doi:10.11959/j.issn.1000-436x.2017022

混合编码方式的图像聚类算法

赵春晖，李雪源，崔颖
（哈尔滨工程大学信息与通信工程学院，黑龙江 哈尔滨 150001）

摘 要：基于群体智能优化算法的图像聚类分析，大多数都采用单一的编码方式，使搜索空间过于局限，算法很容易陷入局部最优，为了解决这个问题，提出一种混合编码方式的图像聚类分析算法（HEICA）。该算法构建一种基于图像聚类的混合编码模型，在扩大搜索空间范围的同时，与改进的雨林算法（IRFA）和量子粒子群算法（QPSO）相结合，提高全局搜索能力。在仿真实验中，采用 4 组数据集对算法进行聚类有效性测试，并将其与 4 种常用的聚类算法进行对比，实验结果表明该算法具有较强的全局搜索能力，稳定性高、聚类效果好。
关键词：图像聚类分析；混合编码；雨林算法；量子粒子群
中图分类号：TP753 **文献标识码**：A

Image cluster algorithm of hybrid encoding method

ZHAO Chun-hui, LI Xue-yuan, CUI Ying
（College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001,China）

Abstract: In the clustering analysis based on swarm intelligence optimization algorithm, the most of encoding method only used single form, and this method might be limit range of search space, the algorithm was easy to fall into local optimum. In order to solve this problem, image clustering algorithm of hybrid encoding (HEICA) was proposed. Firstly, a hybrid encoding model based on image clustering was established, this method could expand the scope of the search space. Meanwhile, it was combined with two optimization algorithms which improved rain forest algorithm (IRFA) and quantum particle swarm optimization (QPSO), this method could improve the global search capability. In the simulation experiment, it was carried out to illustrate the performance of the proposed method based on four datasets. Compared with results form four measured cluster algorithm. The experimental results show that the algorithm has strong global search capability, high stability and clustering effect.
Key words: image cluster analysis, hybrid encoding, rain forest algorithm, quantum particle swarm optimization

1 引言

聚类分析是将一组数据或对象通过某种特定规则分成不同类的过程，同一类的数据具有较大的相似性，不同类别之间具有较小的相似性，在数据挖掘、图像处理、模式识别等领域具有重要作用。图像聚类可以被定义为不同集合的图像最优划分，使同一组的图像更为相似，而另一组或来自不同组的图像有最大差异[1]，揭示图像分布的规律性。即在给出的图像集合中，根据图像的内容，在无先验知识的条件下，将人们感兴趣的图像按其不同的需求进行聚类。与数据聚类相比，图像聚类方法的目的是更好地组织、表示和浏览图像，以及在图像检索前建立有效的索引以提高图像检索的性能。

K 均值作为一种典型的图像聚类算法，它的初始聚类中心敏感，容易陷入局部最优，针对此问题，文献[2]提出一种基于特征关联度，结合“最小最大”原则选取初始聚类中心的方法；文献[3]采用流形距

收稿日期：2016-05-20；**修回日期**：2016-12-27
基金项目：国家自然科学基金资助项目（No.61405041, No.61571145）；黑龙江省自然科学基金资助项目（No.ZD201216）；黑龙江省博士后特别基金资助项目（No.LBH-TZ0420）
Foundation Items: The National Natural Science Foundation of China(No.61405041, No.61571145), The Natural Science Foundation of Heilongjiang Province (No.ZD201216), Heilongjiang Postdoctoral Special Scholars Foundation(No.LBH-TZ0420)

离的相似性度量方法，提出一种基于流形距离与进化算法相结合的图像聚类方法；文献[4]提出基于指数判别分析的图像聚类方法，算法具有较少的参数局部学习；文献[5]提出一种将蜜蜂交配优化算法用于聚类分析的全局优化算法；文献[6]提出一种基于直觉模糊核与粒子群算法相结合的聚类算法。此外，许多智能优化算法被应用于聚类分析的研究，如 PSO[7]、ACO[8]和 ABC[9]等，被广泛应用于解决 K 均值等聚类分析方法的不足，并且取得了一定的成果。

但在目前的群体智能优化算法的聚类分析中，大多数采用单一编码方式处理优化问题。文献[10]中进化粒子群数据聚类采用的是基于聚类中心的编码（CCE）。文献[11]提出一种将样本编号的简化粒子编码结构，被称为基于样本编号的编码（SNE）。其中，CCE 应用最为广泛，但是在搜索过程中，很容易产生超出搜索范围的解，降低搜索效率。SNE 是在已知的搜索空间（待分类样本编号）内，组成聚类中心的编码结构，这种编码结构能够将搜索空间控制在一定范围内，但同时也限制了在搜索范围外最优解的产生。因此，本文提出了一种混合编码方式的图像聚类算法，该算法采用 CCE 和 SNE 这 2 种编码方式的混合，同时解决搜索空间限制和超出搜索范围的问题。算法不仅扩大了搜索空间，而且避免了无效解的产生，从而进一步提高了全局搜索能力。然后再结合 IRFA 和 QPSO，分别使用不同的编码方式在相应的搜索空间内寻找最优的聚类中心。最后通过仿真实验证明了算法的有效性。

2 相关知识

2.1 编码方式

编码就是把一个实际问题的所有解从其空间转换到算法所能处理的搜索空间，编码过程就是实际问题数学化的过程。使用智能优化算法解决图像的聚类问题时，常用的编码方式包括 CCE、SNE 等，这些编码方式常常是以图像的数据矩阵为基础通过特定规则形成的。

图像的数据矩阵是一组不同类别的像素点组成的二维矩阵，如图 1 所示。假设高光谱图像的数据点为 $\boldsymbol{x}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_M]$，其中，$M$ 为样本数量，$\boldsymbol{x}_i=[x_{i1},x_{i2},\cdots,x_{id}]$代表第 i 个数据点的 d 维像素值，$\boldsymbol{x}_j=[x_{1j},x_{2j},\cdots,x_{Mj}]^{\mathrm{T}}$ 代表第 j 维的 M 个数据点的像素值，其中，d 为波段数，k 为图像的类别数。

