近似骨架导向的归约聚类算法

宗　瑜[①]　　李明楚[①]　　江　贺[①②*]
[①](大连理工大学软件学院　大连　116621)
[②](中国科学院软件研究所计算机科学国家重点实验室　北京　100190)

摘　要：该文针对聚类问题上缺乏骨架研究成果的现状，分析了聚类问题的近似骨架特征，设计并实现了近似骨架导向的归约聚类算法。该算法的基本思想是：首先利用现有的启发式聚类算法得到同一聚类实例的多个局部最优解，通过对局部最优解求交得到近似骨架，将近似骨架固定得到规模更小的搜索空间，最后在新空间上求解。在 26 个仿真数据集和 3 个实际数据集上的实验结果表明，骨架理论对提高聚类质量、降低初始解影响及加快算法收敛速度等方面均十分有效。

关键词：聚类问题；NP-难解；启发式算法；近似骨架
中图分类号：TP181　　**文献标识码**：A　　**文章编号**：1009-5896(2009)12-2953-05

Approximate Backbone Guided Reduction Algorithm for Clustering

Zong Yu[①]　　Li Ming-chu[①]　　Jiang He[①②*]
[①](*School of Software*, *Dalian University of Technology*, *Dalian* 116621, *China*)
[②](*The State Key Laboratory of Computer Science*, *Institute of Software*, *CAS*, *Beijing* 100190, *China*)

Abstract: In this paper, the characteristic of approximate backbone is analyzed and an Approximate Backbone guided Reduction Algorithm for Clustering (ABRAC) is proposed. ABRAC works as follows: firstly, multiple local optimal solutions are obtained by an existing heuristic clustering algorithm; then, the approximate backbone is generated by intersection of local optimal solutions; afterwards, the search space can be dramatically reduced by fixing the approximate backbone; finally, this reduced search space can be efficiently searched to find high quality solutions. Extensively wide experiments on 26 synthetic and 3 real-life data sets demonstrate that the backbone has significantly effects for improving the quality of clustering, reducing the impact of initial solution, and speeding up the convergence rate.

Key words: Clustering issue; NP-hard; Heuristic algorithm; Approximate backbone

1　引言

聚类是模式识别、机器学习及数据挖掘等知识发现任务的重要基础。关于聚类研究的最新发展状况请参阅文献[1]。由于聚类质量评价和应用需求紧密相关，故存在多种聚类标准。对于数值化的数据集，聚类问题可以描述成典型的组合优化问题：给定含 N 个数据对象的数据集 $D \subset R^d$ 和聚类个数 K，寻求 K 个代表点(每个代表点 c_k 对应着一个子簇 C_k)，使得目标函数 $\Phi = \sum_{k=1}^{K} \sum_{x_i \in C_k} || x_i - c_k ||^2$ 取值最小。Drineas 在文献[2]中已经证明即使在 $K = 2$ 的情况下该问题也是 NP-难解的。对于大规模实例，人们的研究目标集中在能用较短时间得到可接受解的启发式算法(heuristic algorithm)方面。现有启发式聚类算法有 k-means[3]，K++-means[4]，Modified_K-means[5]等。

骨架分析(backbone analysis)是近年来启发式算法设计中非常活跃的领域。骨架是指一个 NP-难解问题实例的所有全局最优解的相同部分，对于衡量问题的难度和相变具有重要的意义。由于在很多应用领域骨架获取是 NP-难解的[6,7]，研究者们大多采用近似骨架构造 NP-难解问题的启发式算法：Valnir[8]和 Zhang[9]等分别给出了求解 SAT 问题(SATisfiability problem)的骨架导向局部搜索；江贺等给出了基于偏移实例的近似骨架算法[7]。这些研究表明，近似骨架在提高启发式算法的收敛速度和提高算法质量等方面有着明显的作用。

针对目前缺乏聚类问题的骨架研究成果的情况，本文分析了聚类中近似骨架的性质，并将近似骨架运用于聚类算法的设计。本文工作包括以下方面：(1)通过经典聚类实例的全局最优解和局部最优

