DOI：10.13203/j.whugis20130094 文章编号：1671-8860(2015)02-0238-05

基于最小平方和残差的高阶模糊联合聚类算法

黄少滨[1] 杨欣欣[1]

1 哈尔滨工程大学计算机科学与技术学院，黑龙江 哈尔滨，150001

摘 要：目前，多数高阶联合聚类算法属于硬划分方法，不考虑聚簇重叠问题。为了更有效地分析具有重叠聚簇结构的数据，提出了一种基于最小平方和残差的高阶模糊联合聚类算法(MSR-HFCC)，该算法将聚类问题转化为最小化模糊平方和残差的优化问题，推导出求解优化问题的隶属度迭代更新公式，设计出聚类过程的迭代算法。实验结果表明，MSR-HFCC 算法聚类效果优于目前已有的 5 种硬划分高阶联合聚类算法。

关键词：高阶异构数据；联合聚类；模糊聚类；平方和残差

中图法分类号：P207；TP181 **文献标志码**：A

近年来，现代信息系统产生了大量包含多种类型数据的数据集，并且不同类型数据之间相互关联[1,2]。如在论文出版系统中包含 4 种类型的数据[1,2]，分别是作者、论文、会议和关键词。将这种相互关联的多种类型数据称为高阶异构数据[3,4]。为了有效挖掘高阶异构数据隐藏的模式，有学者提出了高阶异构数据联合聚类（简称为高阶联合聚类）方法[3,4]。已有的高阶联合聚类算法包括基于图划分的方法[3,5,6]，主要有 CBGC[3] 和 CIHC[6]；基于信息论的方法[4,7]，如 CIT[4] 和 AD-HOCC[7]；基于 k 部图学习的方法[8,9]，包括 RSN[8] 和 SKGC[9]；基于矩阵分解的方法[1,2,10,11]，代表算法有 SRC[10] 和 SS-NMF[11]；基于 Goodman Kruskal 的方法 CoStar[12]；基于排名的聚类方法，代表算法有 NetClus[13]。这些聚类算法都是硬划分方法，一个数据要么“属于”或“不属于”一个聚簇。然而在实际应用中，有些数据同时属于多个聚簇，聚簇结构之间存在重叠的部分，硬划分方法并没有考虑这种聚簇重叠问题。为此，本文提出了一种基于最小平方和残差的高阶模糊联合聚类算法（minimum sum-squared residue for high-order fuzzy co-clustering，MSR-HFCC）。MSR-HFCC 算法利用平方和残差衡量不同聚簇数据之间关系的相异度[14,15]，评估聚类质量。为了更好地描述重叠聚簇的聚类结果[16,17]，在平方和残差中引入模糊概念，将聚类问题转化为最小化模糊平方和残差的优化问题。

1 MSR-HFCC 算法

1.1 MSR-HFCC 算法目标函数

由 $X^1=\{x_1^1,\cdots,x_{n_1}^1\}$，$X^2=\{x_1^2,\cdots,x_{n_2}^2\}$，$\cdots$，$X^m=\{x_1^m,\cdots,x_{n_m}^m\}$ 共 m 种类型数据组成高阶异构数据，$n_i(1\leqslant i\leqslant m)$ 表示第 i 种类型数据 X^i 的数据个数。X^i 与 X^j 的关系矩阵是 $\boldsymbol{D}^{(ij)}=(d_{pq}^{(ij)})_{n_i\times n_j}$，$p$ 行 q 列元素 $d_{pq}^{(ij)}$ 表示 x_p^i 与 x_q^j 之间的关系强度。设 $X^i(1\leqslant i\leqslant m)$ 划分为 K_i 个聚簇，残差 $h_{pqk_ik_j}^{(ij)}$ 是一种衡量关系 $d_{pq}^{(ij)}$ 与 x_p^i 和 x_q^j 所在聚簇内所有数据之间的关系相异度的指标[14,15]。高阶联合聚类的目标是寻找每种类型数据最优的划分，使得不同聚簇内数据之间的关系相异度最低。可见，高阶联合聚类问题可以转化为不同聚簇内的数据之间关系残差 $h_{pqk_ik_j}^{(ij)}$ 的平方和最小化问题。另外，MSR-HFCC 算法利用隶属度描述对象属于某聚簇的程度。其目标函数定义如下：

$$J=\sum_{1\leqslant i<j\leqslant m}\beta_{ij}\sum_{k_i=1}^{K_i}\sum_{k_j=1}^{K_j}\sum_{p=1}^{n_i}\sum_{q=1}^{n_j}u_{k_ip}^{(i)}u_{k_jq}^{(j)}(h_{pqk_ik_j}^{(ij)})^2+\sum_{i=1}^{m}T_u^{(i)}\left(\sum_{k_i=1}^{K_i}\sum_{p=1}^{n_i}(u_{k_ip}^{(i)})^2\right)\tag{1}$$

