

基于 PSO 的云计算环境中大数据优化聚类算法

朱亚东¹,高翠芳²

(1. 江苏联合职业技术学院 信息中心,江苏 南京 211135;

2. 江南大学 理学院,江苏 无锡 214122)

摘要:在云计算环境下,对大数据进行优化聚类是实现数据优化访问和挖掘的基础。传统方法采用模糊 C 均值聚类算法进行云计算中的大数据聚类,易陷入局部极值,产生聚类偏移,效果不佳。提出一种基于优化粒子群(PSO)算法的大数据聚类算法。分析了云计算环境中的大数据结构模型,计算大数据的离散样本频谱特征,实现聚类样本的特征提取和信息模型构建。由于粒子群在搜索过程中经常会陷入局部最优解,采用混沌映射方法,带领粒子逃离局部最优解,设计粒子群优化算法进行特征聚类,达到大数据优化聚类的目的。仿真结果表明,采用该算法进行数据聚类,误分率降低,寻优性能较好,具有较好的应用价值。

关键词:粒子群;数据聚类;云计算;大数据

中图分类号:TP391.9

文献标识码:A

文章编号:1673-629X(2016)09-0178-05

doi:10.3969/j.issn.1673-629X.2016.09.040

Big Data Optimization Clustering Algorithm Based on PSO in Cloud Computing Environment

ZHU Ya-dong¹,GAO Cui-fang²

(1. Information Center, Jiangsu Union Technical Institute, Nanjing 211135, China;

2. School of Science, Jiangnan University, Wuxi 214122, China)

Abstract: In the cloud computing environment, the optimization of big data is the basis for the data optimized access and mining. In the traditional method, the fuzzy C means clustering algorithm is used to cluster the big data in the cloud computing, which is easy to fall into local extremum. A big data clustering algorithm based on Particle Swarm Optimization (PSO) is proposed. The big data structure model in cloud computing environment is analyzed, and the discrete sample spectrum characteristics of big data are calculated, realizing feature extraction and information model construction of clustering sample. The particles are often fallen into local extremum in searching. The chaotic mapping is used to take the particles against the local extremum. The PSO is designed to carry on the feature clustering for the purpose of optimization clustering for big data. Simulation shows that the proposed algorithm is used for data clustering, and the error rate is reduced, and the optimization performance is better, and it has good application value.

Key words: particle swarm; data clustering; cloud computing; big data

0 引言

各种云计算系统的出现使得信息处理和计算向着云计算方向发展。在云计算系统中,允许开发者将写好的程序放在“云”里运行,实现云计算系统的程控扩展和智能共享。在云计算环境中,海量的大数据需要进行调度和访问,达到数据挖掘的目的。实现云计算中大数据挖掘的基础在于数据聚类,因此研究云计算环境中大数据优化聚类算法具有重要意义。

聚类算法的本质是将海量大数据信息流通过统计

信息分析的方法分成若干个层次的子集,提取数据信息流的属性特征量,调整聚类中心实现数据聚类优化。传统大数据聚类算法主要有分割聚类算法、融合法和分裂法、层次类别算法以及神经网络控制算法^[1-3]。其中,采用粒子群聚类的聚类粒度分割算法具有典型性,取得了一定的研究成果。文献[4]提出一种基于 K-means 算法的云计算环境中的大数据聚类算法,基于互联网的相关服务的增加、使用和交付模式,实现大数据聚类;但是该算法存在对内存空间需求太大、计算开

收稿日期:2015-12-07

修回日期:2016-04-12

网络出版时间:2016-08-01

基金项目:国家自然科学基金青年基金(61402202)

作者简介:朱亚东(1976-),男,硕士,副教授,研究方向为计算机网络、信息安全。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20160801.0907.050.html>