2008-12-08 收到，2009-06-29 改回
国家自然科学基金(60805024)和教育部博士点基金(20070141020)资助课题
*通信作者

解的数据对象隶属关系分析，发现对局部最优解进行简单的相交操作，能以较大的概率得到全局最优解中的数据对象，故此可将其作为聚类问题的近似骨架。(2)给出了近似骨架导向的归约聚类算法 ABRAC (Approximate Backbone guided Reduction Algorithm for Clustering)。新算法首先调用现有的启发式聚类算法获得多个局部最优解，然后提取局部最优解中相同部分作为近似骨架，再通过固定近似骨架实现搜索空间的归约，最后在归约后的新搜索空间上求解。本文以经典的 K-means 算法作为从属算法，在 26 个仿真和 3 个实际数据集上与 4 种 K-means[3] (分别以 FA，McQueen，SCS 及 KKZ 作为初始策略)算法进行实验对比。实验结果表明，新算法在收敛速度、紧密度、分离度等方面均优于其余 4 种算法。

2 预备知识

本节首先给出聚类和聚类质量的定义，然后给出聚类问题的骨架定义。

定义 1 给定数据集 D 和子簇个数 K，聚类的可行解就是指一种数据对象的分配方法 $\pi(x_i)\to\{C_1,C_2,\cdots,C_K\}, i=1,2,\cdots,N$，其目标函数值记为 $\Phi_\pi(\{C_1,C_2,\cdots,C_K\})=\sum_{k=1}^{K}\sum_{x_i\in C_k}\|x_i-c_k\|^2$，其中 c_k 是簇 C_k 的代表点($k\in\{1,2,\cdots,K\}$)。聚类的目标是寻求解 $\pi(x_i)\to\{C_1,C_2,\cdots,C_K\}, i=1,2,\cdots,N$，使得目标函数值最小，即 $\Phi_{\pi^*}(\{C_1,C_2,\cdots,C_K\})=\min(\Phi_\pi(\{C_1,C_2,\cdots,C_K\})\,|\,\pi\in\Pi)$，其中 Π 是聚类实例的所有可行解集合。

目标函数 Φ 的取值表达了簇中数据对象与其代表点之间的偏差平方和最小的特征，但不能反映每个子簇的紧密度和簇间分离度特征。因此，He等人给出了的簇内紧密度 Cmp (cluster compactness) 和簇间分离度 Sep (cluster separation) 的聚类质量的评价标准 [10,11]。

定义2 给定数据集合 D，其偏离度定义为 $\mathrm{dev}(D)=\sqrt{(1/N)\sum_{i=1}^{N}\|x_i-\bar{x}\|}$，其中 N 是 D 中包含的元素个数，$\bar{x}=(1/N)\sum_i x_i$ 是 D 中数据对象的均值。

定义3 给定偏离度函数 $\mathrm{dev}(D)$，则 K 个输出簇 $\{C_1,C_2,\cdots,C_K\}$ 的紧密度 Cmp 定义为 $\mathrm{Cmp}=\frac{1}{K}\sum_k\frac{\mathrm{dev}(C_k)}{\mathrm{dev}(D)}$，其中 K 表示簇个数，$\mathrm{dev}(C_k)$ 表示输出簇 C_k 的偏离度，$\mathrm{dev}(D)$ 则表示数据集 D 的偏离度。

定义4 给定与 K 个输出簇 $\{C_1,C_2,\cdots,C_K\}$ 相对应的代表点集合 $\{c_1,c_2,\cdots,c_K\}$，簇间分离度Sep 定义为 $\mathrm{Sep}=\frac{1}{K(K-1)}\sum_{k=1}^{K}\sum_{k'=1,k'\neq k}^{K}\exp\left(\frac{-\|c_k-c_{k'}\|^2}{2\sigma^2}\right)$，其中 σ 为高斯常数。

Cmp 和 Sep 越小说明聚类结果越好。紧密度和分离度是衡量数据集 D 的划分结果是否是簇内紧密、簇间分离的标准。

下面将给出聚类问题的骨架簇、骨架、近似骨架的概念。

定义 5 给定任意聚类问题，存在有限多个全局最优解 $\pi_1^*,\pi_2^*,\cdots,\pi_Q^*$。记所有全局最优解的集合为 Π^*，其中 $Q=|\Pi^*|$ 为全局最优解的个数。聚类问题的一个骨架簇 bC_i 是指满足以下条件的数据对象集合：(1) $|bC_i|\geq 2$；(2)给定任意的两个数据对象 $x_j,x_k\in bC_i$ 和任意的全局最优解 π_l^*，$\pi_l^*(x_j)=\pi_l^*(x_k)$。所有骨架簇的集合称为聚类问题的骨架，记为 $\mathrm{bone}(\pi_1^*,\pi_2^*,\cdots,\pi_Q^*)$。

