基于模糊 K-harmonic means的谱聚类算法

汪 中[1,2], 刘贵全[1,2], 陈恩红[1,2]
(1. 中国科学技术大学 计算机科学与技术学院,安徽 合肥 230027; 2.安徽省计算与通讯软件重点实验室,安徽 合肥 230027)

摘 要:谱聚类作为一种有效的方法广泛应用于机器学习.通过分析谱聚类初始化敏感的实质,引入对初值不敏感的模糊 K-harmonic means算法来克服这一缺点,提出一种基于模糊 K-harmonic means的谱聚类算法 (FKHM-SC).与传统谱聚类算法以及对初值敏感的 K-means、FCM算法相比,改进算法不仅可以识别有挑战性的人工数据,并且可以得到稳定的聚类中心和聚类结果,同时提高了聚类的精确度.实验结果表明了该算法的有效性和可行性.

关键词:谱聚类;模糊 K-harmonic means;初始化敏感;聚类中心
中图分类号: TP311 **文献标识码**: A **文章编号**: 1673-4785(2009)02-0095-05

A spectral clustering algorithm based on fuzzy K-harmonic means

WANG Zhong[1,2], LIU Gui-quan[1,2], CHEN En-hong[1,2]
(1. School of Computer Science, University of Science and Technology of China, Hefei 230027, China; 2. Key Laboratory of Software in Computing and Communication, Hefei 230027, China)

Abstract: Spectral clustering is an effective method that is widely used in machine learning. After analyzing the essence of initialization sensitivity in spectral clustering, the fuzzy K-harmonic means (FKHM) algorithm was considered to conquer spectral clustering's shortcomings, then an spectral clustering algorithm based on FKHM was developed. Compared with the traditional spectral algorithm and the fuzzy c-means (FCM) algorithm, the suggested algorithm is more sensitive to initial values. The suggested algorithm can not only identify challenging artificial data, but also find stable cluster centers and clustering results, considerably improving clustering precision. Experiments showed that it is an effective and feasible way to improve the performance of spectral clustering algorithms.

Keywords: spectral clustering; fuzzy K-harmonic means; initialization sensitivity; cluster centers

聚类分析是数据挖掘、机器学习、模式识别等众多领域的重要工具.它在执行过程中没有任何关于类别的先验知识和假设,因而被称作是一种无监督的学习方法.近年来产生了大量解决该问题的相关算法,现有的基于产生式模型和基于中心的聚类方法仅在具有凸形结构的数据上有好的效果,且容易陷入局部最优.

为了能在任意形状的样本结构上聚类,且收敛于全局最优,谱聚类算法被广泛应用.谱聚类的思想来源于谱图划分[1],是一种流行的高性能计算方法.该方法基于两点间的相似关系,利用数据的相似矩阵的特征向量进行聚类,通过特征分解可以获得聚类判据在放松了的连续域中的全局最优解.谱聚类算法适用于非测度空间,算法与数据点维数无关,而仅与数据点个数有关,因而可以避免由特征向量的过高维数所造成的奇异性问题.谱聚类算法尽管在实践中取得了很好的效果,但是算法本身仍存在许多值得研究的问题.文献[2]指出当聚类数目大于实际聚类数时,多路谱聚类方法的效果很差;Fischer等提出了依赖于背景的相似性度量方法和尺度参数问题[3];由于谱方法的计算复杂度相当高,Fowlkes等提出使用 Nystrom 逼近方法减少求解特征问题的计算复杂度[4];针对谱聚类初始化敏感的特点,Ekin等提出使用对初始值不敏感的方法[5].

本文提出一种基于模糊 K-harmonic means的谱聚类算法,其主要思想是:首先分析不同数据输入顺序对相似性矩阵、构造矩阵和生成矩阵的影响,得出

收稿日期: 2008-12-16.
基金项目:国家自然科学基金资助项目 (60775037);教育部新世纪优秀人才支持计划资助项目 (NCET-05-0549).
通信作者:汪 中. E-mail: wzspb@mail.ustc.edu.cn.