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{M1} & x_{M2} & \cdots & x_{Md} \end{bmatrix}$$

图 1 图像的数据矩阵

CCE 是最常用的编码方式，其基本结构如图 2 所示。设聚类中心矩阵 $\boldsymbol{C}=[\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_k]$，其中，第 i 个聚类中心为 $\boldsymbol{C}_i=[C_{i1},C_{i2},\cdots,C_{id}]$，维数为 d，CCE 编码结果是一组由 k 个聚类中心组成的矩阵，与传统的 K 均值算法相同，每个聚类中心 $\boldsymbol{C}_i$ 在数据集中随机选出。

$$\begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1d} \\ C_{21} & C_{22} & \cdots & C_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ C_{k1} & C_{k2} & \cdots & C_{kd} \end{bmatrix}$$

图 2 CCE 编码结构

SNE 是将数据点按顺序编号为 1~M，如图 3 所示，设聚类中心矩阵 $\boldsymbol{Z}=[Z_1,Z_2,\cdots,Z_k]$，其中，第 i 个聚类中心为 Z_i，它是在编号 1~M 中随机选出的。SNE 编码方式在搜索过程中，搜索空间是一个 k 维空间，每一维的范围为[1,M]。

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{M1} & x_{M2} & \cdots & x_{Md} \end{bmatrix} \Rightarrow \begin{bmatrix} 1 \\ 2 \\ \vdots \\ M \end{bmatrix} \overset{round(rand(1,M))}{\Rightarrow} \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_k \end{bmatrix}$$

图 3 SNE 编码过程

2.2 雨林算法

雨林算法（RFA）是仿照植物生长模式，结合均匀与非均匀采样原则来权衡优化算法的探索和挖掘，在提高寻找最优解效率的同时，避免在采样过程中无约束性和样本分布信息的缺失，算法通过播种、萌发、生长、竞争、计寿和繁衍等 6 步进行寻优，算法可以有效减少虚拟碰撞的发生，并且能够获得精准性和稳定性较高的全局最优解[12]。

2.3 量子粒子群

量子粒子群[13]（QPSO）算法是将量子力学的概念引入粒子群的粒子群算法的演进。优化算法过程与传统粒子群算法相同，QPSO 算法根据人类群体自组织性和协同性等特点开发，算法易于实现、参数较少，在多个应用领域以及算法改进上得到了一定的关注，QPSO 算法以及其改进算法被广泛应

用于聚类[14~16]。

设种群$\boldsymbol{X}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$是由$N$个粒子组成，其中，第 i 个粒子的位置 $\boldsymbol{X}_i=[X_{i1},X_{i2},\cdots,X_{in}]$，$n$ 为维数。在迭代过程中，第 i 个粒子当前最好位置 $\boldsymbol{P}_i=[P_{i1},P_{i2},\cdots,P_{in}]$，以及群体所有粒子中最好位置 $\boldsymbol{P}_g=[P_{g1},P_{g2},\cdots,P_{gn}]$，由此，粒子 $\boldsymbol{X}_i$ 更新方式如式(1)和式(2)所示。

$$p_{id}=\varphi P_{id}+(1-\varphi)P_{gd},\varphi=rand() \tag{1}$$

$$X_{id}=p_{id}\pm\alpha\,|\,mbest_d-X_{id}\,|\ln\left(\frac{1}{u}\right),u\sim U(0,1) \tag{2}$$

其中，$mbest$ 代表是粒子间的平均最佳位置，由式(3)得出。

$$mbest=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{P}_i=\left(\frac{1}{M}\sum_{i=1}^{M}P_{i1},\frac{1}{M}\sum_{i=1}^{M}P_{i2},\cdots,\frac{1}{M}\sum_{i=1}^{M}P_{in}\right) \tag{3}$$

其中，p_{id} 代表 P_{id} 和 P_{gd} 之间的一个随机点，被称为第 i 个粒子的第 d 维局部吸引粒子，φ 是[0,1]范围内均匀分布的随机数，u 是另一个均匀分布在[0,1]之间的随机数，α 被称为收缩扩张系数，是 QPSO 算法的参数[13]。

3　混合编码方式

3.1　引入混合编码的意义

目前，使用的编码方式都是单一形式的编码，易造成搜索空间范围受到限制，可能会导致 2 种情况发生：1) 如果对搜索范围进行控制（SNE 编码方式），可能导致种群多样性降低；2) 如果没有固定的搜索空间（CCE 编码方式），可能导致在搜索过程中，很容易产生超出搜索空间的无效解，并且很难确定有效的映射方法，这样就降低了搜索效率。为了解决以上问题，引入混合编码的方式，通过 2 种编码方式的混合，扩大搜索空间的范围，提高种群多样性。这个阶段定义了一种基于图像聚类的混合编码的模型。

3.2　基于混合编码方式的种群初始化

基于混合编码方式的种群初始化是通过定义 2 个种群，分别使用不同的编码方式，在相应的搜索空间内产生初始解的过程。采用 SNE 编码方式初始化种群，种群个体定义为 $\boldsymbol{T}$；采用 CCE 编码方式初始化种群，种群个体定义为 $\boldsymbol{X}$。

1) 种群 $\boldsymbol{T}$ 初始化

如图 3 所示，将数据矩阵编号后，从相应的编号中随机选取 k 个作为聚类中心矩阵，重复 N 次，则组成一个种群规模为 N 的种群 $\boldsymbol{T}=\{\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_N\}$。但是，这种随机初始化种群的方法会造成采样空白，文献[12]提出了一种在未知区域均匀采样的方式，可以降低采样空白的影响，在未知区域均匀采样的方法如下。