式中，β_{ij} 为数据 $\{X^i,X^j\}$ 之间关系的权值，满足

收稿日期：2013-04-24

项目来源：国家自然科学基金资助项目(71272216，60903080)；博士后科学基金资助项目(2012M5100480)；国家科技支撑计划资助项目(2009BAH42B02，2012BAH08B02)；中央高校基本科研业务费专项基金资助项目(HEUCFZ1212，HEUCF100603)。

第一作者：黄少滨，教授，博士，现从事数据挖掘、复杂网络和社会网络方面的研究。E-mail：huangshaobin_hrbeu@126.com

$\sum_{1\leqslant i<j\leqslant m}^{m}\beta_{ij}=1$;隶属度 $u_{k_i p}^{(i)}$ 满足如下限制条件:

$$\sum_{k_i=1}^{K_i}u_{k_i p}^{(i)}=1,1\leqslant p\leqslant n_i,1\leqslant i\leqslant m \quad (2)$$

数据项 $\sum_{i=1}^{m}T_u^{(i)}(\sum_{k_i=1}^{K_i}\sum_{p=1}^{n_i}(u_{k_i p}^{(i)})^2)$ 为模糊项;$T_u^{(i)}$ 为模糊度参数,用于调节隶属度 $u_{k_i p}^{(i)}$ 的模糊程度;关系强度 $d_{pq}^{(ij)}$ 的残差 $h_{pqk_i k_j}^{(ij)}$ 有如下两种定义方式[14,15]:

$$h_{pqk_i k_j}^{(ij)}=d_{pq}^{(ij)}-a_{k_i k_j}^{(ij)} \quad (3)$$

$$h_{pqk_i k_j}^{(ij)}=d_{pq}^{(ij)}-a_{pk_j}^{(ij)}-a_{k_i q}^{(ij)}+a_{k_i k_j}^{(ij)} \quad (4)$$

其中,

$$a_{k_i k_j}^{(ij)}=\sum_{p=1}^{n_i}\sum_{q=1}^{n_j}u_{k_i p}^{(i)}u_{k_j q}^{(j)}d_{pq}^{(ij)}/(\sum_{p=1}^{n_i}\sum_{q=1}^{n_j}u_{k_i p}^{(i)}u_{k_j q}^{(j)}) \quad (5)$$

$$a_{pk_j}^{(ij)}=\sum_{q=1}^{n_j}u_{k_j q}^{(j)}d_{pq}^{(ij)}/\sum_{q=1}^{n_j}u_{k_j q}^{(j)} \quad (6)$$

$$a_{k_i q}^{(ij)}=\sum_{p=1}^{n_i}u_{k_i p}^{(i)}d_{pq}^{(ij)}/\sum_{p=1}^{n_i}u_{k_i p}^{(i)} \quad (7)$$

1.2　迭代更新规则

通过拉格朗日乘子法求解 $u_{k_i p}^{(i)}$ 的更新规则,构造如下拉格朗日函数:

$$L=\sum_{1\leqslant i<j\leqslant m}\beta_{ij}\sum_{k_i=1}^{K_i}\sum_{k_j=1}^{K_j}\sum_{p=1}^{n_i}\sum_{q=1}^{n_j}u_{k_i p}^{(i)}u_{k_j q}^{(j)}(h_{pqk_i k_j}^{(ij)})^2+\sum_{i=1}^{m}T_u^{(i)}(\sum_{k_i=1}^{K_i}\sum_{p=1}^{n_i}(u_{k_i p}^{(i)})^2)+\sum_{i=1}^{m}(\sum_{p=1}^{n_i}\lambda_p^{(i)}(\sum_{k_i=1}^{K_i}u_{k_i p}^{(i)}-1)) \quad (8)$$