销大的缺点。文献[5]提出一种基于模糊 C 均值聚类的云计算环境中的大数据聚类算法。算法随着数据量的增加,数据密度和类别距离大小出现非线性偏移,导致聚类中心不稳定,聚类效果不好。文献[6]提出基于分数阶 Fourier 变换特征匹配和 K-L 变换分类的云计算设备中的大数据特征高效分类挖掘算法,实现云计算设备中的大数据特征高效分类挖掘。算法的缺陷在于动态扩展性不好,且对初始聚类中心较为敏感,需要进行改进。

粒子群算法能够通过各个粒子间的合作和竞争关系寻求最优解,并且其算法结构简单,易实现,从而在参数优化方面备受关注。于是文中便利用粒子群的特点,并针对上述问题,提出一种基于改进粒子群(Particle Swarm Optimization, PSO)算法的云计算环境中大数据特征提取和大数据聚类算法。首先分析了云计算环境中的大数据结构模型,进行大数据的特征提取和信息模型构建,设计粒子群优化算法进行特征聚类,并采用混沌搜索对粒子群优化算法进行改进,提高其收敛速度和全局寻优能力,达到大数据优化聚类的目的。

1 云计算环境中大数据存储机制及数据结构分析

1.1 云计算环境中大数据存储机制体系构架

云计算是通过互联网来提供动态易扩展的大数据存储空间和结构模型。为了实现云计算环境中大数据存储聚类 and 分类挖掘,需要首先在云计算环境中构建大数据存储机制体系构架。云计算环境中大数据存储采用虚拟化存储池结构,云计算部署依赖于计算机集群,从上到下分别是: I/O 虚拟计算机, USB 接口层序和磁盘层,企业数据中心通过各种终端获取应用服务,使计算分布在大量的分布式计算机上^[7]。云计算环境中大数据存储总体架构如图 1 所示。

图 1 中,当所有的云计算虚拟机都被分配到物理机之后,利用下述公式能够计算本次聚类中的全局最优解^[8],并能根据最优解将全部云计算中的大数据特征聚类中心 V_{M_i} 分配到物理机 P_{M_i} 上:

$$N = \frac{1}{n} \sum_{j=1}^n |U_{i_j^{cpu}} - U_{i_{avg}^{cpu}}| + \frac{1}{n} \sum_{j=1}^n |U_{i_j^{mem}} - U_{i_{avg}^{mem}}| + \frac{1}{n} \sum_{j=1}^n |U_{i_j^{hw}} - U_{i_{avg}^{hw}}| \quad (1)$$

$$X = [x(t_0), x(t_0 + \Delta t), \cdots, x(t_0 + (K - 1)\Delta t)] = \begin{bmatrix} x(t_0) & x(t_0 + \Delta t) & \cdots & x(t_0 + (K - 1)\Delta t) \\ x(t_0 + J\Delta t) & x(t_0 + (J + 1)\Delta t) & \cdots & x(t_0 + (K - 1)\Delta t + J\Delta t) \\ \vdots & \vdots & \ddots & \vdots \\ x(t_0 + (N - 1)J\Delta t) & x(t_0 + (1 + (m - 1)J)\Delta t) & \cdots & x(t_0 + (N - 1)\Delta t) \end{bmatrix} \quad (4)$$

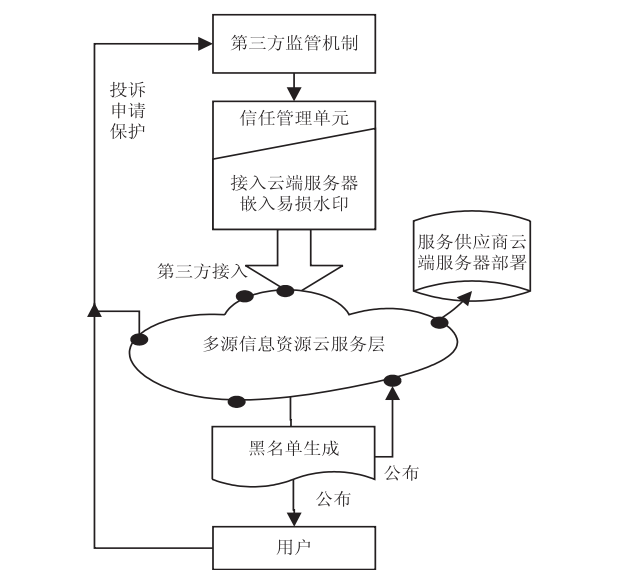


图 1 云计算环境中大数据存储总体架构

对样本进行分析采集,判断样本是否为典型样本,以此样本为数据,设大数据库信数据信息流样本 $S = \overline{X_1}, \overline{X_2}, \cdots, \overline{X_k}$, 分别在时间段 T_1, T_2, \cdots, T_K 进行数据信息采样。

