Journal of Computer Applications
计算机应用,2013,33(5):1330 - 1333
ISSN 1001-9081
CODEN JYIIDU
2013-05-01
http://www.joca.cn

文章编号:1001 - 9081(2013)05 - 1330 - 04
doi:10.3724/SP.J.1087.2013.01330

基于广义内容概率潜在语义分析模型的推荐

张　伟[1*],黄　炜[2],夏利民[1]
(1.中南大学 信息科学与工程学院,长沙 410075;　2.长沙航空职业技术学院 计算机系,长沙 410124)
(*通信作者电子邮箱 360706711@qq.com)

摘　要:针对推荐系统中存在新项目及准确性难以把握等问题,提出一种基于广义内容概率潜在语义模型的推荐方法。该方法以概率潜在语义模型为基础,引入两组潜在变量及项目特征来建立广义内容概率潜在语义模型。该模型中两组潜在变量分别表示用户群体和项目群体,项目特征根据实际情况以特征词的形式进行表示,且通过不对称学习算法完成未知参数的训练及预测。利用三个不同的数据集对所提方法进行实验验证,结果表明该方法具有良好的项目推荐品质。

关键词:概率潜在语义;项目特征;最大期望算法;潜在变量;项目推荐
中图分类号:　　**文献标志码**:A

Recommendation research based on general content probabilistic latent semantic analysis model

ZHANG Wei[1*], HUANG Wei[2], XIA Limin[1]
(1. *School of Information Science and Engineering, Central South University, Changsha Hunan* 410075, *China*;
2. *Department of Computer Engineering, Changsha Aeronautical Vocational and Technical College, Changsha Hunan* 410124, *China*)

Abstract: In the recommendation system, some new items and the accuracy issue cannot be well controlled. Therefore, a new recommendation method based on general content Probabilitistic Latent Semantic Analysis (PLSA) model was proposed. The general content PLSA model contained two latent variables indicating the user groups and item groups, and contained features of items that were trained by asymmetric learning algorithm. The experimental results show that the new method has good quality for recommendation on three different data sets.

Key words: Probabilitistic Latent Semantic Analysis (PLSA); item feature; Expectation-Maximization (EM) algorithm; latent variable; item recommendation

0　引言

随着互联网技术的发展及普及,如何根据需求特征将符合消费期望的信息自动推荐给用户,并为其提供中长期的意向资讯,已成为基于互联网应用的重要技术发展方向之一。信息推荐技术的本质在一定程度上可理解为信息的主动协同过滤技术,该技术可以分为基于内存的协同过滤和基于模型的协同过滤。在基于内存的协同过滤技术方面,其主要研究内容集中于基于用户的协同过滤和基于项目的协同过滤。例如,Kim 等[1]提出了基于协同标签的协同过滤方法来增强推荐的质量;Lee 等[2]提出了两种方法共同预测的协同过滤技术提高了推荐的准确性,并且该方法对稀疏的数据具有鲁棒性。在基于模型的协同过滤研究方面,其模型主要包括贝叶斯网络[3]、奇异值分解[4]、潜在语义分析(Latent Semantic Analysis, LSA)以及概率潜在语义分析(Probabilitistic Latent Semantic Analysis, PLSA)[5-6]等。其中,PLSA 作为协同过滤的方法[6]不仅具有 LSA 方法的优良品性,同时具有良好的概率理论基础,因此能较好满足项目推荐的信息过滤要求,但仍然存在某些不足,如在噪声干扰或在小训练样本的情况下,PLSA 可能出现过拟合[7],从而影响项目推荐的准确性,并同时存在不能对新项目进行预测推荐的问题。

基于 PLSA 方法存在的问题,提出了基于广义内容 PLSA 模型的推荐方法。该方法以 PLSA 模型为基础,引入两个潜在变量,即用户群体和项目群体,同时引入项目的特征。广义内容 PLSA 模型的训练采用不对称学习的方法。实验分别利用三个不同的数据集进行,并与其他方法进行实验对比。结果表明,所提方法具有良好的项目推荐品质。

1　PLSA 模型

Hofmann 首先提出了 PLSA 模型,并应用于信息抽取及推荐系统的协同过滤领域[6],其原理如图 1 所示。其中:u 表示用户,d 表示项目,z 表示隐含的用户群体,r 表示用户群体 z 对项目 d 的评分,则用户、项目和评分可以组成一个三元向量 $\langle u,d,r\rangle$;未知参数 $p(z_k|u)$ 表示用户 u 属于群体 z_k 的概率,且属于各群体概率的和为 1,其中 z_k 可采用年龄、性别、喜好等标准对 z 进行划分获得;$p(r|z_k,d)$ 表示群体 z_k 对项目 d 评分为 r 的概率。