直观上解释，一个骨架簇的数据对象在所有的全局最优解中均处于同一个簇中。在很多应用领域，全局最优解是很难获得的。故此，很多研究人员依据局部最优解和全局最优解的关系，利用局部最优解的交集模拟全局最优解的交集来获得近似骨架。

定义 6 给定任意聚类问题的多个局部最优解为 $\pi_1,\pi_2,\cdots,\pi_M$，聚类问题的一个近似骨架簇 abC_i 是指满足以下条件的数据对象集合：(1) $|abC_i|\geq 2$；(2)给定任意的两个数据对象 $x_j,x_k\in abC_i$ 和局部最优解 π_l，$\pi_l(x_j)=\pi_l(x_k)$。所有近似骨架簇的集合称为聚类问题的近似骨架，记为 $a\_\mathrm{bone}(\pi_1,\pi_2,\cdots,\pi_M)$。

3 近似骨架性质分析

为了研究近似骨架的特性，本文取 UCI 机器学习知识库(MLDB：http://archive.ics.uci.edu/ml/)中的典型数据集 Haberman's survival，Yeast，Ecoli 及 PRHD (Pen-based Recognition of Handwritten Digits)为实验样本。在这些实验样本中每个数据对象都被标识了分类号，即数据样本中的聚类结果是已知的。本文将这些已知的聚类结果看作为聚类问题的全局最优解。针对每个实验样本，本文分别调用经典的启发式聚类算法 K-means 及 CLARANS 各 i 次 $(i=1,2,\cdots,15)$，然后分别对 K-means 与 CLARANS 算法的 i 个局部最优解执行交集运算从而产生了相应实验样本的近似骨架。

图 1 给出了近似骨架的规模和纯度与局部最优解个数 M 之间的关系。其中，近似骨架规模定义为 $\frac{a\_\mathrm{bone}}{N}\times 100\%$，表示近似骨架中有多少数据对象在 M 个局部最优解中共同出现在同一个聚类中。该

特性反映了近似骨架在算法设计中的有效性。近似骨架纯度表示近似骨架中共同出现的数据对象确实属于骨架的比例，它被用来衡量近似骨架在算法中的可用性。 如图 1 所示，近似骨架的纯度随着相交局部最优解的个数的增加呈上升趋势，而近似骨架规模却呈下降趋势。当相交的局部最优解的个数达到一定程度时，近似骨架的规模和纯度的变化都不是很明显。从图 1 可知，当局部最优解个数大于 10 时，表示近似骨架的规模和纯度的曲线的变化比较平缓，因此在基于近似骨架的归约聚类算法中，本文使用 10 个局部最优解来获得近似骨架。同时，图 1 揭示了不同启发式算法对于近似骨架的规模和纯度有一定影响，但是影响幅度有限。

4　近似骨架导向的归约聚类算法

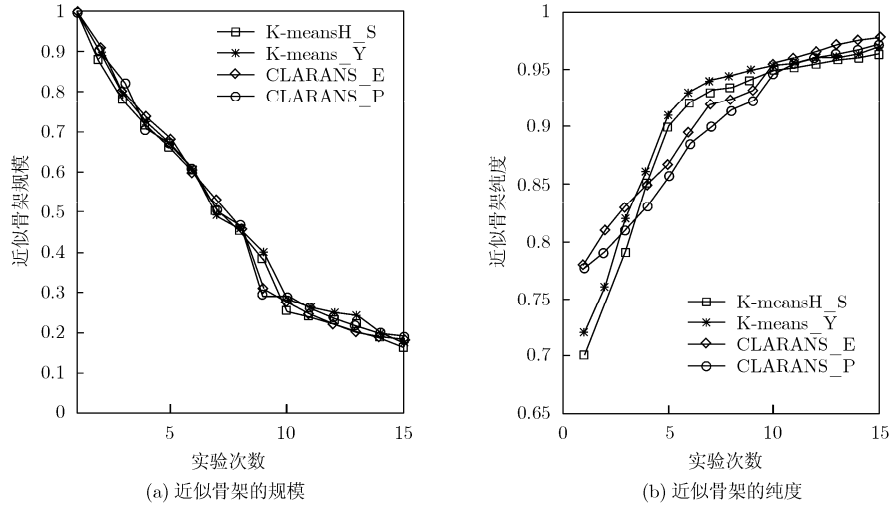
4.1 算法基本思想

根据第 3 节的分析可知，当局部最优解个数达到一定规模时，近似骨架具有较高的纯度。此时，可以固定这些数据对象作为聚类的基本结构，在后续的启发式搜索过程中不再搜索这些数据对象，从而可有效降低启发式搜索算法的搜索空间，提高搜索效率。如图 2 所示，本文采用近似骨架中包含的数据对象的均值来代替近似骨架，从而实现对搜索空间的归约。图 2(a)中黑线连线的数据对象分别是两个近似骨架簇的数据对象，用它们均值代替近似骨架可以缩小数据规模(图 2(a)中有 26 个数据对象，而图 2(b)则只剩下 17 个数据对象，数据规模减少了 34%)。