现在把云计算环境中大数据集合 X 分为 c 类,其中 $1 < c < n$ 。把数据的分割转化为对空间的分割,得到大数据的存储结构中心矢量为:

$$V = \{v_{ij} \mid i = 1, 2, \cdots, c, j = 1, 2, \cdots, s\} \quad (2)$$

其中, V_i 为目标聚类特征的第 i 个矢量(第 i 个聚类中心矢量)。

模糊划分矩阵表示为:

$$U = \{\mu_{ik} \mid i = 1, 2, \cdots, c, k = 1, 2, \cdots, n\} \quad (3)$$

对单个数据源进行冗余数据降维处理,在进行多通道 QoS 需求的虚拟机分簇挖掘的过程中,其输入部分(为虚拟机和物理机的集合)以及相关参数分别为 $V_{MS} = \{V_{M_1}, V_{M_2}, \cdots, V_{M_n}\}$, $P_{MS} = \{P_{M_1}, P_{M_2}, \cdots, P_{M_n}\}$, 启发因子为 α , 启发因子的期望值为 β , 最大挖掘次数为 I_{max} 。由此,客户端上传的数据块提供固定大小的数据块,实现云聚类。通过上述的云计算环境中大数据存储机制体系构架分析,为进行大数据聚类提供准确的数据基础^[9]。

1.2 大数据信息流模型构建与特征提取

假设云计算环境中的信息流时间序列为 $\{x(t_0 + i\Delta t)\}, i = 0, 1, \cdots, N - 1$ 。设 X 和 Y 为属性集合,云计算环境下大数据聚类空间状态矢量表达式为:

式中, $\mathbf{x}(t)$ 为云计算环境下大数据聚类系统信息流时间序列; J 为云计算环境下大数据重构的相空间的时间窗函数; m 为目标聚类调节因子; Δt 为数据采样时间间隔。

计算大数据的离散样本频谱特征 $X_p(u)$, 主特征量为:

$$X_p(u) = s_c(t) e^{j2\pi f_d t} = \frac{1}{\sqrt{T}} \text{rect}\left(\frac{t}{T}\right) e^{j2\pi(f_d + Kt^2)/2} \quad (5)$$

其中, $s_c(t)$ 为大数据的特征标量时间序列; $e^{j2\pi f_d t}$ 为大数据聚类数据的离散样本中心。

数据集为 $\{X_1, X_2, \dots, X_n\}$, (F, Q) 为样本数据高阶贝塞尔函数统计量, 确定节点数据包的置信度, 确立置信区间, 得到的置信度和置信区间分别为:

$$z_{\langle i, d \rangle}^{k+1} = x_{r_1}^k + F * (x_{r_2}^k - x_{r_3}^k) \quad (6)$$

$$u_{\langle i, d \rangle}^{k+1} = \begin{cases} x_{id}^{k+1} & f_{\text{fitness}}^i < f_{\text{fitness}}^* \\ z_{\langle i, d \rangle}^{k+1} & f_{\text{fitness}}^i \geq f_{\text{fitness}}^* \end{cases} \quad (7)$$

数据聚类中心的粒子最优解的向量矩阵为:

$$\mathbf{\Sigma}_r = \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_r) \in \mathbf{R}^{r \times r} \quad (8)$$

其中, σ_r 为粒子在 $k+1$ 时刻的位置; $\mathbf{R}^{r \times r}$ 则为实矩阵。

对角向量可以表述为粒子距离目标解的远近, 并且满足:

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0 \quad (9)$$

基于误差反传的梯度下降训练, 实现对大数据的特征优化提取, 输入得数据聚类系统, 实现模式识别。

2 大数据聚类算法的改进实现

在上述对云计算中的大数据信息流模型进行构建与特征提取的基础上, 进行大数据聚类优化设计与实现。传统方法采用模糊 C 均值聚类算法进行云计算中的大数据聚类, 易陷入局部极值, 产生聚类偏移, 效果不好^[10]。文中提出一种基于粒子群优化 (PSO) 算法的大数据聚类算法。粒子群 (PSO) 优化算法由 Kennedy 和 Eberhart 于 1995 年提出, 是一种新型智能优化算法。利用粒子群算法进行云计算中的大数据聚类处理时, 由于每个个体有不一样的特征, 适应度高的个体更容易进入下一代, 由此可以优化聚类算法的实现效率。

假设在 D 维大数据聚类搜索空间中, 有 m 个粒子组成一个种群, 每个大数据信息特征矢量 \mathbf{X}_i 对应的一个函数为:

$$l_i(k) = (1 - \rho) l_i(k-1) + \gamma f(x_i(k)) \quad (10)$$

其中, f_i 是 \mathbf{X}_i 模因组适应度函数; $P_{ij}(k)$ 表示 k 时刻第 i 个粒子的全局优化粒子权值。

设置收敛阈值 N_{th} , 当 $N_{\text{eff}} < N_{\text{th}}$ 时, 第 j 个粒子移动

的概率为:

$$x_{k+1} = \sin(a/x_k), \quad -1 \leq x_k \leq 1, x_k \neq 0 \quad (11)$$

其中, x_k 为第 k 个动态惯性权重; a 为聚类中心的控制参量。

计算按最优聚类解的概率密度函数 $q(x_k^i/x_{k-1}^i)$, 根据模因组中的更新迭代顺序, 得到:

$$\mathbf{\Sigma}_\tau = \text{diag}(\max(\sigma_i - \tau, 0)) \quad (12)$$

根据不同数据聚类任务^[8], 调整适应度函数内权重, 得到 PSO 聚类的权重系数为:

$$\begin{cases} w = w(t) * w_{\text{start}} & k \geq \alpha \\ w = w(t) * \frac{1}{w_{\text{end}}} & k < \beta \end{cases} \quad (13)$$

其中, $\{\alpha, \beta\}$ 为云计算环境下大数据聚类的分集聚敛目标函数, 得到优化的 PSO 聚类目标函数为:

$$J_m(U, V) = \sum_{k=1}^n \sum_{i=1}^c \mu_{ik}^m (d_{ik})^2 x_i = x_{\text{imin}} + cx_i \cdot (x_{\text{imax}} - x_{\text{imin}}) \quad (14)$$

其中, 粒子的位置对应样本数据的 k 个聚类中心。除了粒子位置外, 对粒子的适应度和速度进行编码。由于样本数据的属性向量维数为 d , 则粒子的位置和速度为 $k \times d$ 维矩阵。

针对粒子群算法容易出现早熟并且收敛速度慢的缺陷^[11], 文中采用混沌映射方法对其进行优化, 带领粒子逃离局部最优解, 加速收敛。混沌搜索表面上显示出毫无规律的遍历, 然而它是凭借着其内在规则随机不重复地对系统中所有状态进行搜索遍历。混沌方法首先要生成混沌序列, 这里采取 Logistic 映射获得混沌序列, 可以通过如下方程进行描述:

$$Z_{n+1} = \mu Z_n (1 - Z_n) \quad (15)$$

在粒子群不断进行迭代计算的过程中, 超过一定代数, 其算法收敛速度便开始降低, 于是为了提高粒子群的收敛速度和全局寻优能力, 通过生成的混沌序列来扰动全局最优粒子。对于前述的 m 个粒子, 将它们的一维度一一映射到 $(0, 1)$ 范围上, 于是便能够得到向量 $\mathbf{D} = (d_1, d_2, \dots, d_m)$ 。其中, d_i 为粒子第 i 维, 其表达式为:

$$d_i = (\text{gbest}_i - a) / (b - a) \quad (16)$$

式中, gbest_i 为适应度最高粒子的第 i 维; a 和 b 分别为粒子在任意维度中的取值下限和上限。

利用混沌扰动重新进行迭代计算, 得到新序列:

$$Z_1 = (Z_{11}, Z_{12}, \dots, Z_{1m}) \quad (17)$$

把得到的新序列 Z_1 当成新粒子, 并进行适应度计算, 如果计算得到 Z_1 适应度高于之前搜索得到的最优解, 那么便令 Z_1 为当前最优解。

通过上述处理, 在云计算系统的大数据聚类中就

代表一个任务调度策略^[12]。改进的 PSO 大数据优化聚类算法流程描述如图 2 所示。

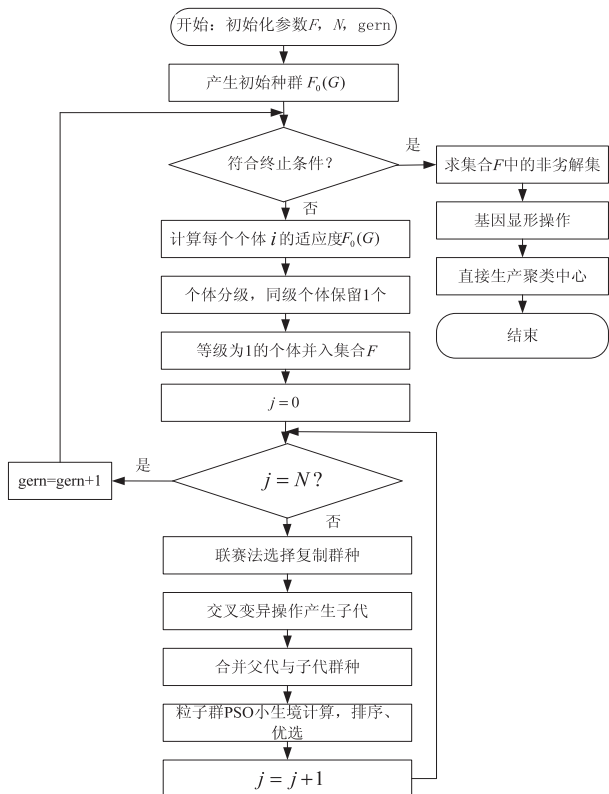


图 2 改进算法实现流程

3 仿真实验

为了验证文中算法在实现云计算环境中大数据优化聚类和数据挖掘中的性能,对其进行仿真实验。

仿真实验硬件环境为:处理器 Intel (R) Core (TM) 2 Duo CPU 主频 2.93 GHz,内存 2 GB;操作系统:Windows 7。仿真软件采用 Matlab 7。

实验中,大数据的采样频率 $f_s = 4f_0 = 20\text{ kHz}$ 。大数据聚类的时间中心 $t_0 = 15\text{ s}$,数据量从 10 MB 到 1 GB,以 10 MB 为单位,粒子群数量 N 为 30 984 个,粒子群聚类过程中的相空间搜索维度设置为 30,粒子移动的概率为 0.34,每次 PSO 运行迭代 5 000 次。大数据聚类的算法处理参数设置见表 1。

表 1 大数据聚类的算法处理参数设置

处理器数目	MIPS	内存/GB	带宽/dB	需求特点
2	400	2	200	网络型
2	600	2	200	储存型
2	400	4	200	计算型
2	400	2	400	普通型
4	400	2	400	储存型
4	600	2	400	网络型
4	400	4	400	计算型
4	200	2	800	存储型