定义 k 维空间 R^k，搜索空间为$[\boldsymbol{r}_1\times\boldsymbol{r}_2\times\cdots\times\boldsymbol{r}_k]$，其中，$\boldsymbol{r}_i$ 代表第 i 维向量，k 代表聚类的类别数。在 R^k 上以阵列的形式均匀采样得到 $N=n_1\times n_2\times\cdots\times n_k$ 个人工树节点（母节点），n_i 代表第 i 维内均匀采样的个数。搜索空间的每一维的范围都是$[1,M]$，M 是样本点的个数。种群 $\boldsymbol{T}$ 采用如图 3 所示的编码结构，将聚类中心矩阵 $\boldsymbol{Z}$ 作为个体的位置，即第 i 个个体的位置 $\boldsymbol{T}_i=\boldsymbol{Z}=[Z_1,Z_2,\cdots,Z_k]$。如图 4 所示，以类别数 k=2 情况为例，每个点代表一个个体 $\boldsymbol{T}_i$，对应的是数据点的标号，如[3,58]、[96,32]、[33,89]等。

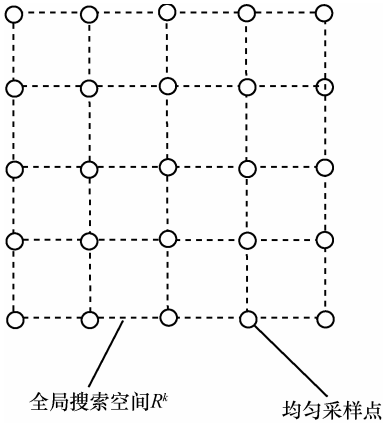


图 4　种群 $\boldsymbol{T}$ 初始化示例

2) 种群 $\boldsymbol{X}$ 初始化

定义种群 $\boldsymbol{X}=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$，$N$ 为种群规模，粒子采用如图 2 所示的编码结构，将聚类中心矩阵 $\boldsymbol{C}$ 作为粒子的位置，即第 i 个粒子的位置 $\boldsymbol{X}_i=\boldsymbol{C}=[\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_k]$。种群 $\boldsymbol{X}$ 中所有粒子的位置是通过种群 $\boldsymbol{T}$ 中每个样本点所对应的数据点的位置决定的，即 $\boldsymbol{T}_i=[Z_1,Z_2,\cdots,Z_k]\rightarrow\boldsymbol{X}_i=[\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_k]$，通过式(4)将 Z_i 所代表的相应数据点 $\boldsymbol{x}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_M]$所代表的位置，对应为相应的聚类中心。以此为例，将 $\boldsymbol{T}_i$ 全部映射为 $\boldsymbol{X}_i$ 后，得到种群 $\boldsymbol{X}$。

$$[\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_k]=[\boldsymbol{x}_{z_1},\boldsymbol{x}_{z_2},\cdots,\boldsymbol{x}_{z_k}] \tag{4}$$

其中，$\boldsymbol{C}_i$ 代表粒子 $\boldsymbol{X}_i$ 中第 i 个聚类中心，Z_i 代表 $\boldsymbol{T}_i$ 中第 i 个聚类中心，$\boldsymbol{x}_{z_i}$ 代表第 Z_i 个数据点。

4 混合编码方式的图像聚类算法设计

根据混合编码的思想，为了实现混合编码，引入 RFA 和 QPSO 这 2 种群体智能优化算法，根据各自不同的搜索策略，结合不同的编码方式，在各自的搜索空间内寻找最优解。提出一种混合编码方式的图像聚类算法（HEICA）。算法分为 2 个阶段，第 1 阶段引入改进的雨林算法，结合 SNE 编码方式，采用均匀采样的方式在搜索空间范围内进行初始化，避免了采样空白[12]的产生，丰富了种群的多样性，在此基础上进行局部寻优，提高了全局搜索能力。第 2 阶段引入 QPSO 算法，结合 CCE 编码规则，扩展算法的搜索空间，弥补了上一阶段个体独立搜索造成的个体间信息的缺失问题，通过群体中个体之间的交互，以达到快速收敛的目的。

4.1 算法原理

算法的第 1 阶段是根据 SNE 编码规则，结合改进的雨林算法（IRFA）原理中的局部领域搜索方法，算法原理如下。

雨林算法采用分级寻优采样的方式，每个母节点产生的枝叶以特定的规则保留，算法复杂度高，为降低算法复杂度，算法取消分级采样的模式，选择最优枝叶进行保留。如图 5 所示，在全局搜索空间 R^k 内，采用均匀采样的方式初始化种群。在迭代过程中，采用非均匀采样的方法对局部最优区域进行挖掘，保留最优新生枝叶作为下一代的母节点，具体实现方法如下。

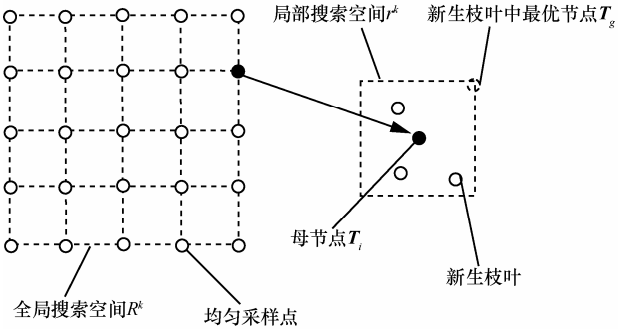


图 5 局部采样

在 k 维空间 R^k 中，首先，以均匀采样的方式初始化种群 $\boldsymbol{T}=\{\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_N\}$，其中，$\boldsymbol{T}_i$ 为第 i 个人工树的母节点的位置 $\boldsymbol{T}_i=[Z_1,Z_2,\cdots,Z_k]$，然后对于每个母节点 $\boldsymbol{T}_i$，在其局部搜索空间 r^k 内通过非均匀采样的方式产生新的枝叶，首先计算新生枝叶数量 n_i 及其伸展范围 r，计算式为

$$n_i = m\frac{f(\boldsymbol{x}_i)-y_{\min}+\xi}{\sum_{i=1}^{N}(f(\boldsymbol{x}_i)-y_{\min})+\xi} \tag{5}$$