4.2 ABRAC 算法框架

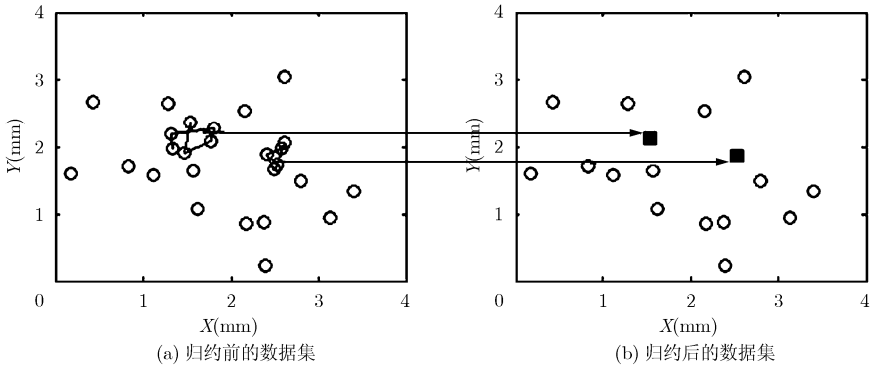
图 3 给出了近似骨架导向的归约聚类算法 ABRAC 的基本框架。步骤(1)调用现有启发式算法 M 次，获得聚类问题的 M 个局部最优解 $\pi_1,\pi_2,\cdots,\pi_M$。步骤(2)对 M 个局部最优解进行求交操作，获得聚类实例的近似骨架。第(3)步则固定前一步骤的近似骨架产生新的搜索空间。最后调用启发式算法 Λ 在新的搜索空间中搜索聚类结果。

算法 ABRAC 具有以下优点:(1)高度的灵活性，启发式算法 ABRAC 并没有限定从属算法 Λ 的类型，而仅仅提供了一种框架。故此，可以将现有的



(a) 近似骨架的规模　　(b) 近似骨架的纯度

图 1 近似骨架的特性



(a) 归约前的数据集　　(b) 归约后的数据集

图 2 近似骨架固定方法

各类聚类问题的启发式算法与之相结合；(2)实现简单，简单的相交操作就能发现 M 个局部最优解中的近似骨架对象，从而产生近似骨架结构；(3)提高了从属算法 Λ 的收敛速度，固定近似骨架降低了原始聚类问题的规模，在归约后的搜索空间上可以更加高效求解。

算法：ABRAC
输入：D，K，M
输出：Cluster
begin
(1)M 次调用启发式算法 Λ，获得聚类问题的局部最优解 $\pi_1,\pi_2,\cdots,\pi_M$；
(2)对 M 局部最优解求交获得近似骨架 $a\_\text{bone}(\pi_1,\pi_2,\cdots,\pi_M)$；
(3)固定近似骨架 $a\_\text{bone}(\pi_1,\pi_2,\cdots,\pi_M)$，获得新的搜索空间；
(4)在新的搜索空间中调用 Λ 产生聚类结果 Cluster。
End

图 3 ABRAC 算法框架

5 实验及分析

本文以快速、简单且应用广泛的K-means算法作为ABRAC算法的从属算法，得到算法ABRAC (K-means)。在26个仿真数据集和3个实际数据集上分别与以FA，McQueen，SCS及KKZ为初始策略的K-means算法进行实验对比。ABRAC及4种对比算法均由Matlab 7.0编程语言实现，并在Intel 2.0/ 1 GM/80 G兼容机、windows XP环境下执行。算法评价标准除了采用Cmp， Sep外，本文还对比了在给定初始代表点的情况下，算法 Λ 的收敛速度。I 是算法的收敛时间，其值越小，算法 Λ 的收敛速度就越快[10]。

