大数据中基于熵加权的稀疏分数特征选择聚类算法*

魏霖静[1†], 宁璐璐[2], 郭 斌[3], 侯振兴[4]
(1. 甘肃农业大学 信息科学技术学院, 兰州 730070; 2. 南洋理工大学 生物科学学院, 新加坡 639798; 3. 河海大学 计算机与信息学院, 南京 210094; 4. 南京大学 信息管理学院, 南京 210093)

摘 要: 为了提高大数据统计及分析的效率,有必要对数据集合进行聚类,以减少数据集合维度,并去掉相似数据冗余。采用熵加权和稀疏分数特征选择相结合,一方面对异构数据进行局部结构划分,降低数据维度,对局部结构的特征重要性标记并排序,提高聚类精度,另一方面,提高聚类稳定性。实验证明,该方法对不同种类的大数据聚类具有较强的适用性。

关键词: 数据聚类; 熵加权; 稀疏分数; 特征选择; 数据维度; 大数据
中图分类号: TP274 **文献标志码**: A **文章编号**: 1001-3695(2018)08-2293-02
doi:10.3969/j.issn.1001-3695.2018.08.013

Clustering algorithm based on entropy weighted and sparse fractional feature selection in big data

Wei Linjing[1†], Ning Lulu[2], Guo Bin[3], Hou Zhenxing[4]
(1. *School of Information Science & Technology, Gansu Agriculture University, Lanzhou* 730070, *China*; 2. *School of Biological Sciences, Nanyang Technological University, Singapore City* 639798, *Singapore*; 3. *School of Computer Science & Engineering, Hohai University, Nanjing* 210094, *China*; 4. *School of Information Science & Engineering, Nanjing University, Nanjing* 210093, *China*)

Abstract: In order to improve the efficiency of data statistics and analysis, it is necessary to cluster data sets, for reduces the data sets collection dimension and removes similar data redundancy. This paper used entropy weighted and sparse fractional feature selection. On the one hand, it divided the local structure of heterogeneous data, reduced the data dimension, marked and sorted the feature importance of local structure, and improved the clustering accuracy. Experimental results show that the method has strong applicability to different kinds of large data clustering.

Key words: data clustering; entropy weighted; sparse fraction; feature selection; data dimension; big data

随着数据密集型时代的到来,大数据已经成为重要的资产。而数据数量巨大,价值密度低,实时在线,多源异构,造成了数据挖掘的难度。面对浩如烟海的数据,大数据的挖掘常用的方法有分类、回归分析、聚类、关联规则、神经网络方法、Web数据挖掘等。大数据分析和挖掘的首要任务是聚类,聚类是一个划分数据对象集的过程。文献[1]是常见聚类算法 K-means 的改进在大数据并行处理中的应用,文献[2]将分布式处理技术与聚类算法结合,实现了分布式多视图聚类,文献[3]是大规模数据聚类的算法改进[3-6]。这些研究都致力于提高聚类精度,但在算法时间复杂度和收敛速度方面涉猎较少,本文将熵加权和稀疏分数应用于数据集合的特征选择和聚类,以求提高数据聚类的精度和稳定性。

1 竞争合并熵加权

在介绍竞争合并熵加权之前,本文有必要先介绍在线熵加权算法。熵作为信息源保护信息量多少的测度,可以用来度量系统的不确定性程度[7]。首先定义聚类的目标函数为

$$J(t)=\sum_{j=1}^{N}\sum_{i=1}^{C}u_{ij}^{m}\sum_{k=1}^{D}w_{ik}(x_{jk}-v_{ik})^2+\gamma\sum_{i=1}^{C}\sum_{k=1}^{D}w_{ik}\log w_{ik} \tag{1}$$

其中,有 $0\leqslant u_{ij}\leqslant 1$, $\sum_{i=1}^{C}u_{ij}=1$, $0\leqslant w_{ik}\leqslant 1$, $\sum_{k=1}^{D}w_{ik}=1$。

设定被聚类的对象 $X=\{x_1,x_2,\cdots,x_N\}\subset R^D$,聚类个数定义为 C,算法遍历次数为 M[8]。

然后进行初始化 $w_{ik}(0)$,然后进行重复迭代。迭代过程主要是计算隶属表述程度、特征系数等[9],计算公式如下:

由式(1)及最小聚类算法目标函数,计算当前集合的隶属表述程度:

$$u_{ij}=\frac{(d_{ij})^{-1/m-1}}{\sum_{s=1}^{C}(d_{sj})^{-1/m-1}} \tag{2}$$

然后该数据集合的特征系数为

$$v_{ik}=\frac{\sum_{j=1}^{N}u_{ij}^{m}x_{ik}}{\sum_{j=1}^{N}u_{ij}^{m}} \tag{3}$$

根据目标函数及式(2)可以计算隶属迭代公式为

$$u_{i(Nt)}=\frac{(d_{i(Nt)})^{-1/m-1}}{\sum_{s=1}^{C}(d_{s(Nt)})^{-1/m-1}} \tag{4}$$

然后根据式(3)计算的结果计算聚类中心距离:

$$d_{i(Nt)}=\sum_{k=1}^{D}w_{ik}(t-1)(x_{(Nt)k}-v_{ik})^2 \tag{5}$$

进一步,计算 t 时刻聚类中心值,方法如式(6)所示。

$$v_{ik}(t)=v_{ik}(t-1)-\eta^{(t)}\times u_{i(Nt)}^{m}\times(v_{ik}(t-1)-x_{(Nt)k}) \tag{6}$$

其中,有 $\eta^{(t)}=\eta_0(\eta_f/\eta_0)^{\frac{t}{NM}}$。

接着,计算熵加权系数,计算方法如式(7)所示。

$$w_{ik}(t)=\frac{\exp(-q_{ik}(t)/\gamma)}{\sum_{s=1}^{D}\exp(-q_{is}(t)/\gamma)} \tag{7}$$

其中,有 $q_{ik}(t)=q_{ik}(t-1)-u_{i(Nt)}^{m}(v_{ik}(t)-x_{(Nt)k})^2$。

考虑到在线熵加权迭代次数及算法时间复杂度问题,引入竞争合并策略熵加权。在处理高维度数据集合过程中,能减少时间复杂度[10]。竞争合并熵加权与在线熵加权在算法中最大的区别是其引入了竞争合并参数与特征系数加权矩阵。主要

收稿日期: 2017-04-10; **修回日期**: 2017-06-02 **基金项目**: 国家自然科学基金资助项目(61063028,31560378);江苏省自然科学基金青年基金资助项目(BK20150784);中国博士后面上资助项目(2015M581800);甘肃省科技支撑计划项目(1604WKCA011);陇原青年创新创业人才项目(2016-47);2016 年度甘肃省高校重大软科学(战略)研究项目计划资助项目(2016F-10)

作者简介: 魏霖静(1977-),女(通信作者),甘肃嘉峪关人,副教授,博士(后),主要研究方向为智能计算、算法应用研究、图像分析(wlj@gsau.edu.cn);宁璐璐(1989-),女,河南新乡人,博士(后),主要研究方向为计算生物学、生物信息学、图像数据分析;郭斌(1981-),男,辽宁沈阳人,博士研究生,主要研究方向为模式识别、嵌入式系统等;侯振兴(1977-),男,甘肃武威人,副教授,博士研究生,主要研究方向为人机交互与用户行为、算法应用研究.

算法如下:设定被聚类的对象 $X=\{x_1,x_2,\cdots,x_N\}\subset R^D$,聚类个数定义为 C,竞争合并参数 η,然后进行初始化设置,包括隶属表述程度 $u_{ij}^{(1)}$[11]。

然后计算聚类中心矩阵表示为

$$v_{ik}=\frac{\sum_{j=1}^{n}u_{ij}^2x_{jk}}{\sum_{j=1}^{n}u_{ij}^2} \tag{8}$$

给定数据集合 $U=[u_1,u_2,\cdots,u_n]$ 和 $U=[v_1,v_2,\cdots,v_n]$,并设定 w_{ik} 计算如式(9)。

$$w_{ik}=\frac{\exp(-\sum_{j=1}^{n}u_{ij}^2(x_{jk}-v_{ik})^2/\gamma)}{\sum_{s=1}^{d}\exp(-\sum_{j=1}^{n}u_{ij}^2(x_{js}-v_{is})^2/\gamma)} \tag{9}$$