其中，m 是控制数量级的参数，$f(\boldsymbol{x}_i)$是第 i 个适应度函数值，$y_{\max}$ 是最优适应度函数值，ξ 是极小值。通过每个节点的适应度函数值来控制新生枝叶的数量，适应度值越大（最好值），相应产生新的枝叶的数量越多，通过 m 调节 n_i 的数量级，然后通过式(6)控制新生枝叶的伸展范围 r。

$$r=\left(\frac{t-h}{t}\right)^{expo}(r_{\text{init}}-r_{\text{end}})+r_{\text{end}} \tag{6}$$

其中，t 是最大迭代次数，h 是当前迭代次数，$expo$ 是控制变量，一般为 2，r_{init} 和 r_{end} 分别表示迭代开始时和迭代结束时的伸展范围，r 是以迭代次数呈指数递减的，即随着迭代次数的增加，伸展范围逐渐缩小。通过 r_{init} 和 r_{end} 控制最大和最小伸展范围。

然后，以母节点 $\boldsymbol{T}_i$ 为中心，根据新生枝叶伸展范围 r，在局部搜索空间 r^k 中，根据式(7)产生新的枝叶 $\boldsymbol{Tn}_g$，重复 n_i 次，产生 n_i 个新生枝叶，定义为

$$\boldsymbol{Tn}_g^k=\boldsymbol{T}_i^k+rand[0,1]^k\,rand[0,r]^k \tag{7}$$

其中，k 为维度，$\boldsymbol{Tn}_g$ 代表第 $g(1\leqslant g\leqslant n_i)$个新生枝叶的位置，$\boldsymbol{T}_i$ 为树的母节点位置，$rand[0,1]$为 0~1 之间的随机数，用来控制搜索方向，$rand[0,r]$为 0~r 的随机数，用来控制搜索范围。

使用类内距离来评价新生枝叶 $\boldsymbol{Tn}_g$ 的聚类内聚程度，即

$$f=\frac{k_0}{\sum_{i=1}^{k}\sum_{\boldsymbol{x}_j\in\boldsymbol{c}_i}d(\boldsymbol{x}_j,\boldsymbol{z}_i)} \tag{8}$$

其中，k_0 为常数，k 为聚类类别数，$\boldsymbol{x}_j$ 为第 j 个数据点，$\boldsymbol{z}_i$ 是第 i 个聚类中心的位置，$\boldsymbol{c}_i$ 代表第 i 类集合，$d(\boldsymbol{x}_j,\boldsymbol{z}_i)$为第 j 个数据与第 i 个聚类中心之间的欧式距离。

对新生的 n_i 个枝叶，保留最优的枝叶 $\boldsymbol{Tn}_{best}$ 作为下一轮的母节点 $\boldsymbol{T}_i(h+1)$，即

$$\begin{cases} best=g \\ \boldsymbol{Tn}_{best}=\boldsymbol{Tn}_g \end{cases},\ f_g=f_{\max} \tag{9}$$

其中，f_g 代表第 g 个新生枝叶的适应度函数值，$f_{\max}$ 代表其中最优的适应度函数值，$best$ 代表 1 到 n_i 个新生枝叶中最优枝叶的编号，$\boldsymbol{Tn}_{best}$ 代表最优枝叶的位置。

算法的第 2 阶段是根据 CCE 编码规则，结合 QPSO 算法的全局寻优方法。算法原理如下。

在第 1 阶段中，通过改进的雨林算法，采用 SNE 编码方式，在固定的搜索空间 R 范围内进行了一次局部邻域搜索，为了进一步扩大搜索空间的范围，引入 QPSO，采用 CCE 编码方式，在上述搜索空间 R^k 的基础上，在新的搜索空间 S^n 内搜索最优解，如图 6 所示。算法数学描述如下。

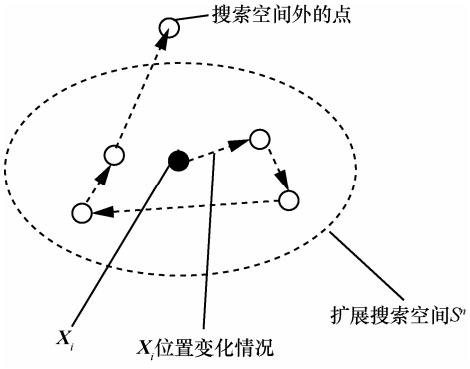


图 6　扩展空间寻优示意

首先，按照 3.2 节中的方法初始化种群 $\boldsymbol{X}$，初始时，粒子的位置定义为 $\boldsymbol{X}_i(0)$，粒子的最好位置 $\boldsymbol{P}_i(0)=\boldsymbol{X}_i(0)$以及群体所有粒子中最好位置 $\boldsymbol{P}_g(0)$。

对于改进的雨林算法局部邻域搜索新产生的母节点 $\boldsymbol{T}_i$，由式(4)转变其编码方式，得到 CCE 编码方式的位置，定义为 $\boldsymbol{M}_i$，并与当前个体最好位置 $\boldsymbol{P}_i$ 比较，即

$$\begin{cases} \boldsymbol{P}_i = \boldsymbol{M}_i \\ f(\boldsymbol{P}_i) = f(\boldsymbol{M}_i) \end{cases}, \quad f(\boldsymbol{M}_i) > f(\boldsymbol{P}_i) \tag{10}$$