根据上述仿真环境和参数设定结果,对云计算中

的大数据聚类进行仿真,其中大数据的特征分布如图 3 所示。

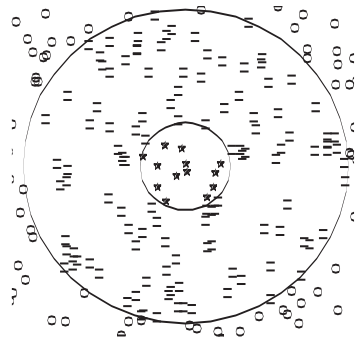


图 3 云计算中的大数据二维特征分布

由图 3 可见,原始的大数据二维特征分布具有随机性,在二维空间中难以实现对其规律性的特征提取和分类。采用文中算法进行特征提取和数据聚类处理,进行大数据的特征提取和信息模型构建,设计粒子群优化算法进行特征聚类,得到的特征提取结果如图 4 所示。

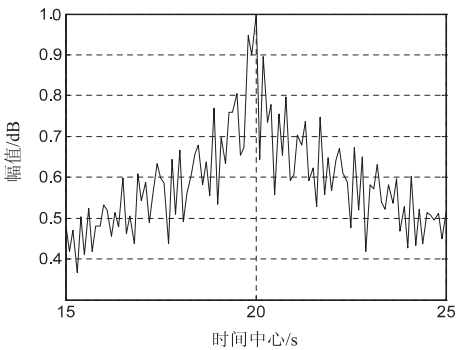


图 4 特征提取结果

由图 4 可见,文中算法能有效实现对云计算中的大数据的特征提取,波束的聚焦性能较好,为数据优化聚类提供准确的特征依据,以此为基础实现数据聚类。

采用不同算法分析大数据聚类的寻优性能,得到聚类中心寻优性能曲线如图 5 所示。

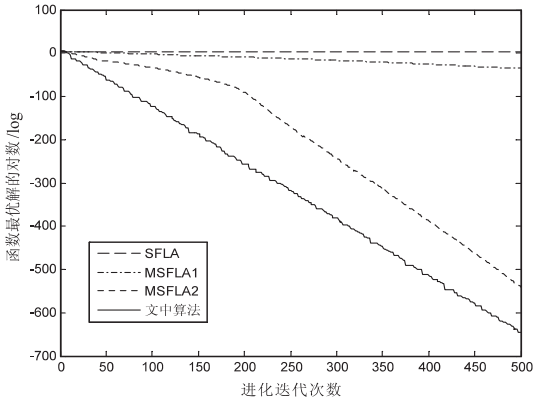


图 5 不同算法的寻优能力对比

由图 5 可见,文中算法在连续不断迭代的计算过程中,以稳定的收敛速度向最优解逼近,相比其他算法,具有明显的全局最优解搜寻优势和较好的收敛速

度,提高了数据聚类的寻优能力^[13],从而提高了大数据聚类精度,降低了误分率。通过定量分析可知,采用文中算法的误分率比传统算法降低了 13.56%,展示了较好的大数据聚类挖掘能力。

4 结束语

在云计算环境中,海量的大数据需要进行调度和访问,达到数据挖掘的目的。实现云计算中大数据挖掘的基础在于数据聚类,文中提出一种基于改进 PSO 算法的大数据聚类算法。首先分析了云计算环境中的大数据结构模型,进行大数据的特征提取和信息模型构建^[14],设计粒子群优化算法进行特征聚类,达到大数据优化聚类的目的。仿真结果表明,文中算法在提高云计算环境中的大数据聚类性能方面表现优异。通过文中算法进行数据聚类,降低了误分率,具有较好的寻优性能。

参考文献:

- [1] 谭鹏许,陈越,兰巨龙,等.用于云存储的安全容错编码[J].通信学报,2014,35(3):109-115.
- [2] 魏理豪,王甜,陈飞,等.基于层次分析法的信息系统实用化评价研究[J].科技通报,2014,30(2):143-145.
- [3] 吴涛,陈黎飞,郭躬德.优化子空间的高维聚类算法[J].计算机应用,2014,34(8):2279-2284.
- [4] 辛宇,杨静,汤楚衡,等.基于局部语义聚类的语义重

叠社区发现算法[J].计算机研究与发展,2015,52(7):1510-1521.