5.1 仿真数据集

本文实验的仿真数据集是由文献[11]给出的随机数据生成器产生，该生成器可以从 http://www.jihe.net/research/ijcnn04 下载。随机数据生成器可以产生多个服从高斯分布的二维仿真数据集。这些高斯混合分布数据集是测试K-means算法的最佳数据集[11]。本文利用随机数据生成器产生了 26 个仿真数据集，其中每个数据集是这样产生的：给定 15 个相对分散的聚类代表点，利用可变方差 v 及噪声等级 r，随机产生 150, 000 个服从高斯分布的数据对象。方差 v 表示数据对象与代表点之间的差异，而噪声等级 r 则反映了数据集中包含的噪声数据的多少。限于文章篇幅，本文仅给出 3 个仿真数据集：dataset1($v=0.05, r=0.0$)、dataset10($v=0.10, r=0.2$)及 dataset20($v=0.15, r=0.4$)上的实验结果。

为了实现 ABRAC 算法，必须先调用算法 Λ 产生使目标函数取极小值的局部最优解，即聚类结果 $\pi_1,\pi_2,\cdots,\pi_M$。然后算法对 M 局部最优解执行求交操作获得近似骨架数据对象，并产生近似骨架。固定近似骨架结构构造新的搜索空间。最后利用近似骨架构造启发式算法的初始解，并调用 K-means 算法在新的搜索空间中搜索聚类结果。因此，ABRAC (K-means)算法共执行了 11 次(本文取 $M=10$)K-means 算法。为了在相同条件下对比 ABRAC (K-means) 与其他 4 种算法。以 FA、McQueen、SCS 及 KKZ 为初始策略的 K-means 算法都被执行 11 次并保留最好的聚类结果。在数据集 Dataset $i(i=1,2,\cdots,26)$ 上，每个算法的聚类个数 K 都被设定为 $\{10,15,20\}$。表 1 仅给出了 $K=15$ 时在仿真数据集 dataset1，dataset10 及 dataset20 上执行 5 种算法的对比结果。除了 ABRAC (K-means) 外，其余 4 种算法的 I 值都是 11 次执行 K-means 算法的平均收敛时间。从表中可以看出，ABRAC (K-means) 算法在 3 个数据集上得到的 Cmp 和 Sep 值都是都小于其他 4 种方法，这是因为近似骨架是从使得算法 Λ 收敛的局部最优解 $\pi_1,\pi_2,\cdots,\pi_M$ 中产生的，保存了全局最优解的高质量特性。因此，ABRAC (K-means) 算法能获得高内聚、高分离的聚类结果。另外，固定近似骨架就等于是固定了算法 Λ 的部分解，减小了算法 Λ 的搜索的范围。因此，ABRAC (K-means) 的收敛速度远远超过了其余 4 种算法。

表 1 3 种仿真数据集上的实验结果

	I	Cmp($\times 10^{-5}$)	Sep
	Dataset1, $\sigma=0.5, K=15$		
FA	86.95	5.17	0.6068
McQueen	164.95	3.28	0.6328
SCS	27.68	3.4	0.6294
KKZ	36.7	3.4	0.6294
ABRAC(K-means)	8.65	0.734	0.6162
	Dataset10, $\sigma=0.5, K=15$		
FA	204.54	6.43	0.6490
McQueen	184.90	3.28	0.6328
SCS	30.75	3.4	0.6294
KKZ	40.71	3.4	0.6294
ABRAC(K-means)	9.72	0.734	0.6062
	Dataset20, $\sigma=0.5, K=15$		
FA	627.82	2.91	0.6019
McQueen	422.55	3.31	0.5980
SCS	385.73	2.11	0.6045
KKZ	1216.46	2.16	0.6030
ABRAC(K-means)	138.00	0.355	0.5608

5.2 实际数据集

除了仿真数据集外，本文还在 3 个实际数据集上对 ABRAC (K-means)算法与其余 4 种算法进行实验比较。这 3 个实际数据集是从 UCI 机器学习知识库(MLDB)上下载的。它们分别是数据集 Iris，ImgSeg(Image Segmentation) 及 LtrRec(Letter Recognition)。表 2 给出了它们的基本描述。

表 2　3 种实际数据集的统计描述

	Iris	ImgSeg	LtrRec
实例个数	150	2,310	20,000
特征个数	4	19	16
簇数	4	7	26

在 3 个实际数据集上的实验方案与上节相似，不同之处是：(1)在实际数据集中我们将 K 设置为数据集的真实聚类个数；(2)计算 Sep 时，高斯常数的设置不同 (见表 3)。表 3 给出了分别在 3 个实际数据集上执行 5 种算法在紧密度、分离度及收敛时间 3 个方面的对比结果。从表中可以明显发现，ABRAC(K-means) 优于其它 4 种算法。