其中,$\lambda_p^{(i)}$ 为对应限制条件(2)的拉格朗日乘子。由拉格朗日最优解的必要条件,分别求 L 关于 $\lambda_p^{(i)}$ 和 $u_{k_i p}^{(i)}$ 的偏导数,并令其为零,则:

$$\frac{\partial L}{\partial\lambda_p^{(i)}}=\sum_{k_i=1}^{K_i}u_{k_i p}^{(i)}-1=0 \quad (9)$$

$$\frac{\partial L}{\partial u_{k_i p}^{(i)}}=\sum_{j=1,j\neq i}^{m}\beta_{ij}\sum_{k_j=1}^{K_j}\sum_{q=1}^{n_j}u_{k_j q}^{(j)}(h_{pqk_i k_j}^{(ij)})^2+2T_u^{(i)}u_{k_i p}^{(i)}+\lambda_p^{(i)}=0 \quad (10)$$

结合式(9)和式(10)可得 $u_{k_i p}^{(i)}$ 的更新迭代规则:

$$u_{k_i p}^{(i)}=\frac{1}{K_i}+\frac{1}{2T_u^{(i)}}(\frac{1}{K_i}\sum_{j=1,j\neq i}^{m}\beta_{ij}\sum_{k_i=1}^{K_i}\sum_{k_j=1}^{K_j}\sum_{q=1}^{n_j}u_{k_j q}^{(j)}(h_{pqk_i k_j}^{(ij)})^2-\sum_{j=1,j\neq i}^{m}\beta_{ij}\sum_{k_j=1}^{K_j}\sum_{q=1}^{n_j}u_{k_j q}^{(j)}(h_{pqk_i k_j}^{(ij)})^2) \quad (11)$$

综上所述,根据式(3)、式(4)两种残差计算方式,设计 MSR-HFCC 算法的两种计算形式 MSR-HFCC_H1 和 MSR-HFCC_H2。其中,MSR-HFCC_H1 算法的计算步骤描述如下:

Input:relational matrix $\{\boldsymbol{D}^{(ij)}\}_{i,j=\{1,2,\cdots,m\}}$, number of clusters $\{K_i\}_{i=\{1,2,\cdots,m\}}$, parameter $\{T_u^{(i)}\}_{i=\{1,2,\cdots,m\}}$ and $\tau_{\max}$, positive weights $\{\beta_{ij}\}_{i,j=\{1,2,\cdots,m\}}$。

Output: Memberships $u_{k_i p}^{(i)}$ for each type data X^i。

Set parameter $T_u^{(i)}$ and $\tau_{\max}$; Set $\tau=0$;

Randomly initialize$(u_{k_i p}^{(i)})^{\tau}$;

Compute $(a_{k_i k_j}^{(ij)})^{\tau}$ using Eq.(5);

REPEAT

　for $1\leqslant i\leqslant m$

　　update $(u_{k_i p}^{(i)})^{\tau+1}$ using Eq.(11);

　　update $(a_{k_i k_j}^{(ij)})^{\tau+1}$ using Eq.(5);

　　compute $(h_{pqk_i k_j}^{(ij)})^{\tau+1}$ using Eq.(3);

　end for

$\tau=\tau+1$;

UNTIL $\max_{i,k_i,p}|(u_{k_i p}^{(i)})^{\tau+1}-(u_{k_i p}^{(i)})^{\tau}|\leqslant\varepsilon$ or $\tau=\tau_{\max}$

MSR-HFCC_H2 与 MSR-HFCC_H1 的计算步骤类似,不同之处在于 MSR-HFCC_H2 利用式(5)、式(6)、式(7)和式(4)分别更新$(a_{k_i k_j}^{(ij)})^{\tau+1}$、$(a_{pk_j}^{(ij)})^{\tau+1}$、$(a_{k_i q}^{(ij)})^{\tau+1}$和$(h_{pqk_i k_j}^{(ij)})^{\tau+1}$。

根据式(11)计算 $u_{k_i p}^{(i)}$ 的时间复杂度为 $O(n_i K_i\sum_{j=1,j\neq i}^{m}K_j n_j)$;根据式(5)、式(6)、式(7)计算 $a_{k_i k_j}^{(ij)}$ 的时间复杂度分别为 $O(n_i n_j)$、$O(n_j)$ 和 $O(n_i)$。高阶异构数据涉及 m 种类型的数据,MSR-HFCC_H1 和 MSR-HFCC_H2 的时间复杂度均为 $O(\tau\sum_{i=1}^{m}(n_i K_i\sum_{j=1,j\neq i}^{m}K_j n_j))$,其中 τ 为迭代次数。