将式(9)与目标聚类函数式(1)两者结合,得到如下方程。

$$\psi(w_{ik})=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^2\sum_{k=1}^{d}w_{ik}(x_{jk}-v_{ik})^2+\gamma\sum_{i=1}^{c}\sum_{j=1}^{n}w_{ik}\log w_{ik}-\sum_{i=1}^{c}\lambda_i^w(\sum_{k=1}^{d}w_{ik}-1) \tag{10}$$

式(10)分别对 w_{ik} 和 λ_i^w 求偏导数并令结果等于0,则有

$$\frac{\partial\psi(w_{ik})}{\partial w_{ik}}=\sum_{j=1}^{n}u_{ij}^2(x_{jk}-v_{ik})^2+\gamma(\log w_{ik}+1)-\lambda_i^w=0 \tag{11}$$

$$\frac{\partial\psi(w_{ik})}{\partial w_{ik}}=\sum_{k=1}^{d}w_{ik}-1=0 \tag{12}$$

结合式(11)(12)可得

$$w_{ik}=\frac{\exp(-\sum_{j=1}^{n}u_{ij}^2(x_{jk}-v_{ik})^2/\gamma)}{\sum_{s=1}^{d}\exp(-\sum_{j=1}^{n}u_{ij}^2(x_{js}-v_{is})^2/\gamma)} \tag{13}$$

所以式(13)是式(1)的必要条件。然后定义竞争合并熵加权的隶属表示计算方法如下:

$$u_{ij}=u'+u''_{ij} \tag{14}$$

其中:$u'_{ij}=\frac{1}{\sum_{s=1}^{C}[(\sum_{k=1}^{d}w_{ik}(x_{jk}-v_{ij})^2)/(\sum_{k=1}^{d}w_{sk}(x_{jk}-v_{sk})^2)]}$ (15)

$$u''_{ij}=\alpha\frac{N_i-N_j}{\sum_{k=1}^{d}w_{ik}(x_{jk}-v_{ik})^2} \tag{16}$$

在式(16)中有

$$N_j=\alpha\frac{\sum_{s=1}^{C}[1/(\sum_{k=1}^{d}w_{sk}(x_{jk}-v_{sk})^2)N_s]}{\sum_{s=1}^{C}[1/\sum_{k=1}^{d}w_{sk}(x_{jk}-v_{sk})^2]} \tag{17}$$

2 稀疏分数特征选择聚类

稀疏表示方法最早用于模式识别领域,特别适用于人脸识别[12],在近年来的研究中,逐渐将稀疏分数表示用于其他的特征选择中,对于特征样本集的处理有独特之处,因此,该方法适合对大数据样本的特征判断和分类,在选择判断过程中,能较好地过滤不需要的特征,保留需要的特征。具体的特征选择步骤如下:

首先,对数据集合进行 l_1 范数最小化,给定集合 $\{x_i\}_{i=1}^{n}$,$x_i\in\mathrm{R}^d$,令 $\boldsymbol{X}=[x_1,x_2,\cdots,x_n]\in\mathrm{R}^{d\times n}$,然后对 x_i 求解稀疏分数系数[13],计算方法如式(18)。

$$\min_{S_i}\|S_i\|_1\quad\text{s.t. }x_i=\boldsymbol{X}'s_i \tag{18}$$

其中:$\boldsymbol{X}'$ 为除了第 i 列的 $\boldsymbol{X}$ 矩阵。得到系数之后,接着计算系数分数目标函数,如式(19)。

$$S(r)=\frac{\sum_{i=1}^{n}(x_{ir}-(\boldsymbol{X}s_i)_r)^2}{\mathrm{var}(\boldsymbol{X}(r))} \tag{19}$$

其中:$\sum_{i=1}^{n}(x_{ir}-(\boldsymbol{X}s_i)_r)^2$ 表示第 r 维特征与该维重构特征的误差;$\mathrm{var}(\boldsymbol{X}(r))$ 表示累积方差[13]。

最后,根据 $S(r)$ 进行特征标签并排序,特征重要性与 $S(r)$ 的值成反比,为了选取重要特征,这里应选取 $S(r)$ 值最小的特征。

3 实验仿真

本文实例仿真主要是对基于竞争合并的熵加权聚类算法和稀疏分数特征选择的聚类性能进行验证分析。在性能分析中,主要是对算法的聚类效果及稳定性仿真。为了方便显示数据集合的聚类结果,采用二维可视化方法显示,而在稳定性分析中,以算法的收敛速度作为评价标准。