其中，$f(\boldsymbol{M}_i)$、$f(\boldsymbol{P}_i)$为 $\boldsymbol{M}_i$ 和 $\boldsymbol{P}_i$ 的适应度函数值，更新 $\boldsymbol{P}_i$ 的位置，并且找出群体的全局最好位置 $\boldsymbol{P}_g$。

然后，根据 QPSO 算法的位置更新规则，在扩展搜索空间 S^n 内通过多次迭代更新 $\boldsymbol{X}_i$ 的位置，最后得到全局最优解。

4.2　算法步骤及复杂度分析

4.2.1　算法步骤

根据算法原理，基于混合编码的图像聚类算法是通过 IRFA 和 QPSO 这 2 种群体智能优化算法，分别使用 SNE 和 CCE 编码方式，在搜索空间 R^k 及其扩展空间 S^n 范围内，寻找全局最优解的过程，算法主要步骤如下。

Step1　播种。实现种群初始化过程，定义种群 $\boldsymbol{T}$ 人工树的母节点位置 $\boldsymbol{T}_i$，种群 $\boldsymbol{X}$ 的粒子 $\boldsymbol{X}_i$，每一个粒子经历的最好位置记为 $\boldsymbol{P}_i$，初始时 $\boldsymbol{P}_i=\boldsymbol{X}_i$；种群中所有粒子当前经历的全局最好位置记为 $\boldsymbol{P}_g$，并且初始化参数群体规模为 N，学习因子为 α，控制数量级参数为 m，最大迭代次数为 T，极小值为$\S$，迭代开始时伸展范围为 r_{init}，迭代结束时伸展范围为 r_{end}。以及量子粒子群收缩扩张系数为 β。

Step2　萌发。根据式(5)和式(6)计算围绕人工树节点 $\boldsymbol{T}_i$而产生的新生枝叶的数量 n_i和伸展范围 r。

Step3　生长。根据式(7)以母节点为中心，产生新的枝叶 $\boldsymbol{Tn}_g$。

Step4　竞争。通过式(8)的类内距离来评价新生枝叶的聚类内聚程度，并通过式(9)保留最优的枝叶 $\boldsymbol{Tn}_{best}$ 作为下一轮的人工树节点 $\boldsymbol{T}_i(h+1)$。

Step5　传递。新生的枝叶的位置 $\boldsymbol{X}_i$，由式(4)转换编码方式，转换为 $\boldsymbol{M}_i$，并与当前个体最好位置 $\boldsymbol{P}_i$ 比较，通过式(10)更新 $\boldsymbol{P}_i$ 的位置，并且找出群体的全局最好位置 $\boldsymbol{P}_g$。

Step6　根据式(1)和式(3)计算每个粒子的局部吸引粒子 p_{id} 和种群的重心位置 $mbest$；根据式(2)更新粒子 $\boldsymbol{X}_i$ 的位置，记为 $\boldsymbol{X}_i(h+1)$。

Step7　根据式(8)评价 $\boldsymbol{X}_i(t+1)$ 的目标函数 $f(\boldsymbol{X}_i(t+1))$，根据式(9)更新 $\boldsymbol{P}_i$ 的位置，并且找出群体的全局最好位置 $\boldsymbol{P}_g$。

Step8　结束。判断是否满足迭代终止条件，如果满足，则结束；否则跳到 Step2 执行。

4.2.2　算法复杂度分析

根据算法的时间复杂度及空间复杂度的理论分析，k 均值聚类算法的一次迭代需要运行的计算次数为 nc，其算法的时间复杂度为 $O(ncl)$，其中，n 为数据集样本数目，c 为聚类类别数，l 为算法的迭代次数，因此，K 均值聚类算法的时间复杂度也可记为 $O(n)$。同理，空间复杂度也为 $O(n)$，本文采用 IRFA 和 QPSO 这 2 种群体智能算法，IRFA 算法每次迭代需要计算新生枝叶数量、伸展范围、枝叶的位置，计算次数为 $3Nc$。QPSO 算法每次迭代需要计算局部吸引粒子、种群的重心位置、粒子的位置，计算次数同样为 $3Nc$，其中，N 为种群规模。因此，这 2 种群体智能优化算法的时间复杂度为 $O(6Ncl)$，可以记为 $O(N)$，同理，其空间复杂度也为 $O(N)$[5]。

由上述分析可知，本文所提出的混合编码方式的图像聚类算法时间复杂度为 $O(n+N)$，空间复杂度为 $O(n+N)$，但相较于其他对比算法，本文算法具有较高的时间复杂度，影响时间复杂度的主要因素为

种群规模 N、聚类类别数 c 和迭代次数 l。因此，为降低算法复杂度，需要降低种群规模以及迭代次数。

本文所选择的种群规模和数据集样本数目是固定参数，在所选择对比算法中，除雨林算法外，种群规模和数据集样本数目都是固定参数。雨林算法是通过规模可变的种群代替规模限定种群进行分区分级寻优采样的方式，在迭代过程中，种群规模会模拟林木的生长进化特点，种群规模逐级递增，因此，雨林算法的算法复杂度高，本文通过改进的雨林算法，取消了种群规模可变的方式，降低算法复杂程度，以便更好地应用于聚类分析。

5 实验结果与分析

5.1 实验环境及数据

测试软件平台为 Windows 10，Matlab 2012a，处理器为 Intel Core i7-6700HQ、CPU 为 2.6 GHz，内存为 8 GB，硬盘 1 TB。所有实验结果是各聚类算法在所有数据集上分别独立运行 10 次之后进行统计平均得到的。