2　实验分析

2.1　数据集

采用文献[13]的方法,利用 Cora 论文数据集,由会议、论文和单词组成高阶异构数据集 $T1$ 和 $T2$,每个类别包含 300 篇论文;采用文献[12]的方法,利用 Corel 图像数据集,由图像、单词和图像分割组构建高阶异构数据集 $I1$ 和 $I2$,每个类别包含 100 张图像;采用文献[5]的方法,利用 IAPR TC-12 Benchmark 数据集,由图像、特征和单词构建高阶异构数据集 $P1$,每个类别包含 300 张图像。论文和图像类别、单词数见表 1。

2.2　结果分析

2.2.1　重叠聚簇结构分析

图 1 为 MSR-HFCC 算法在 $I1$ 数据集上挖掘的聚簇结构可视化结果,可以看出存在明显的聚簇重叠结构,$T1$、$T2$、$I2$ 和 $P1$ 数据集中也具有

表1 标准数据集

Tab. 1 Benchmark Datasets

简称	数据集	单词数	类别
*T*1	Cora	2 000	database, artificial intelligence
*T*2	Cora	2 000	operating systems , architecture
*I*1	Corel	116	cow, grass, horses
*I*2	Corel	114	tree, bird, sky
*P*1	IAPR	1 000	traveler, animal

类似的现象。图2为*P*1数据集中图像对于第一个聚簇的隶属度值,260～376之间的图像为聚簇重叠部分,隶属度值接近于0.5。表2为*P*1数据集中聚簇重叠部分图像的隶属度示例。这些图像既属于traveler聚簇,又属于animal聚簇,因此难以严格地将数据划分到traveler聚簇或者animal聚簇中。可见,MSR-HFCC_H算法更客观地描述了现实世界数据的聚类结果,能够有效地分析具有聚簇重叠结构的数据。

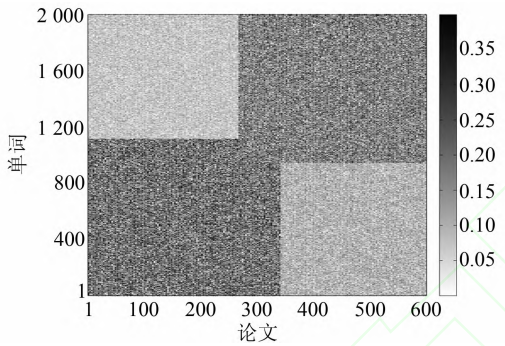


图1 聚簇结构可视化结果

Fig. 1 Results of Visual Cluster Structures

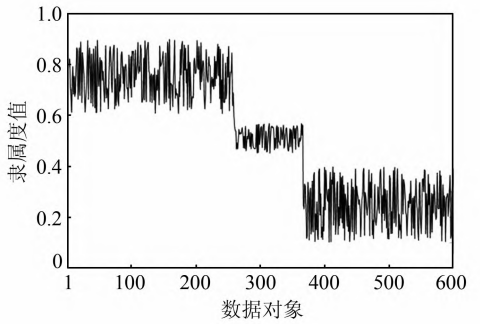


图2 *P*1数据集图像隶属度值分布

Fig. 2 Distribution of Memberships of Images in *P*1 Dataset

2.2.2 准确性分析

将MSR-HFCC_H1和MSR-HFCC_H2算法与目前已有的5种"硬划分"高阶联合聚类算法进行准确率对比,包括CBGC[3]、CIT[4]、RSN[8]、NMF[11]和CoStar[12]。图3所示为以上算法聚类结果的正则化互信息(normalized mutual information, NMI)值。首先,MSR-HFCC_H1和MSR-HFCC_H2算法的聚类结果明显优于目前已有的5种"硬划分"高阶联合聚类算法,这一定程度上是由于模糊方法能更有效地描述具有重叠聚簇结构数据的聚类结果。其次,MSR-HFCC_H2算法的聚类结果略优于MSR-HFCC_H1的。其原因在于MSR-HFCC_H2算法在计算残差时不仅考虑了某对数据之间的关系与数据所在聚簇内其他关系的差异,而且考虑了与此对数据相关的其他关系的差异。

表2 *P*1数据集中重叠聚簇部分图像隶属度值示例

Tab. 2 Examples of Membership of Images at the Overlaps of Clusters

												
簇1	0.485	0.493	0.513	0.481	0.509	0.524	0.497	0.518	0.507	0.513	0.519	0.507
簇2	0.515	0.507	0.487	0.519	0.491	0.476	0.503	0.482	0.493	0.487	0.481	0.493