- [5] 许成鹏,朱志祥.一种基于云计算平台的数据库加密保护系统[J].电子设计工程,2015,23(19):97-100.
- [6] 陶新民,宋少宇,曹盼东,等.一种基于流形距离核的谱聚类算法[J].信息与控制,2012,41(3):307-313.
- [7] 刘少伟,孔令梅,任开军,等.云环境下优化科学工作流执行性能的两阶段数据放置与任务调度策略[J].计算机学报,2011,34(11):2121-2130.
- [8] 许丞,刘洪,谭良.Hadoop 云平台的一种新的任务调度和监控机制[J].计算机科学,2013,40(1):112-117.
- [9] 张洁.云计算环境下的数据存储保护机制研究与仿真[J].计算机仿真,2013,30(8):254-257.
- [10] 张彬桥.云环境下计算资源调度策略与仿真研究[J].计算机仿真,2013,30(11):392-395.
- [11] 王德政,申山宏,周宁宁.云计算环境下的数据存储[J].计算机技术与发展,2011,21(4):81-84.
- [12] Qin Z R, Wang G Y, Wu L Y, et al. A scalable rough set knowledge reduction algorithm[C]//Proceedings of rough sets and current trends in computing, [s. l.]:[s. n.], 2004:445-454.
- [13] Liao Lüchao, Jiang Xinhua, Zou Fumin, et al. A spectral clustering method for big trajectory data mining with latent semantic correlation[J]. Chinese Journal of Electronics, 2015, 43(5):956-964.
- [14] 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.

(上接第 177 页)

- design[J]. IEEE Transactions on Communications, 1980, 28(1):84-95.
- [2] 孙圣和,陆哲明.矢量量化技术及应用[M].北京:科学出版社,2002.
- [3] Shen F, Hasegawa O. An adaptive incremental LBG for vector quantization[J]. Neural Networks, 2006, 19:694-704.
- [4] Hagan M T, Demuth H B. 神经网络设计[M].戴葵,译.北京:机械工业出版社,2002.
- [5] Amerijckx C, Legaty J D, Verle-Ysen M. Image compression using self organizing maps[J]. Systems Analysis Model Simulation, 2003, 43(11):1529-1543.
- [6] Seo S, Oberayer K. Self organizing maps and clustering methods for matrix data[J]. Neural Networks, 2004, 17:1211-1230.
- [7] Lau K W, Yin H, Hubbard S. Kernel self-organizing maps for classification[J]. Neurocomputing, 2006, 69:2033-2040.
- [8] McAulie J D, Atlas L E, Rivera C. A comparison of the LBG algorithm and Kohonen neural network paradigm for image vector quantization[C]//Proc of ICASSP. [s. l.]:[s. n.], 1990:2293-2296.

- [9] Nasrabadi N M, King R A. Image coding using vector quantization: a review[J]. IEEE Transactions on Communications, 1988, 36(8):957-971.
- [10] Lancini R, Tubaro S. Adaptive vector quantization for picture coding using neural networks[J]. IEEE Transactions on Communications, 1995, 43(2):534-544.
- [11] 王茂芝,徐文哲. LBG 算法对初始码书敏感的实验性能分析[J].物探化探计算技术,2004,26(4):375-378.
- [12] Huang H, Chen S H. Fast encoding algorithm for VQ-based image coding[J]. Electronics Letters, 1990, 26:1618-1619.
- [13] Ra S W, Kim J K. A fast mean-distance-ordered partial codebook search algorithm for image vector quantization[J]. IEEE Transactions on Circuits and Systems-II: Analog and Digital Signal Processing, 1993, 40(9):576-579.
- [14] Chang C C, Chang R F, Lee W T, et al. Fast algorithms for vector quantization[J]. Journal of Information Science and Engineering, 1996, 12(4):593-602.
- [15] Chang C C, Lee W T, Chen T S. Two improved codebook search methods of vector quantization based on orthogonal checking and fixed range search[J]. Journal of Electronic Imaging Representation, 1997, 8(1):27-37.