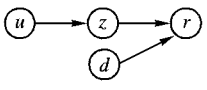

图 1　协同过滤中的 PLSA 模型

则用户和商品的联合概率为:

收稿日期:2012-11-01;**修回日期**:2013-01-01。
基金项目:教育部博士点基金资助项目(20080533005 9,20090162110057);湖南省科技计划项目(2011GK3213)。
作者简介:张伟(1972 -),男,黑龙江哈尔滨人,博士研究生,主要研究方向:模式识别、计算视觉;　黄炜(1972 -),女,湖南长沙人,讲师,硕士,主要研究方向:数据挖掘;　夏利民(1963 -),男,湖南长沙人,教授,博士生导师,博士,主要研究方向:模式识别、计算视觉。

$$p(r|u,d) = \sum_z p(z|u)p(r|z,d) \tag{1}$$

Hofmann 又进一步将群体 z 对项目 d 评分 r 的条件概率 $p(r|z,d)$ 假定为高斯分布[6],即 $p(r|z,d) = N(\mu_{zd},\sigma_{zd}) = p(r;\mu_{zd},\sigma_{zd})$,则用户和项目的联合概率是一个高斯混合模型:

$$p(r|u,d) = \sum_z p(z|u)p(r;\mu_{zd},\sigma_{zd}) \tag{2}$$

为了求解式(2),对于参数 $P(z|u)$,μ_{zd},σ_{zd} 的估计采用最大期望(Expectation-Maximization, EM)算法,其中对数似然函数为:

$$L(\theta) = -\sum_{\langle u,d,r,z\rangle}[\lg p(r,d|z) + \lg p(z|u)] \tag{3}$$

对于每个向量 $\langle u,d,r\rangle$,$p(z|u,d,r)$ 是未知的,由此定义:

$$R(\theta,\hat{\theta}) = -\sum_{\langle u,d,r\rangle}\sum_z p(z|u,d,r)[\lg p(r,d|z) + \lg p(z|u)] \tag{4}$$

将参数 $p(z|u)$、μ_{zd} 和 σ_{zd} 进行随机初始化(如 $P(z|u)$,$\mu_{zd},\sigma_{zd} \in (0,1]$),然后交替实施下列 E 阶段和 M 阶段,通过迭代获得参数 $P(z|u)$,μ_{zd},σ_{zd} 的估计值。

E 阶段:对每一个评分向量$\langle u,d,r\rangle$,计算每个潜在变量 z 的后验概率 $p(z|u,d,r)$:

$$p(z|u,d,r) = \frac{p(z|u)p(r;\mu_{zd},\sigma_{zd})}{\sum_z p(z|u)p(r;\mu_{zd},\sigma_{zd})} \tag{5}$$

M 阶段:结合 E 阶段,对式(4)期望函数 $R(\theta,\hat{\theta})$ 进行拉格朗日最优化极值求解,分别求其偏导数得到:

$$p(z|u) = \frac{\sum_{\langle u',d,r\rangle;u'=u} p(z|u,d,r)}{\sum_{z'}\sum_{\langle u',d,r\rangle;u'=u} p(z'|u,d,r)} \tag{6}$$

$$\mu_{zd} = \frac{\sum_{\langle u,d',r\rangle;d'=d} rp(z|u,d,r)}{\sum_{\langle u,d',r\rangle;d'=d} p(z|u,d,r)} \tag{7}$$

$$\sigma_{zd} = \frac{\sum_{\langle u,d',r\rangle;d'=d} (r-u_{zd})^2 p(z|u,d,r)}{\sum_{\langle u,d',r\rangle;d'=d} p(z|u,d,r)} \tag{8}$$

交替执行 E、M 阶段,直到 $R(\theta^{new},\theta) - R(\theta,\theta)$ 小于给定值而收敛,迭代结束,得到模型各参数。最后,通过模型的计算可得到用户 u 对商品 d 的预测评分:

$$R_{u,d} = \sum_z p(z|u)u_{zd} \tag{9}$$

传统的协同过滤技术不能对新项目进行预测,同时对于推荐结果的准确性也有待提高。因此本文提出基于广义内容 PLSA(General Content Probabilitistic Latent Semantic Analysis, GC-PLSA)模型的推荐方法。该方法通过考虑项目本身的属性,可以有效地解决新项目问题,且明显提高推荐结果的准确性。