谱聚类初始化敏感的实质原因,通过引入模糊 K-Harmonic means(FKHM)算法来解决对初值敏感的问题,从而得到更稳定的聚类性能. 在人工数据和真实数据上进行实验模拟,结果表明,FKHM-SC算法相对于原有谱聚类算法(SC)和传统的 K-means(KM)算法和 fuzzy C-means(FCM)算法在聚类性能和稳定性上有了显著的提高.

1 谱聚类算法的初始化敏感分析

谱聚类算法将数据聚类看成是一个无向图的多路划分问题,由于图划分问题的组合本质,求图划分判据的最优解是一个 NP难题[6]. 一个有效的求解方法是考虑问题的连续放松形式,将原有问题转换为求解矩阵的特征值和特征向量问题,利用这些特征向量构造一个简化了的数据空间,在该空间中的数据的分布结构更加明显,代表性算法有 Ng等人提出的 NJW 算法[7].

NJW 算法本质上是利用相似矩阵的特征向量进行聚类,选取构造矩阵的前 K个最大特征值所对应的特征向量,从而在 K维空间中构成与原数据一一对应的表述,进而在 K维空间中利用 K-means或其他简单算法进行聚类.

现有的各种谱聚类算法的差异之处在于: 1)构造的相似矩阵不同; 2)使用的特征向量不同; 3)从特征向量获得最终的聚类方法不同; 4)从连续变量放松到离散变量的方法不同. NJW 算法是基于上述的第 3种多类实现方法给出的一种简单有效方法,由于传统的 K-means算法对初始值敏感,故 NJW 算法对初值敏感是否归结于最终的聚类方法. 下面给出谱聚类算法对初值敏感的证明.

定理 1 设数据集 $S = \{s_1, \cdots, s_n\}$,以 2种不同顺序输入得到的相似矩阵为 A_1、A_2,对角矩阵为 D_1、D_2,构造矩阵为 L_1、L_2,生成矩阵为 Y_1、Y_2,则 A_1 和 A_2 相似,D_1 和 D_2 相似,L_1 和 L_2 相似,Y_1 和 Y_2 相似.

证 明 设以 $s_1, s_2, \cdots, s_n$ 顺序输入得到的相似矩阵 $A_1 = \begin{pmatrix} \mathbf{0} & A_{12} & \cdots & A_{1n} \\ A_{21} & \mathbf{0} & \cdots & A_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ A_{n1} & A_{n2} & \cdots & \mathbf{0} \end{pmatrix}$,以 $s_1, s_2, \cdots, s_n$ 逆序输入得到的相似矩阵

$$A_2 = \begin{pmatrix} \mathbf{0} & A_{n(n-1)} & \cdots & A_{n1} \\ A_{(n-1)n} & \mathbf{0} & \cdots & A_{(n-1)1} \\ \cdots & \cdots & \cdots & \cdots \\ A_{1n} & A_{2n} & \cdots & \mathbf{0} \end{pmatrix},$$

由矩阵知识易知,矩阵 A_1 经过若干次初等行或者列变换则可以得到矩阵 A_2,故矩阵 A_1 和 A_2 相似. 同理,当以任意顺序输入得到的相似矩阵 A_2 均与 A_1 相似. 由于矩阵 D_1 为对角矩阵,其主对角元素为相似矩阵 A_1 的相应各行元素之和,故矩阵 D_1 和 D_2 相似.

因为 $L_1 = D_1^{-1/2}/A_1 D_1^{-1/2}$, $L_2 = D_2^{-1/2} A_2 D_2^{-1/2}$,且 D_1、D_2 为对角矩阵,由上面的相似关系可知矩阵 L_1 和 L_2 相似,其对应的特征向量 x_1 经过若干次初等变换可以得到特征向量 x_2,则生成矩阵 Y_1 和 Y_2 亦相似. 故结论成立.