本文采用 Indian Pines 高光谱图像数据，这个高光谱图像是通过 AVIRIS 传感器在印第安西北测试地点采集而来的，图像为 145×145 像素，有 224 个光谱反射率波段，波长范围 0.4×10^{-6}~2.5×10^{-6} m。其中，农作物占整幅图像的$\frac{2}{3}$，另外的$\frac{1}{3}$为森林或其他自然植被。图像中有 2 个主要的铁路线，还有一些低密度的住宅、其他建筑和小型公路，由于现场拍摄于 6 月，部分农作物（玉米、大豆等）均处于生长发育期，覆盖率不到 5%。真实地面可被指定为 16 类，波段数降低到 200，其中，去除了水分子吸收的区域覆盖波段：104~108、150~163。如图 7 所示，是由波段 50、27 和 17 构成的假彩色合成图像。从中选取不同地物，每 3 类作为一组、每类选取 100 个样本点，作为测试数据集。表 1 为 4 种测试数据集所选取的类别。

图 7 波段 50、27、17 合成的假彩色图像

表 1 测试数据集

数据名称	样本类别
数据集 1	大豆—免耕、大豆—小、大豆—干净
数据集 2	牧场、灌木、干草
数据集 3	玉米—免耕、玉米—小、玉米
数据集 4	小麦、树林、建筑物

5.2 实验参数及性能评价

在实验中，算法的参数设置采用多次测试值和文献[12]参考值，具体参数为：雨林的群体数量为 64，学习因子 α=0.5，控制数量级参数 m=400，最大迭代次数 T=200，极小值$\delta=10^{-9}$，迭代开始时伸展范围 r_{init}=30，迭代结束时伸展范围 r_{end}=10，量子粒子群收缩扩张系数 β=0.8。

文中的评价指标采用精度评价，精度评价是指比较实际数据和分类结果，以确定分类过程的准确度。最常用的精度评价方法是总体分类精度 OA 和 Kappa 系数，采用文献[1]的式子，如式(11)、式(12)所示。

$$OA=\sum_{i=1}^{k}\frac{n_i}{n} \tag{11}$$

$$k=\frac{n\sum_{i=1}^{k}n_i-\sum_{i=1}^{k}n_{i+}n_{+i}}{n^2-\sum_{i=1}^{k}n_{i+}n_{+i}} \tag{12}$$

其中，n_i 为第 i 类正确分类的样本个数，n 为样本的总个数。n_{i+}表示第 i 类的真实像元的总和，n_{+i} 表示第 i 类被分类的像元总数。

5.3 实验设计及结果分析

为验证算法的有效性，本文采用 HEICA 与 4 种常见的聚类分析方法进行比较，分别是 PSO 聚类算法[7]、FPSO 聚类算法[11]、QPSO 聚类算法[14]和 RFA 聚类算法。

如表 2 所示，除了在数据集 3 情况下，本文算法相较于其他聚类分析方法总体精度有所提高，而在数据集 2 和数据集 4 相较于 PSO 算法总体精度分别提高了 2.16%和 2.37%。数据集 1 和数据集 3 相较于 PSO 算法总体精度分别提高了 0.23%和 0.6%。实验结果表明，本文算法比较适用于相似度较低的地物，具有较高的聚类准确度，并且证明混合编码方式的聚类效果要优于分别使用不同编码方式的 FPSO 算法和 QPSO 算法。

表 2　各种聚类算法总体精度比较

数据集	PSO	FPSO	QPSO	RFA	本文算法
数据集 1	76.57%	75.87%	75.10%	75.77%	76.80%
数据集 2	92.46%	94.63%	94.60%	94.67%	95.07%
数据集 3	68.63%	69.33%	69.33%	69.43%	69.23%
数据集 4	81.10%	68.20%	79.13%	79.47%	83.47%

表 3 为各种算法运行 10 次总体精度的标准差，代表数据分散程度的一种度量，标准差越小，代表数据越接近平均值，算法的稳定性越高。如表 3 所示，除了数据集 1 中 PSO 算法以及数据集 4 中 QPSO 算法之外，本文算法的标准差均小于其他算法，表明本文算法相较于其他算法具有较高的稳定性。

表 3　算法总体精度标准差

数据集	PSO	FPSO	QPSO	RFA	本文算法
数据集 1	0.006 5	0.017 8	0.015 3	0.013 0	0.009 1
数据集 2	0.072 3	0.003 7	0.005 6	0.017 7	0.003 4
数据集 3	0.005 2	0.004 4	0.006 8	0.030 4	0.003 2
数据集 4	0.107 9	0.151 7	0.094 9	0.120 4	0.098 9

如表 4~表 7 所示，通过比较类内距离，本文算法相较于其他算法，类内距离在一个很小的范围内变动，表明本文算法的稳定性高，不易陷入局部最优，从一定程度上解决了初始聚类中心敏感问题。此外，通过比较最大类内距离，发现 PSO 算法类内距离高于其他算法，但是总体精度并不高于本文算法，可能的原因是算法容易陷入局部最优，或者文中所选用的适应度函数并不能准确地反映出地物的类别，需要进一步的研究。如表 4~表 7 所示的运行时间，本文算法具有较高的时间复杂度（除了 RFA 迭代终止条件不同，算法平均迭代次数约为 20，其余算法迭代次数为 200），因此，本文算法是以效率为代价获取更高精度的方法。为解决此问题，算法也可以从降低迭代次数或降低种群规模的角度进行改进。

如图 8 所示，本文算法由于在迭代过程中，在 2 个搜索空间内不断寻优，不断地跳出局部最优解，收敛较慢，但在大多数数据集中，得到的最优值较高，表明算法全局搜索能力强，并且具有较强的顽健性。另外，数据集 4 虽然略低于 PSO 算法，但是其收敛速度优于 PSO 算法，并且相较于 PSO 算法，本文算法具有更高的总体精度。