6　结论

本文利用局部最优解的交集模拟近似骨架，并通过固定近似骨架来对搜索空间进行归约。根据这一思想，提出了近似骨架导向的归约聚类算法 ABRAC。利用 K-means 算法为从属算法，本文将其与 4 种经典的启发式聚类算法进行实验对比。大量实验结果表明，新算法在收敛速度、紧密度、分离度方面均优于其余 4 种算法。

表 3　3 种实际数据集上的实验结果

	I	Cmp	Sep
Data Set: Iris, $\sigma = 1$, $K = 4$			
FA	28.34	0.0079	0.1552
McQueen	38.95	0.0062	0.1675
SCS	23.90	0.0065	0.1337
KKZ	45.20	0.0064	0.1379
ABRAC(K-means)	7.713	0.0041	0.1375
Data Set: ImgSeg, $\sigma = 500$, $K = 7$			
FA	41.98	0.4057	0.6726
McQueen	22.19	0.3400	0.6280
SCS	62.27	0.5827	0.9403
KKZ	22.91	0.4813	0.6746
ABRAC(K-means)	0.011	0.3334	0.5399
Data Set: LtrRec, $\sigma = 5$, $K = 26$			
FA	98.27	0.0056	0.1387
McQueen	143.2	0.0066	0.1320
SCS	54.07	0.0046	0.1467
KKZ	65.17	0.0075	0.1338
ABRAC(K-means)	41.69	0.0022	0.1230

参 考 文 献

[1] 孙吉贵, 刘杰, 赵连宇. 聚类算法研究. 软件学报, 2008, 19(1): 48−61.
Sun J G, Liu J, and Zhao L Y. Clustering algorithms research. *Journal of Software*, 2008, 19(1): 48−61.

[2] Drinesa P, Frieze A, and Kannan R, *et al.*. Clustering large graphs via the singular value decomposition [J]. *Machine Learning*, 2004, 56(1−3): 9−33.

[3] Jain A K and Dubes R C. Algorithms for Clustering Data [M]. Prentice Hall, Englewood Cliffs, New Jersey, 1998: 10−89.

[4] David A and Sergei V. k-means++: the advantages of careful seeding[C]. SODA 2007, New Orleans France, 2007: 1027−1035.

[5] Amir A and Lipoka D. A K-mean clustering algorithm for mixed numeric and categorical data [J]. *Data and Knowledge Engineering*, 2007, 63(2): 503−527.

[6] 江贺, 张宪超, 陈国良. 图的二分问题唯一全局最优解实例与骨架计算复杂性[J]. 科学通报, 2007, 52(17): 2077−2081.
Jiang H, Zhang X C, and Chen G L. Exclusive optimal solution instance and backbone computation complexity of graph bi-partition problem. *Chinese Science Bulletin*, 2007, 52(17): 2077−2081.

[7] 江贺, 张宪超, 陈国良, 李明楚. 二次分配问题的骨架分析与算法设计[J]. 中国科学 E 辑, 2008, 38(2): 209−222 .
Jiang H, Zhang X C, Chen G L, and Li M C. Backbone analysis and algorithm design for the quadratic assignment problem. *Science in China Series E*: *Information Sciences*, 2008, 28(2): 209−222.

[8] Valnir F J. Backbone guided dynamic local search for propositional satisfiability[C]. Proceeding of 9th International Symposium on Artificial Intelligence and Mathematics (AI & Math-06). Florida America, 2006: 100−108.

[9] Zhang W X. Configuration landscape analysis and backbone guided local search: Part I: Satisifiability and maximum satisfiability [J]. *Artificial Intelligence*, 2004, 158(1): 1−26.

[10] He J, Tan A H, and Tan C L, *et al.*. On quantitative evaluation of clustering systems[C]. Information Retrieval and Clustering. Kluwer Academic Publishers, ISBN 1-4020-7682-7, 2003.

[11] He J, Lan M, and Tan C L, *et al.*. Initialization of cluster refinement algorithms: a review and comparative study[C]. Proceedings of International Joint Conference on Neural Networks (IJCNN). Budapest Hungary, 2004: 297−302.

宗　瑜:　男，1976 年生，博士生，研究领域为数据挖掘.
李明楚:　男，1963 年生，教授，研究领域为数据挖掘、入侵检测.
江　贺:　男，1980 年生，副教授，研究领域为智能计算、数据挖掘.