3.1 基于熵加权的同规模聚类结果仿真

实现仿真硬件环境主要配置为,2.6 GHz CPU 主频,4 GB 内存,500 GB 硬盘。为了节省算法初始化时间成本,对数据集合先进行常见归一化处理,并设置初始熵加权系数 $\gamma=1$。为了排除数据集合的主观性,本文数据集合由系统随机发生器产生。

首先,将聚类数据集合个数设置为 15 个,即初始目标聚类中心为 15 个,初始的二维视图如图 1 所示。

经过熵加权聚类处理,实验要求目标聚类中心个数降至 3 个。为了验证算法收敛速度,将算法的迭代次数分别设置为 2、3、9、16 次。每次设置完毕初值,仿真结果如图 2 ~ 5 所示。

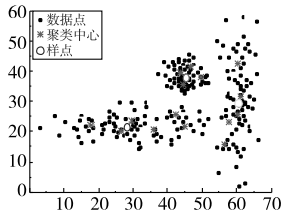

图1 初始聚类中心二维视图分布

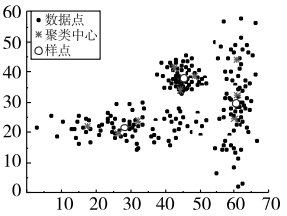

图2 2次迭代后聚类结果

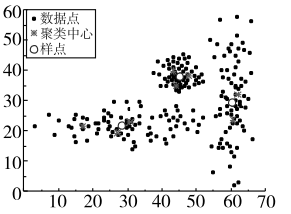

图3 3次迭代后聚类结果

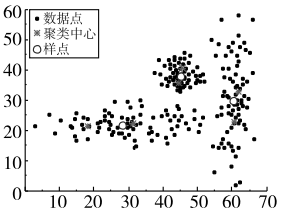

图4 9次迭代后聚类结果

从图 2 ~ 5 可以看出,当迭代次数设置为 2、3、9、16 次时,聚类中心分别变为 9 个、8 个、6 个和 4 个。迭代次数越多,越接近目标聚类中心个数 3 个。迭代次数与聚类中心个数呈非线性关系。为了达到目标聚类中心个数 3,继续迭代,实验发现,当迭代次数为 17 次时,聚类中心个数变为 3 个,如图 6 所示。

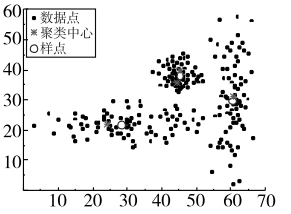

图5 16次迭代后聚类结果

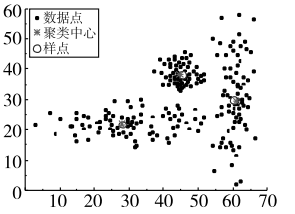

图6 17次迭代后聚类结果

当迭代次数达到26 次时,发现聚类中心个数收敛稳定,聚类中心不再随着迭代次数的增加而变化,验证了该算法的稳定性。

3.2 基于熵加权的不同规模聚类结果仿真

下面将考虑不同聚类集合规模下,该算法的表现性能。给定初始数据集合分别为 9、12、15、18、21。分别验证在该聚类个数时,经过迭代计算,本文算法是否能够达到目标聚类中心个数。仿真结果如图 7 所示。由图 7 可以看出,五种不同规模的数据集合均能达到目标聚类中心个数 3,只是迭代次数和收敛速度有差别。初始聚类中心分别为 9、12、15、18、21 的数据集合,经过计算得到目标聚类中心所经过的迭代次数分别为 3、18、14、20、24。虽然从整体上来说,初始聚类中心个数越多,算法收敛需要的迭代次数越多,但是从以上仿真结果不难发现,初始聚类中心为 12 的数据集合达到收敛时迭代 18 次,而初始聚类中心为 15 的数据集合达到收敛时只迭代 14 次,迭代次数与初始聚类中心个数并不成正比。

3.3 基于熵加权与稀疏分数结合的聚类结果仿真

最后对熵加权与稀疏分数结合的特征选 (下转第 2303 页)