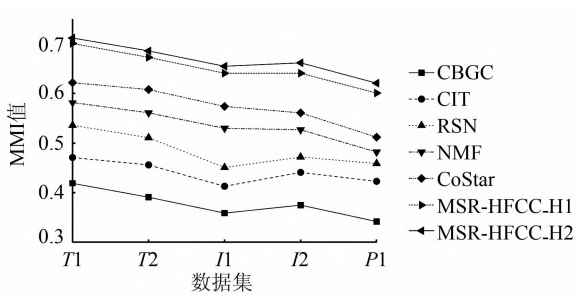


图3 不同算法聚类结果的NMI值

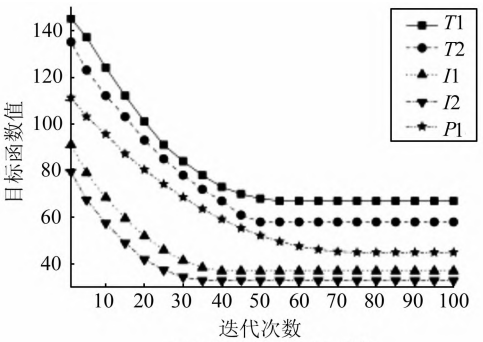
Fig. 3 NMI Values of Clustering Results

2.2.3 收敛性分析

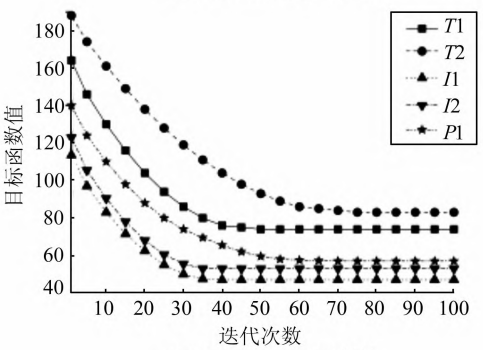
从图4可以看出,MSR-HFCC_H1和MSR-HFCC_H2算法在5个数据集上迭代40次左右达到收敛状态,具有较好的收敛性。

3 结 语

本文提出了基于最小平方和残差的高阶模糊联合聚类算法MSR-HFCC_H1和MSR-HFCC_H2,该算法能够更有效地描述和分析具有聚簇重叠结构的高阶异构数据,获得更加符合实际的聚类结果。



(a) MSR-HFCC_H1算法



(b) MSR-HFCC_H2算法

图 4　MSR-HFCC 算法的收敛性

Fig. 4　Convergence of MSR-HFCC Algorithm

参　考　文　献

[1] Wang Hua, Nie Feiping, Huang Heng, et al. Nonnegative Matrix Tri-Factorization Based High-Order Co-clustering and Its Fast Implementation[C]. The 11th IEEE International Conference on Data Mining, Arlington, USA, 2011

[2] Wang Hua, Nie Feiping, Ding C. Simultaneous Clustering of Multi-type Relational Data via Symmetric Nonnegative Matrix Tri-factorization[C]. The 20th ACM International Conference on Information and Knowledge Management, Glasgow, UK, 2011

[3] Gao Bin, Liu Tieyan, Zheng Xin, et al. Consistent Bipartite Graph Co-Partitioning for Star-Structured High-Order Heterogeneous Data Co-Clustering[C]. The 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, Chicago, USA, 2005

[4] Liu Tieyan, Ma Weiying. Star-Structured High-Order Heterogenous Data Co-Clustering Based on Consistent Information Theory[C]. The 6th IEEE International Conference on Data Mining, Hong Kong, China, 2006

[5] Gao Bin, Liu Tieyan, Qin Tao, et al. Web Image Clustering by Consistent Utilization of Visual Features and Surrounding Texts[C]. The 13th Annual ACM International Conference on Multimedia, Singapore, 2005

[6] Rege M, Dong Ming, Hua Jing. Graph Theoretical Framework for Simultaneously Integrating Visual and Textual Features for Efficient Web Image Clustering[C]. The 17th International Conference on World Wide Web, Beijing, China, 2008

[7] Greco G, Guzzo A. Co-clustering Multiple Heterogeneous Domains: Linear Combinations and Agreements[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22(12): 1 649-1 663

[8] Long B, Wu Xiaoyun, Zhang Zhongfei, et al. Unsupervised Learning on *K*-Partite Graphs[C]. The 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, USA, 2006