2 改进的谱聚类算法

通过定理 1可知,不同输入顺序得到的相似矩阵 A和最终的生成矩阵 Y是相似的,故谱聚类算法对初始值敏感的实质在于 NJW 算法从特征向量获得最终的聚类方法是否对初始化敏感. 传统的 K-means和 FCM算法对初始中心敏感[5],对于不同的初始值,可能得到不同的聚类结果. 在此基础上,通过引入对初值不敏感的模糊 K-harmonic means算法来克服这一缺点.

2.1 模糊 K-harmonic means算法

模糊 K-harmonic means(FKHM)算法[8]是一种基于中心的聚类算法,将模糊概念应用到 KHM 算法[9]中,应用数据点对不同聚类的隶属度对目标函数中的距离测度进行模糊加权. 设 $X = \{x_i \mid i = 1, \cdots, n\}$ 为 n个数据点的集合, $C = \{C_j \mid j = 1, \cdots, k\}$ 为 k个聚类中心的集合, $d_{ij} = \| X_i - C_j \|$ 表示为数据对象到聚类中心的距离表示,采用欧式距离表示. 其目标函数为

$$E_{\mathrm{FHKM}} = \sum_{i=1}^{n} \frac{k}{\sum_{j=1}^{k} \frac{1}{w_{ij}^{\alpha} d_{ij}^2}}. \tag{1}$$

式中: $w_{ij} \in [0, 1]$ 表示数据对象 X_i 对聚类中心 C_j 的隶属度,且 $\sum_{j=1}^{k} w_{ij} = 1, \alpha \geq 0$ 为模糊算子. 隶属度函数和聚类中心更新公式为

$$w_{ij} = \frac{(1/d_{ij}^2)^{1/(\alpha+1)}}{\sum_{z=1}^{k} (1/d_{iz}^2)^{1/(\alpha+1)}}, \tag{2}$$

$$C_j = \frac{\sum_{i=1}^{n} \frac{w_{ij}^{\alpha}}{\sum_{z=1}^{k} \left(\frac{w_{ij}^{\alpha} d_{ij}^2}{w_{iz}^{\alpha} d_{iz}^2}\right)^2} X_i}{\sum_{i=1}^{n} \frac{w_{ij}^{\alpha}}{\sum_{z=1}^{k} \left(\frac{w_{ij}^{\alpha} d_{ij}^2}{w_{iz}^{\alpha} d_{iz}^2}\right)^2}}. \tag{3}$$

算法描述为:初始化聚类中心,通过式(2)和式

(3)分别计算隶属度和聚类中心,然后进行迭代,直到最大的迭代次数或者目标函数式(1)稳定为止.该算法的时间复杂度为 $O(m \times n \times k)$,其中 m 为数据属性的个数.

2.2 基于模糊 K-harmonic means的谱聚类算法(FKHM-SC)

由于传统谱聚类算法对初始值敏感,本文在 NJW 算法的框架下提出基于模糊 K-harmonic means 的谱聚类算法,具体实现步骤如下:

输入: n个数据点 $S = \{s_1, \cdots, s_n\}$,聚类数目 k

输出:数据点集的划分

步骤:

1) 构造相似矩阵 $A \in R^{n \times n}$,其中 $A_{ij} = \exp(-\| s_i - s_j \|^2 / 2\sigma^2)$, $i \neq j$, $A_{ii} = 0$. 其中 σ 是参数;

2)构造矩阵 $L = D^{-1/2} A D^{-1/2}$. 其中 D 是对角矩阵,对角元素为 $D_{ii} = \sum_{j=1}^{n} A_{ij}$;

3)采用 Nystrom 逼近方法[4]计算 L 的前 k个特征值所对应的特征向量 $x_1, x_2, \cdots, x_n$ (重复特征值取其相互正交的特征向量) 构造矩阵 $X = [x_1\ x_2\ \cdots\ x_n] \in R^{n \times k}$,规范化矩阵 X的行向量,得到矩阵 $Y = X_{ij} / (\sum_j X_{ij}^2)^{1/2}$;

4)将矩阵 Y的每行当作 R^k空间中一点,采用上述对初值不敏感的 FKHM 算法聚为 k类,如果 k的第 i行数据为第 j类,则原数据 s_i划分到第 j类.