为文本中句子的分类问题。通过对句子的语义角色进行识别,将每一个包含数值信息的分句与五种核心语义角色中的一个角色相对应,最终实现问题的求解。为了实现上述目标,本文提出了一种基于特征词与 *n*-gram 模型相结合的方法对应用题文本中的句子进行角色识别。为了验证方法的有效性,本文采集了 189 道分层抽样应用题,其中 150 道作为训练集,39 道作为测试集。实验结果表明,与仅通过特征词和模式进行判断的方法相比,基于特征词与 *n*-gram 模型相结合的方法有效地提高了句子角色判定的准确率,尤其是整题的识别率从 17.95% 提高到了 64.1%,这也证明了本文所提出方法的有效性。但是该结果依然存在很大的可提升空间。如何增强模型的适应性、拓展模型的覆盖率、进一步提升角色判定的准确率是下一步的工作重点。

参考文献:

[1] Feigenbaum E A, Feldman J. Computers and thought[M]. Cambridge, MA: MIT Press, 1963.

[2] Schoenfeld A H. Mathematical problem solving[M]. [S. l.]: Elsevier, 2014.

[3] 吴文俊. 初等几何判定问题与机械化证明[J]. 中国科学:数学, 1977, 20(6): 507-516.

[4] 张景中, 杨路, 高小山, 等. 几何定理可读证明的自动生成[J]. 计算机学报, 1995, 18(5): 380-393.

[5] Schoenfeld A H. Reflections on problem solving theory and practice [J]. The Mathematics Enthusiast, 2013, 10(1): 9-34.

[6] Hosseini M J, Hajishirzi H, Etzioni O, *et al*. Learning to solve arithmetic word problems with verb categorization[C]//Proc of Conference on Empirical Methods in Natural Language Processing. 2014: 523-533.

[7] Kushman N, Artzi Y, Zettlemoyer L, *et al*. Learning to automatically solve algebra word problems[C]//Proc of the 52nd Annual Meeting of the Association for Computational Linguistics. 2014.

[8] Bobrow D G. Natural language input for a computer problem solving system[J]. Semantic Information Processing, 1964, 9(3): 281-288.

[9] Charniak E. Computer solution of calculus word problems[C]//Proc of the 1st International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Publishers Inc, 1969: 303-316.

[10] Riley M S. Development of children's problem-solving ability in arithmetic[R]. [S. l.]: National Institute of Education, 1984.

[11] Kintsch W, Greeno J G. Understanding and solving word arithmetic problems[J]. Psychological Review, 1985, 92(1): 109-129.

[12] Dellarosa D. A computer simulation of children's arithmetic word-problem solving[J]. Behavior Research Methods, Instruments & Computers, 1986, 18(2): 147-154.

[13] Wong W K, Hsu S C, Wu S H, *et al*. LIM-G: learner-initiating instruction model based on cognitive knowledge for geometry word problem comprehension[J]. Computers & Education, 2007, 48(4): 582-601.

[14] 程志. 小学数学应用题自动解答系统的研究——以整数一、二步和分数基本应用题为例[D]. 北京: 北京师范大学, 2008.

[15] 马玉慧. 小学算术应用题自动解答的框架表征及演算方法研究[D]. 北京: 北京师范大学, 2010.

[16] Yu Xinguo, Wang Mingshu, Zeng Zhizhong, *et al*. Solving directly-stated arithmetic word problems in Chinese[C]//Proc of International Conference of Educational Innovation through Technology. Washington DC: IEEE Computer Society, 2015: 51-55.

[17] Wang Yingxu, Chiew V. On the cognitive process of human problem solving[J]. Cognitive Systems Research, 2010, 11(1): 81-92.

[18] Krawec J, Huang Jia, Montague M, *et al*. The effects of cognitive strategy instruction on knowledge of math problem-solving processes of middle school students with learning disabilities[J]. Learning Disability Quarterly, 2013, 36(2): 80-92.

[19] Singer F M, Voica C. A problem-solving conceptual framework and its implications in designing problem-posing tasks[J]. Educational Studies in Mathematics, 2013, 83(1): 9-26.

[20] Hooshyar D, Ahmad R B, Yousefi M, *et al*. SITS: a solution-based intelligent tutoring system for students' acquisition of problem-solving skills in computer programming[J]. Innovations in Education and Teaching International, 2016, 64(4): 787-808.

[21] Qin Yulin, Carter C S, Silk E M, *et al*. The change of the brain activation patterns as children learn algebra equation solving[J]. Proceedings of the National Academy of Sciences of the United States of America, 2004, 101(15): 5686-5691.