[9] Shao Jian, Yin Wentao, Ma Shuai, et al. Topic Discovery of Web Video Using Star-Structured *K*-Partite Graph[C]. The International Conference on Multimedia, Firenze, Italy, 2010

[10] Long B, Zhang Zhongfei, Wu Xiaoyun, et al. Spectral Clustering for Multi-type Relational Data[C]. The 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006

[11] Chen Yanhua, Wang Lijun, Dong Ming. Non-Negative Matrix Factorization for Semisupervised Heterogeneous Data Co-clustering[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2010, 22(10): 1 459-1 474

[12] Dino I, Robardet C, Pensa R G, et al. Parameterless Co-clustering for Star-Structured Heterogeneous Data[J]. *Data Mining Knowledge Discovery*, 2012, 26: 217-254

[13] Sun Yizhou, Yu Yintao, Han Jiawei. Ranking-Based Clustering of Heterogeneous Information Networks with Star Network Schema[C]. The 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 2009

[14] Hartigan J A. Direct Clustering of a Data Matrix [J]. *Journal American Statistical Association*, 1972, 67 (337): 123-129

[15] Cheng Yizong, Church G M. Biclustering of Expression Data[C]. The 8th International Conference of Intelligent Systems for Molecular Biology, San Diego, USA, 2000

[16] Zhong Yexun, Hu Baoqing, Qiao Junjun. Fuzzy Clustering of Multi-factors Evaluated System[J]. *Geomatics and Information Science of Wuhan University*, 2010, 35(6): 752-755(钟业勋,胡宝清,乔俊军. 多因素评价体系的模糊聚类分析[J]. 武汉大学学报・信息科学学报, 2010, 35(6): 752-755)

[17] Kong Lingqiao, Qin Kun, Long Tengfei. Global SST Data Mining Based on Fuzzy Clustering[J].

Geomatics and Information Science of Wuhan University, 2012,37(2):215-219(孔令桥,秦昆,龙腾飞.利用二阶模糊聚类进行全球海表温度数据挖掘[J].武汉大学学报·信息科学学报,2012,37(2):215-219)

A Minimum Sum-squared Residue for High-order Fuzzy Co-clustering Algorithm

HUANG Shaobin[1] *YANG Xinxin*[1]

1 College of Computer Science & Technology, Harbin Engineering University, Harbin 150001, China

Abstract: Most existing high-order co-clustering algorithms focus on hard clustering methods, which ignore the problem of overlaps in the clustering structures. In order to analyze the clustering results of data with overlapping clusters more efficiently, we developed a minimum sum-squared residue for high-order fuzzy co-clustering algorithm (MSR-HFCC). The clustering problem is formulated as the problem of minimizing fuzzy sum-squared residue. The update rules for fuzzy memberships were derived, and an iterative algorithm was designed for a co-clustering process. Finally, experimental results show that the qualities of clustering results of MSR-HFCC are superior to five existing algorithms.

Key words: high-order heterogeneous data; co-clustering; fuzzy clustering; sum-squared residue

First author: HUANG Shaobin, professor, PhD. He is engaged in data mining, complex and social network. E-mail: huangshaobin_hrbeu@126.com

Foundation support: The National Natural Science Foundation of China, Nos. 71272216, 60903080; the Science Foundation for Post Doctorate Research, No. 2012M5100480; the National Key Technology R&D Program, Nos. 2009BAH42B02, 2012BAH08B02; the Fundamental Research Funds for the Central University, Nos. HEUCFZ1212, HEUCF100603.

(上接第237页)

noise energy contained in each IMF is approximately estimated by using the IMF noise energy distribution model, and then, decomposing the each IMF by KPCA, and adaptively selecting the principle components which are should be retained. At last, the denoised gyro signal is obtained by accumulating the each processed IMF by KPCA. A detailed comparison between the proposed method and the wavelet methods is given. The denoising effect of different methods is analyzed by the overlapping Allan variance. Experimental results show that the proposed method performs better in removing noise than classic wavelet methods and can more efficiently suppress the gyro random drift.

Key words: gyro random drift; empirical mode decomposition; kernel principal component analysis; de-noise

First author: YU Min, PhD candidate, specializes in the application of fractal and wavelet. E-mail: yufeng3378@yahoo.com.cn

Foundation support: The Fund of State Key Laboratory of Satellite Ocean Environment Dynamics, No. SOED1405.