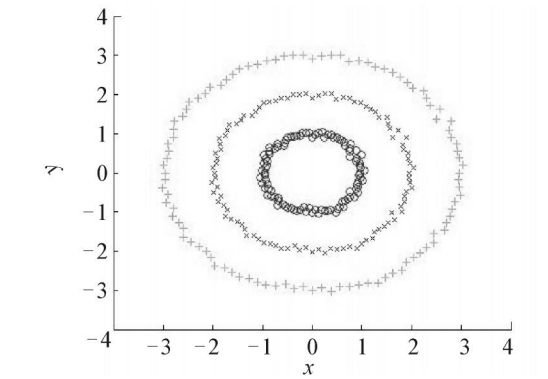
由于谱聚类初始化敏感的特点,考虑到传统的 K-means和 FCM 算法依赖于初始 k个对象的选择;而 FKHM 算法通过对数据点到聚类中心的调和平均进行加权,动态加权对初值不敏感起到重要的作用,故改进算法在理论上能得到更稳定的聚类性能.利用 Nystrom 逼近方法求解矩阵的特征值和特征向量减少了计算的复杂度.

3 实　验

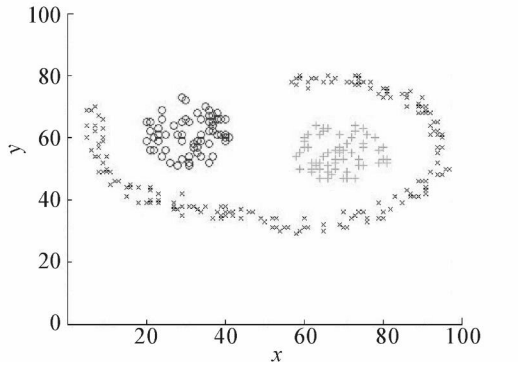
3.1 人工数据

文献[7]给出一些有挑战性的人工数据聚类问题.图 1分别给出了 FKHM-SC在 4组人工数据集上的聚类结果图,FKHM-SC可以成功地识别这 4组有聚类数据问题,同时给出了相应的核参数 σ的值.

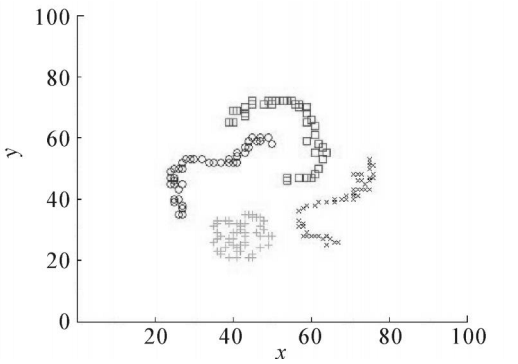
FKHM-SC可以成功地识别上面 4组数据,利用多个特征向量构造了一个简化的数据空间,在该空间中的数据的分布结构更加明显,通过每组数据的特征向量图可以直观地观察数据的分布.