2　基于 GC-PLSA 模型的推荐

GC-PLSA 模型以 PLSA 模型为基础,引入用户组和项目组两个潜在变量,同时引入项目的特征。为此首先介绍广义内容 PLSA 模型及其训练和预测过程,然后给出推荐算法的一般流程。对于模型中所使用的项目特征,根据推荐项目的种类所提取的特征有所不同,但最终表示方法均是以词的形式来描述的。

2.1　广义内容 PLSA 模型

本文给出的 GC-PLSA 模型原理如图 2 所示,其中 u 为用户,d 为项目,Z_U 为潜在用户群体,Z_D 为潜在项目群体,r 为评分等级,w 为项目特征词,$p(Z_U|u)$ 为用户 u 属于用户组的概率,$p(Z_D|d)$ 为项目 d 属于项目组的概率,$p(r|Z_U,Z_D)$ 为用户组和项目组评分为 r 的概率,$p(w|Z_D)$ 为项目组包含特征词 w 的概率。

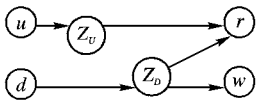

图 2　GC-PLSA 模型原理

不对称学习算法能在潜在空间的定义中更好地控制各个模态数据产生的影响,为此,采用不对称的学习算法[8]估计未知的模型参数,其建模和预测推荐过程如图 3 所示。在建模阶段:1)根据用户和项目的评分矩阵,采用 EM 算法训练出模型参数 $p(Z_U|u)$、$p(Z_D|d)$ 和 $p(r|Z_U,Z_D)$;2)根据项目特征的表示方法,将每一项目用特征词进行表示;3)根据得到的模型参数 $p(Z_D|d)$ 和项目特征,采用 folding-in 算法[9]训练出参数 $p(w|Z_D)$。

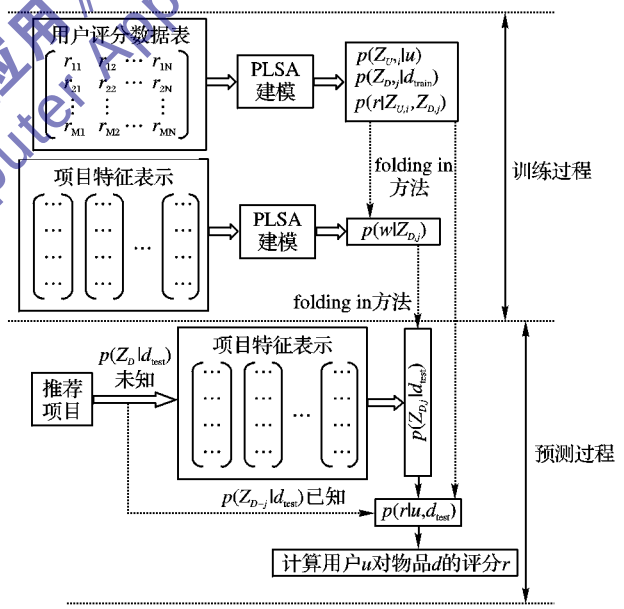

图 3　GC-PLSA 模型的学习和预测过程

在预测阶段,模型参数 $p(Z_U|u)$、$p(Z_D|d)$、$p(r|Z_U,Z_D)$ 和 $p(w|Z_D)$ 已经使用不对称学习算法估计得到;对于推荐项目,若其在项目群体中的主题分布 $p(Z_D|d)$ 已得到,则对用户 u 推荐项目 d 的等级为 r 的概率可直接用式(10)计算;若其在项目群体中的主题分布 $p(Z_D|d)$ 未知,而其特征可以自动提取得到,即可获得该项目的特征词描述,则相应的主题分布 $p(Z_D|d)$ 可以使用 folding-in 算法推出,然后对用户 u 推荐项目 d 的等级为 r 的概率也可用式(10)计算:

$$p(r|u,d) = \sum_{Z_U}\sum_{Z_D} p(Z_D|u)p(Z_D|d)p(r|Z_D,Z_U) \tag{10}$$

最终根据预测得到的评分等级,可以将前 N 个项目推荐给用户。

2.2 推荐算法流程

根据图 3 所给出的 GC-PLSA 模型的学习和预测过程，确定推荐算法的具体步骤如下，其中输入为用户评分矩阵和项目，输出结果为预测等级排序的前 N 个项目。

1）根据项目特征提取方法，提取所有项目的特征，并将项目用特征词进行表示；

2）根据用户评分矩阵和提取项目特征向量训练模型参数；

3）根据训练好的模型预测用户对项目的评分等级；

4）根据预测的评分等级选出前 N 个项目作为推荐结果。

3 实验

3.1 实验环境、实验数据和度量准则

实验所用 PC 机的配置为 Intel Pentium 4 2.66 GHz CPU；4 GB RAM；操作系统为 Windows XP；编程环境为 VC 6.0。