表 4　各种聚类算法在数据集 1 下聚类效果比较

算法	类内距离			总体精度	Kappa 系数	运行时间/s
	最小	平均	最大			
PSO	1.538 4	1.569 9	1.608 1	76.57%	0.648 5	8.534 4
FPSO	1.502 2	1.521 6	1.530 6	75.87%	0.638 0	8.371 4
QPSO	1.403 2	1.487 8	1.543 4	75.10%	0.626 5	8.390 3
RFA	1.380 9	1.406 3	1.482 9	75.77%	0.636 5	4.666 0
本文算法	1.570 5	1.583 2	1.600 2	76.80%	0.652 0	90.823 8

表 5　各种聚类算法在数据集 2 下聚类效果比较

算法	类内距离			总体精度	Kappa 系数	运行时间/s
	最小	平均	最大			
PSO	0.006 1	1.164 9	1.316 2	92.46%	0.886 8	7.864 8
FPSO	1.236 9	1.246 1	1.254 7	94.63%	0.919 5	8.205 0
QPSO	1.176 8	1.234 3	1.266 7	94.60%	0.919 0	8.345 0
RFA	1.170 8	1.214 9	1.237 7	94.67%	0.914 0	5.254 8
本文算法	1.261 4	1.282 8	1.307 3	95.07%	0.926 0	86.771 4

表 6　各种聚类算法在数据集 3 下聚类效果比较

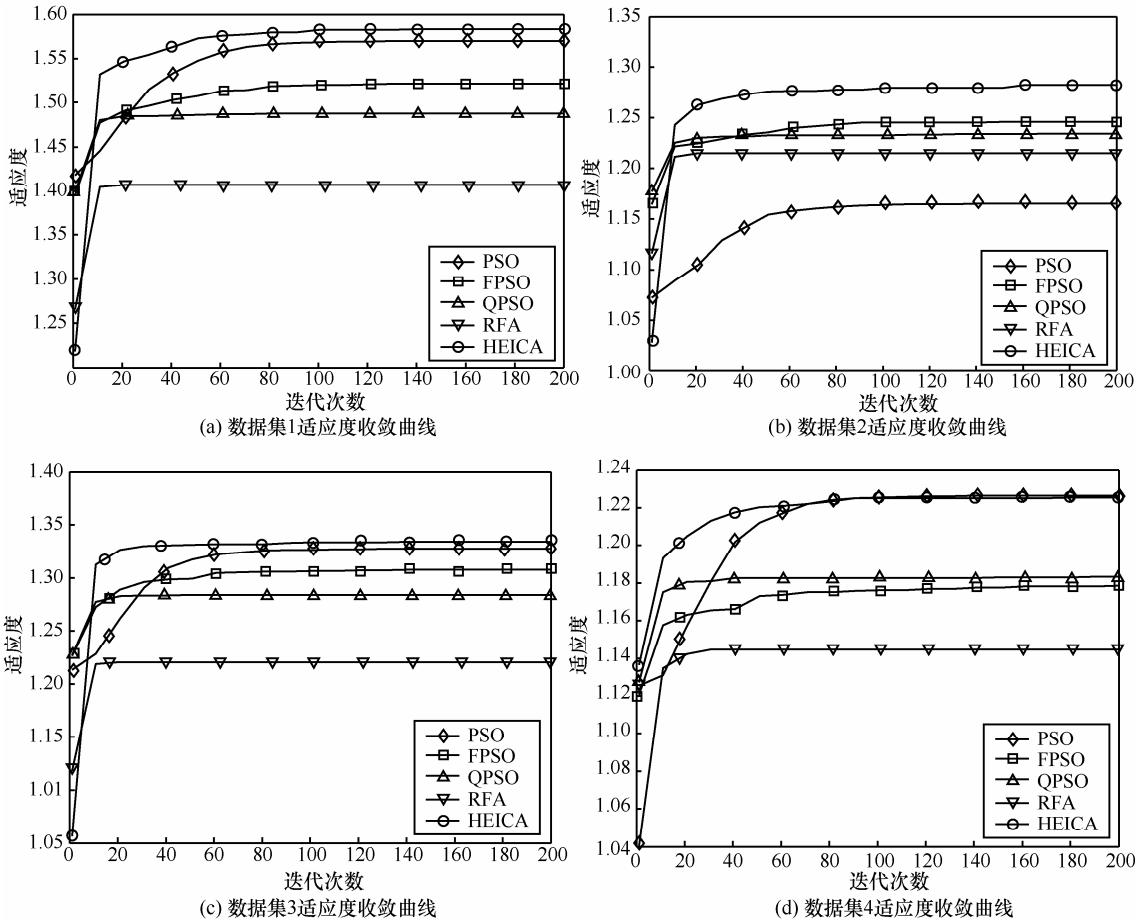
算法	类内距离			总体精度	Kappa 系数	运行时间/s
	最小	平均	最大			
PSO	1.304 3	1.327 6	1.355 4	68.63%	0.529 5	7.733 2
FPSO	1.290 5	1.307 9	1.318 5	69.33%	0.550 0	9.236 7
QPSO	1.202 9	1.284 3	1.332 6	69.33%	0.540 0	8.281 6
RFA	1.180 6	1.220 7	1.256 7	69.43%	0.541 5	8.347 0
本文算法	1.319 7	1.334 2	1.353 2	69.23%	0.538 5	92.577 4

表 7　各种聚类算法在数据集 4 下聚类效果比较

算法	类内距离			总体精度	Kappa 系数	运行时间/s
	最小	平均	最大			
PSO	1.192 2	1.226 6	1.249 3	81.10%	0.716 5	8.290 1
FPSO	1.173 1	1.178 5	1.185 7	68.20%	0.523 0	9.135 4
QPSO	1.139 2	1.183 3	1.217 8	79.13%	0.687 0	7.988 8
RFA	1.108 7	1.144 5	1.176 8	79.47%	0.692 0	11.060 0
本文算法	1.212 9	1.225 3	1.240 1	83.47%	0.752 0	91.747 6

