蜂的条件是否满足，如果采蜜蜂的适应度值fitness(i)在limit次迭代中均未发生变化，则认为该采蜜蜂的适应度值fitness(i)为局部最优解，放弃局部最优解，同时该采蜜蜂转变为侦查蜂，按照V_i＝rand(c,s).*(max(X_New)-min(X_New))+min(X_New)在解空间进行新的搜索；否则，该采蜜蜂不转变；s为新样本集X_New中每个样本元素的维数，max(X_New)为行向量，由新样本集X_New中每列的最大值组成，min(X_New)也为行向量，由新样本集X_New中每列最小值组成，rand(c,s)为由[0,1]构成的c*s矩阵，*表示矩阵之间对应元素相乘；

[0085] (7)重复步骤(3)至(6)，直到最大迭代次数maxcycle或者满足||Vbest(iter+1)-Vbest(iter)||<ε；||Vbest(iter+1)-Vbest(iter)||表示第iter+1次迭代得到最优聚类中心和第iter次迭代得到最优聚类中心的欧式距离；

[0086] (8)得到最终最优聚类中心Vbest，并由Vbest得到对应的隶属矩阵U，并按照c_i＝argmax(u_{ij})得到最终聚类。

[0087] 采用本实施例所述方法及采用现有的PCM、PFCM、ABC_KPFECM四种算法的得到的实

验结果的聚类精度如表1所示：

[0088] 表1

[0089]

算法	FCM	PCM	PFCM	ABC_KPFECM
精确度	68.54%	64.51%	69.01%	93.45%
时间(s)	0.5895	0.5078	0.6405	14.7613

[0090] 由表1可知，本发明提出的方法，以较少的时间代价，显著提高了聚类算法的精度，使得聚类算法的性能得到大幅提高。

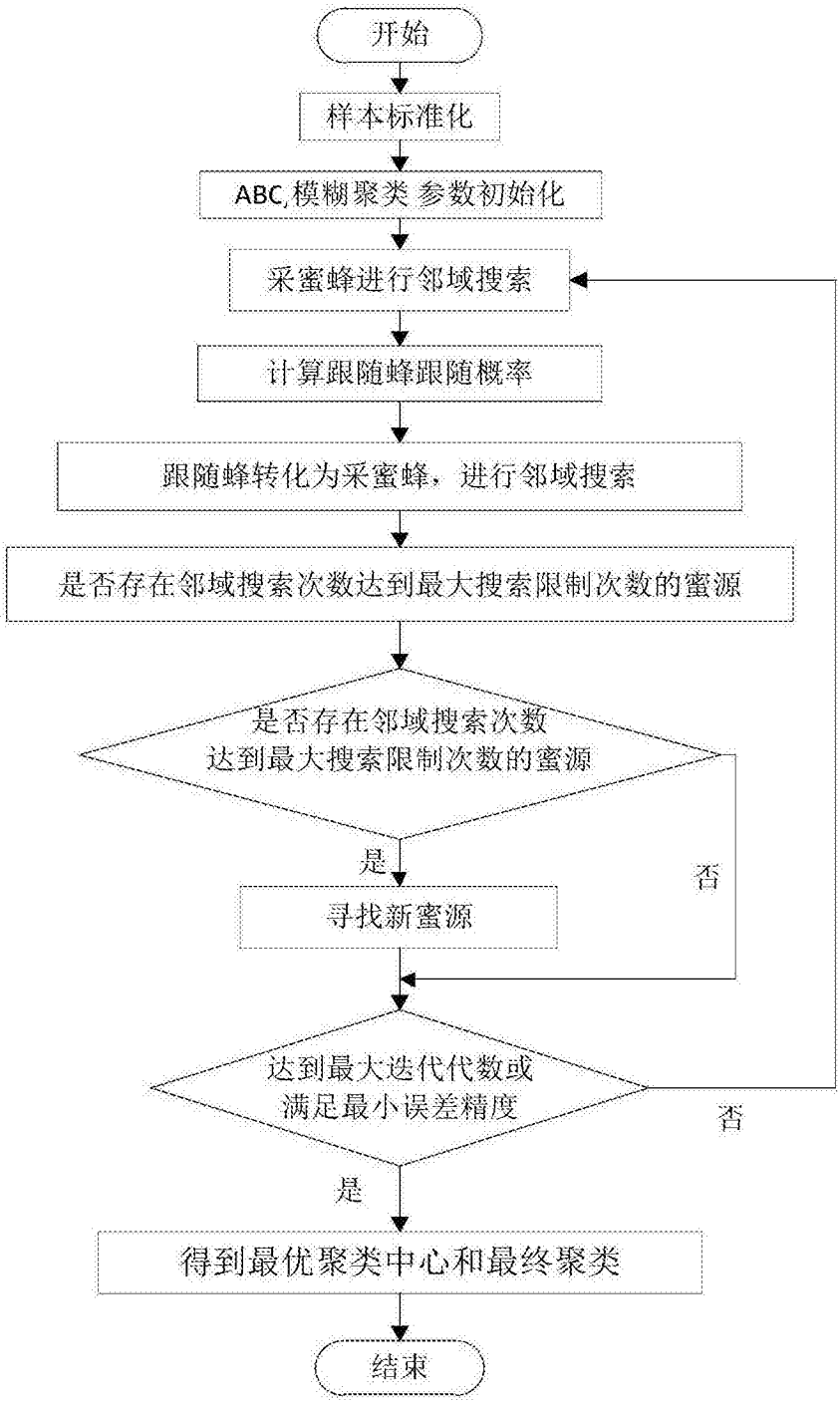
开始
样本标准化
ABC_模糊聚类 参数初始化
采蜜蜂进行邻域搜索
计算跟随蜂跟随概率
跟随蜂转化为采蜜蜂，进行邻域搜索
是否存在邻域搜索次数达到最大搜索限制次数的蜜源
是否存在邻域搜索次数
达到最大搜索限制次数的蜜源
是
否
寻找新蜜源
达到最大迭代代数或
满足最小误差精度
否
是
得到最优聚类中心和最终聚类
结束

图1