为验证本文方法的推荐效果，实验采用三个不同的数据集分别进行验证。实验数据集 1 为文档评分数据集，这些文章均来自 CiteULike，包括 5 551 位用户、16 980 篇文章以及 204 986 组评分数据。对所有文章，首先提取其标题和摘要，然后去掉停顿词并采用 tf-idf 的方法[10] 来选取 8 000 个区别词构成词典；那么每一篇文档便可以用词典中的词进行表示。实验数据集 2 为电影评分数据集（Grouplen 提供的 MovieLens 数据集），包括 943 位用户、1 682 部电影和 100 000 组评分数据。对所有电影，首先提取其文本描述内容（包括电影类别、导演、演员等），然后选取 2 400 个区别词作为词典，那么每一部电影便可使用词典中的词进行表示。实验数据集 3 为自建的图片评分数据集，其中实验图片为自然风景图片且数量为 2 300 张，实验用户为中南大学的学生共 420 名；学生定期（每两天）为给定的 3 张物品图像按照 5 分制进行打分，如此持续三个月共得到 56 700 组评分数据。对所有图片，首先提取方块特征和局部不变特征，然后采用 K 均值聚类算法进行量化，最后运用 BOW（Bag-Of Words）原理可以得到一个视觉词集合；那么每一幅图片便可用视觉词进行表示[11-14]，该过程如图 4 所示。

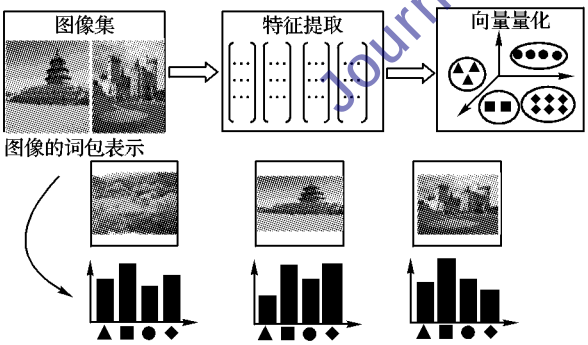

图 4 图像的词包表示

对于推荐质量的度量，实验采用统计精度度量方法中的平均绝对误差（Mean Absolute Error，MAE）。平均绝对误差通过计算预测的用户评分与实际用户评分之间偏差来度量预测的准确性，MAE 值越小，推荐准确度越高。该实验将推荐度作为预测等级进行 MAE 值的统计。MAE 计算如下所示：

$$MAE = \left(\sum_{i=1}^{N} | p_i - q_i | \right) / N \tag{11}$$

其中：p_i 为预测的用户的第 i 项评分，q_i 为实际用户的第 i 项评分，N 为测试物品的数量。

3.2 实验结果及分析

为验证 GC-PLSA 方法的有效性，与 G-PLSA 方法[6] 和基于项目的协同过滤方法[15]（Item-based Collaborative Filtering，IBCF）分别从新项目问题和准确性两个方面进行比较分析。

对于新项目问题，通过三种方法分别在三个数据集上计算 *MAE*，并观察 *MAE* 与项目评分用户数量的关系。实验过程为：1）从数据集合中随机抽取一个项目 i，对于项目 i，从对其评分的用户集合中随机抽取 N 位作为其评分用户集合，其余用户作为未评分用户集合；2）训练模型并预测将项目 i 推荐给其未评分用户集合中用户的等级；3）重复 1）和 2）共 M 次；4）根据前三步骤统计 *MAE*。实验结果如图 5 所示。