6　结束语

本文提出了一种混合编码方式的图像聚类算法，采用 SNE 和 CCE 编码方式结合，扩大了搜索空间范围，丰富了种群多样性，通过在不同搜索空间内寻优，提高了算法的全局搜索能力。通过引入雨林算法和均匀采样的方式，在一定程度上解决了初始聚类中心敏感问题，避免算法陷入局部最优。



(a) 数据集1适应度收敛曲线　(b) 数据集2适应度收敛曲线

(c) 数据集3适应度收敛曲线　(d) 数据集4适应度收敛曲线

图 8　各种数据集在不同聚类算法中的适应度收敛曲线

实验结果表明，本文算法全局寻优能力强，顽健性高，能够较好地解决初始聚类中心敏感问题，较大程度避免了陷入局部最优。但是，使用类内距离最小准则来判别样本所属类别误差较大，通过多次运行发现，类内距离越小，并不能代表聚类效果越好，如何设定有效的判别准则是进一步研究的重点。本文算法具有较高的时间复杂度，为解决此问题，可以从设定迭代终止条件或舍弃质量差个体的角度进一步研究。

参考文献：

[1] AHMED N. Recent review on image clustering[J]. IET Image Processing, 2015, 9(11): 1020-1032.

[2] 陈兴蜀, 吴小松, 王文贤, 等. 基于特征关联度的 *K*-means 初始聚类中心优化算法[J]. 四川大学学报(工程科学版), 2015, 47(1): 13-19.
CHEN X S, WU X S, WANG W X, et al. An improved initial cluster centers selection algorithm for *K*-means based on features correlative degree[J]. Journal of Sichuan University (Engineering Science Edition), 2015, 47(1): 13-19.

[3] 李阳阳, 石洪竺, 焦李成, 等. 基于流形距离的量子进化聚类算法[J]. 电子学报, 2011, 39(10): 2343-2347.
LI Y Y, SHI H Z, JIAO L C, et al. Quantum-inspired evolutionary clustering algorithm based on manifold distance[J]. Acta Electronica Sinica, 2011, 39(10): 2343-2347.

[4] AHMED N. Image clustering using exponential discriminant analysis[J]. IET Computer Vision, 2015, 9(1): 1-12.

[5] 罗可, 李莲, 周博翔. 一种蜜蜂交配优化聚类算法[J]. 电子学报, 2014, 42(12): 2435-2441.
LUO K, LI L, ZHOU B X. A honey-bee mating optimization clustering algorithm[J]. Acta Electronica Sinica, 2014, 42(12): 2435-2441.

[6] 余晓东, 雷英杰, 岳韶华, 等. 基于粒子群优化的直觉模糊核聚类算法研究[J]. 通信学报, 2015, 36(5): 74-80.
YU X D, LEI Y J, YUE S H, et al. Research on PSO-based intuitionistic fuzzy kernel clustering algorithm[J]. Journal on Communications, 2015, 36(5): 74-80.

[7] MERWE D W, ENGELBRECHT A P. Data clustering using particle swarm optimization[C]//The 2003 Congress on Evolutionary Computation. 2003, 1: 215-220.

[8] TIWARI R, HUSAIN M, GUPTA S, et al. Improving ant colony optimization algorithm for data clustering[C]//International Conference and Workshop on Emerging Trends in Technology, 2010: 529-534.

[9] 罗可, 李莲, 周博翔. 基于变异精密搜索的蜂群聚类算法[J]. 控制与决策, 2014, 29(5): 838-842.
LUO K, LI L, ZHOU B X. Artificial bee colony rough clustering algorithm based on mutative precision search[J]. Control and Decision, 2014, 29(5): 838-842.

[10] ALAM S, DOBBIE G, REHMAN S U. Analysis of particle swarm optimization based hierarchical data clustering approaches[C]// Fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control. 2015:1-4.

[11] 王纵虎, 刘志镜, 陈东辉. 一种改进的粒子群优化快速聚类算法[J]. 西安电子科技大学学报, 2012, 39(5): 61-65.
WANG Z H, LIU Z J, CHEN D H. Improved PSO-based fast clustering algorithm[J]. Journal of Xidian University, 2012, 39(5): 61-65.

[12] 高维尚, 邵诚, 高琴. 群体智能优化中的虚拟碰撞：雨林算法[J]. 物理学报, 2013, 62(19): 28-43.
GAO W S, SHAO C, GAO Q. Pseudo-collision in swarm optimization algorithm and solution: rain forest algorithm[J]. Acta Physica Sinica, 2013, 62(19): 28-43.

[13] SUN J, FENG B, XU W. Particle swarm optimization with particles having quantum behavior[C]//Evolutionary Computation. 2004: 1571-1580.

[14] SUN J, XU W, YE B. Quantum-behaved particle swarm optimization clustering algorithm[C]//Advanced Data Mining and Applications. 2006: 340-347.

[15] WANG M, FANG W, LI C. Clustering quantum-behaved particle swarm optimization algorithm for solving dynamic optimization problems[C]//Bio-Inspired Computing-Theories and Applications. 2015: 411-421.

[16] 陈伟, 傅毅, 孙俊, 等. 一种改进二进制编码量子行为粒子群优化聚类算法[J]. 控制与决策, 2011, 26(10): 1463-1468.
CHEN W, FU Y, SUN J, et al. Improved vinery quantum-behaved particle swarm optimization clustering algorithm[J]. Control and Decision, 2011, 26(10): 1463-1468.

作者简介：

赵春晖（1965-），男，黑龙江汤原人，博士，哈尔滨工程大学教授、博士生导师，主要研究方向为数字信号与图像处理、数字形态学与高光谱遥感图像处理等。

李雪源（1989-），女，辽宁盘锦人，哈尔滨工程大学博士生，主要研究方向为高光谱图像处理。

崔颖（1979-），女，黑龙江哈尔滨人，博士，哈尔滨工程大学副教授，主要研究方向为遥感图像处理、智能信号处理、无线传感器网络优化等。