[22] Kao Y S, Douglass S A, Fincham J M, *et al*. Traveling the second bridge: using fMRI to assess an ACT-R model of geometry proof[J]. Ajr American Journal of Roentgenology, 2012, 135(1): 164-166.

(上接第 2294 页)择性能进行仿真,实验随机选取六组数据进行仿真测试,六组数据的最大特征数为 20 个,将稀疏分数特征选择算法运用于这六组数据,以便对六组数据的特征重要性进行标签并排序。实验仿真结果如图 8 所示。

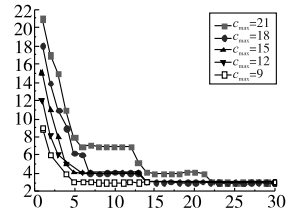

图7　不同数据集合规模的聚类结果

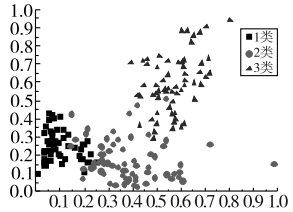

图8　稀疏分数特征选择聚类结果

从图 8 可以看出,六组数据大致被聚类成三类,聚类的依据是根据六组数据最重要的两个特征,即特征 7 和 13,将六组数据分为了三类。稀疏分数特征选择在聚类中最重要的作用,是在 20 个特征中找出了最重要的特征 7 和 13 两个特征。这是稀疏分数特征选择算法降低高维度数据独特的优势。

4　结束语

本文采用熵加权的方法对大数据集的局部结构进行划分加权运算,提高了聚类算法的稳定性,在局部结构的特征选择过程中,采用稀疏分数表示法,将高维度数据的相似局部结构进行去冗余,降低数据维度,以求得到更有效的聚类结果。

参考文献:

[1] 李晓瑜, 俞丽颖, 雷航, 等. 一种 K-means 改进算法的并行化实现与应用[J]. 电子科技大学学报, 2017, 46(1): 61-68.

[2] 邓强, 杨燕, 王浩. 一种改进的多视图聚类集成算法[J]. 计算机科学, 2017, 44(1): 65-70.

[3] Serdah A M, Ashour W M. Clustering large-scale data based on modified affinity propagation algorithm[J]. Journal of Artificial Intelligence & Soft Computing Research, 2016, 6(1): 23-33.

[4] Li Yangyang, Yang Guoli, He Haiyang, *et al*. A study of large-scale data clustering based on fuzzy clustering[J]. Soft Computing, 2016, 20(8): 3231-3242.

[5] Si Fuming, Bu Tianran. Design of a large data clustering algorithm based on Hadoop cloud computing platform[J]. Journal of Chuxiong Normal University, 2016, 31(3): 49-55.

[6] Zhang Yanfeng, Chen Shimin, Yu Ge. Efficient distributed density peaks for clustering large data sets in MapReduce[J]. IEEE Trans on Knowledge and Data Engineering, 2016, 28(12): 3218-3230.

[7] Delgado A, Romero I. Environmental conflict analysis using an integrated grey clustering and entropy-weight method[J]. Environmental Modelling & Software, 2016, 77(C): 108-121.

[8] Zhang Lijun, Zhao Fangfang. Application for technological achievements evaluations model based on entropy weight and matter-element analysis[J]. Science & Technology Management Research, 2016(6): 70-73.

[9] 邱保志, 贺艳芳, 申向东. 熵加权多视角核 K-means 算法[J]. 计算机应用, 2016, 36(6): 1619-1623.

[10] 高翠芳, 黄珊维, 沈莞蔷, 等. 基于信息熵加权的协同聚类改进算法[J]. 计算机应用研究, 2015, 32(4): 1016-1018.

[11] 蒋亦樟, 邓赵红, 王骏, 等. 熵加权多视角协同划分模糊聚类算法[J]. 软件学报, 2014, 25(10): 2293-2311.

[12] 吴杰祺, 李晓宇, 袁晓彤, 等. 利用坐标下降实现并行稀疏子空间聚类[J]. 计算机应用, 2016, 36(2): 372-376.

[13] 岳温川, 王卫卫, 李小平. 基于加权稀疏子空间聚类多特征融合图像分割[J]. 系统工程与电子技术, 2016, 38(9): 2184-2191.