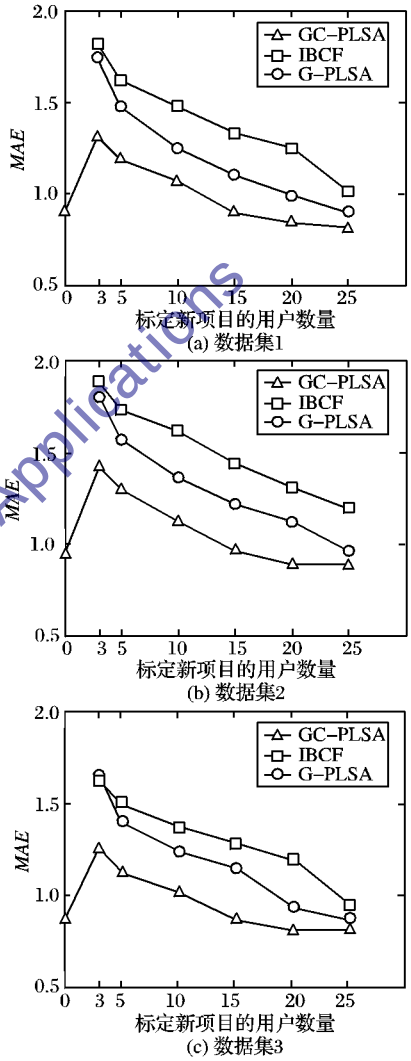

图 5 *MAE* 与项目评分用户数量关系对比

对于准确性问题，实验过程为：1）将实验数据按 4∶1 的比例分为训练集和测试集；2）根据训练集进行模型训练，得到个模型参数；3）预测测试集中的评分等级；4）根据式（11）统计 *MAE*。统计三种方法的平均 *MAE* 值如表 1 所示。

表 1 三种方法在三个数据集中的平均 *MAE*

数据集序号	GC-PLSA	G-PLSA	IBCF
1	0.74	0.77	0.80
2	0.72	0.76	0.78
3	0.76	0.79	0.83

根据实验结果可知：1）GC-PLSA 方法在新项目推荐品质上明显优于其他两种方法，尤其在没有任何评分记录的情况下也能进行准确的预测推荐，因为此时项目在群体中的分布是根据其特征获得而非评分数据；2）其他两种方法在评分用

户数量为 0 时无法进行预测,且在评分用户数量相同时其 *MAE* 值也明显高于 GC-PLSA 方法;3)通过表 1 可知,GC-PLSA 方法的平均 *MAE* 值明显低于其他两种方法,GC-PLSA 方法具有良好的项目推荐品质。

4　结语

针对 PLSA 模型中存在的新项目难以推荐及推荐准确性较低等问题,引入了用户组和项目组两个潜在变量以及项目的特征,提出一种基于广义内容的推荐方法 GC-PLSA,并通过实验与 G-PLSA、IBCF 等方法进行了对比,分别从新项目和准确性两个方面说明该方法具有良好的物品推荐品质。

参考文献:

[1] KIM H-N, JI A-T, HA I, *et al*. Collaborative filtering based on collaborative tagging for enhancing the quality of recommendation[J]. Electronic Commerce Research and Applications, 2010, 9(1): 73 – 83.

[2] LEE J-S, OLAFSSON S. Two-way cooperative prediction for collaborative filtering recommendations[J]. Expert Systems with Applications, 2009, 36(3): 5353 – 5361.

[3] WANG K B, TAN Y. A new collaborative filtering recommendation approach based on naive Bayesian method[C]// Proceedings of the Second International Conference on Advances in Swarm Intelligence. Berlin: Springer, 2011: 218 – 227.

[4] 杨阳, 向阳, 熊磊. 基于矩阵分解与用户近邻模型的协同过滤推荐算法[J]. 计算机应用, 2012, 32(2): 395 – 398.

[5] 宋晓雷, 王素格, 李红霞. 基于概率潜在语义分析的词汇情感倾向判别[J]. 中文信息学报, 2011, 25(2): 89 – 93.

[6] HOFMANN T. Collaborative filtering via Gaussion probilistic lantent semantic analysis [C] // Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2003: 259 – 266

[7] 张玉芳, 朱俊, 熊忠阳. 改进的概率潜在语义分析下的文本聚类算法[J]. 计算机应用, 2011, 31(3): 674 – 676.

[8] MONAY F, GATICA-PEREZ D. Modeling semantic aspects for cross-media image indexing [J]. IEEE Transactions on Pattern Anglysis and Machine Intelligence, 2007, 29(10): 1802 – 1817.

[9] WANG X, JIN X M. Understanding and enhancing the folding-in method in latent semantic indexing[C] // Proceedings of the 17th International Conference on Database and Expert Systems Applications, LNCS 4080. Berlin: Springer, 2006: 104 – 113.

[10] BLEI D M, LAFFERTY J D. Dynamic topic models [C] // ICML'06: Proceedings of the 23rd International Conference on Machine Learning. Washington, DC: IEEE Computer Society, 2006: 113 – 120.

[11] FENG S L, MANMATHA R, LAVRENKO V. Multiple Bernoulli relevance models for image and video annotation[C] // Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2004: 1002 – 1009.

[12] HUANG J, KUMAR S R, MITRA M. Spatial color indexing and applications[J]. International Journal of Computer Vision, 1999, 35(3): 245 – 268.

[13] MANJUNATH B S, MA W Y. Texture features for browsing and retrieval of image data[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18(8): 837 – 842.

[14] LOWE D C. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91 – 110.

[15] MIRANDA C, JORGE A M. Item-based and user-based incremental collaborative filtering for Web recommendations[C]// Proceedings of the 14th Portuguese Conference on Artificial Intelligence. Berlin: Springer-Verlag, 2009: 673 – 684.

(上接第 1316 页)

综上可知,本文提出的 ACFOA 总体来说比 FOA 以及权威文献中算法具有更快的收敛速度、更高的收敛精度和收敛可靠性。

表 6　基于 f_3 的算法时间比较　　s

f_3 维数	FOA	ACFOA	PSO[14]	GA[14]	BFO[14]
30	0.348	2.024	1.203	21.563	2.484
50	0.423	2.516	1.360	32.938	2.719

5　结语

本文将混沌算法融入基本果蝇优化算法,提出自适应混沌果蝇优化算法。通过群体适应度方差,判定果蝇优化算法处于局部收敛状态时,利用混沌算法进行全局寻优,从而跳出局部极值,提高了果蝇优化算法的收敛精度和收敛速度。6 个基准测试函数的对比实验结果表明新算法具有更好的全局搜索能力,在收敛速度、收敛可靠性及收敛精度上均比基本果蝇优化算法有很大的提高。

参考文献:

[1] PAN W T. A new fruit fly optimization algorithm: Taking the financial distress model as an example[J]. Knowledge-Based Systems, 2012, 26: 69 – 74.

[2] 潘文超. 果蝇最佳化演算法[M]. 台北: 沧海书局, 2011.

[3] 胡旺, 李志蜀. 一种更简化而高效的粒子群优化算法[J]. 软件学报, 2007, 18(4): 862 – 863.

[4] 相征, 张太镒, 孙建成. 基于混沌吸引子的快衰落信道预测算法[J]. 西安电子科技大学学报, 2006, 33(1): 145 – 149.

[5] 吕晓明, 黄考利, 连光耀. 基于混沌粒子群优化的系统级故障诊断策略优化[J]. 系统工程与电子技术, 2010, 32(1): 217 – 220.

[6] 杨俊杰, 周建中, 喻菁, 等. 混合混沌优化方法及其在非线性规划问题中的应用[J]. 计算机应用, 2004, 24(10): 119 – 120.

[7] 刘道华, 原思聪, 兰洋, 等. 混沌映射的粒子群优化方法[J]. 西安电子科技大学学报, 2010, 37(4): 764 – 769.

[8] 张劲松, 李歧强, 王朝霞. 基于混沌搜索的混和粒子群优化算法[J]. 山东大学学报, 2007, 37(1): 48 – 50.

[9] 莫愿斌, 陈德钊, 胡上序. 混沌粒子群算法及其在生化过程动态优化中的应用[J]. 化工学报, 2006, 57(9): 2123 – 2127.

[10] 王凌. 智能优化算法及其应用[M]. 北京: 清华大学出版社, 2001: 148 – 149.

[11] 林川, 冯全源. 一种新的自适应粒子群优化算法[J]. 计算机工程, 2008, 34(7): 181 – 183.

[12] 王联国, 洪毅, 施秋红. 全局版人工鱼群算法[J]. 系统仿真学报, 2009, 21(23): 7483 – 7486.

[13] 逄珊, 杨欣毅, 张小峰. 混沌映射的多种群量子粒子群优化算法[J/OL]. [2011 – 11 – 14]. http://cnki.net/kcms/detail/11.2127.TP.20111114.0947.044.html.

[14] 胡洁. 细菌觅食优化算法的改进及应用研究[D]. 武汉: 武汉大学, 2012.

[15] 杨萍, 孙延明, 刘小龙, 等. 基于细菌觅食趋化算子的 PSO 算法[J]. 计算机应用研究, 2011, 28(10): 3640 – 3642.