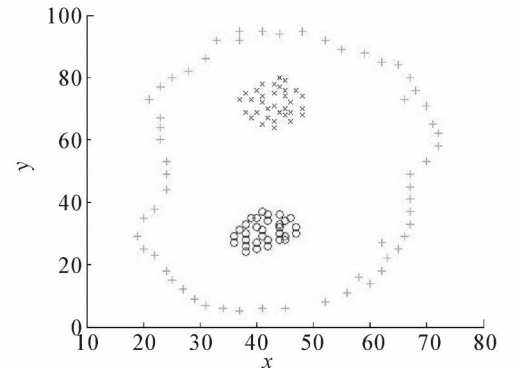
(a) three circles-joined聚类结果 (σ = 0.05)



(b) line and balls聚类结果 (σ = 1)



(c) squiggles聚类结果 (σ = 1)



(d) blobs and circles聚类结果 (σ = 2)

图 1　FKHM-SC在 4组人工数据集上的聚类结果图

Fig 1　Cluster results of FKHM-SC on four artificial data sets

3.2 真实数据

聚类算法性能的评价一直是一个具有挑战性的问题,为了评价改进算法性能,采用 ARI(adjusted rand index)[10]来评价算法,它将聚类划分看作是样本之间的一种关系,每一个样本要么划分在同一类,要么在不同类.准确度就等于正确匹配对数与两两比较次数的比值,其中准确度在 0和 1之间取值,其值越大表明聚类结果与被分析数据越匹配,即算法的有效性越高.

为了验证算法的有效性,采用 UCI数据库[11]上的 Iris、Glass、Ionosphere、Sonar这 4组数据作为测试数据. UCI数据库是一个专门用于测试机器学习、数据挖掘算法的数据库,库中的数据都有确定的分类,因此可以直观的表示聚类结果的质量. 表 1给出真实数据集的数据特征.

表 1 真实数据集

Table 1 Real data sets

数据集	属性数目	类的个数	数据点总数
Iris	4	3	150
Glass	9	6	214
Ionosphere	34	2	351
Sonar	60	2	208

3.2.1 稳定性仿真

在稳定性仿真实验中,采用 Iris数据集作为测试数据. 该数据集共有 3类,其中第 1类和其他 2类有较好的分离,另外两类之间存在交迭. 它的实际聚类中心位置分别为 (6.588 2.974 5.552 2.026)、(5.006 3.418 1.464 0.244)、(5.936 2.770 4.260 1.326)分别采用 KM、FCM和 FKHM-SC算法对 Iris数据集聚类 10次,每次 3种算法均采用相同的随机初始值,取其平均值,聚类中心结果如表 2.

表 2 3种算法对 Iris数据聚类的结果

Table 2 Cluster results of three algorithms in Iris data

聚类算法	聚类中心			
KM	6.721 8	3.054 2	5.512 1	1.990 6
	5.023 2	3.439 3	1.467 3	0.247 4
	5.772 5	2.718 0	4.082 2	1.275 8
FCM	6.081 4	3.070 2	5.602 5	2.013 4
	5.059 3	3.217 4	1.800 0	0.379 9
	6.074 4	2.861 8	4.640 1	1.553 3
FKHM-SC	6.595 7	2.993 4	5.354 9	1.908 8
	5.009 1	3.411 3	1.482 6	0.250 4
	5.756 0	2.707 2	4.222 0	1.326 0

从表 2中可以看出,FKHM-SC算法明显更接近于实际的聚类中心,即更接近与原始数据的类别分布;而 KM和 FCM算法由于对初始值敏感,故稳定性较差. FKHM-SC算法采用对初值不敏感的 FKHM算法,实验结果表明 FKHM-SC算法提高了聚类结果的稳定性.

3.2.2 聚类性能仿真

图 2给出了 KM、FCM、SC和 FKHM-SC 4种算法在表 1中 4个真实数据集上的聚类性能曲线图. 从图 2的 4个图中可以观察到:

1) KM和 FCM算法在 4个数据集上出现了明显的性能波动,SC算法在 Glass数据集上稳定性较差,而 FKHM-SC算法性能曲线平稳. 故可以说明 KM和 FCM算法对初值敏感性较强,SC算法对初值敏感性稍弱,而 FKHM-SC算法没有出现任何波动,稳定性较好.

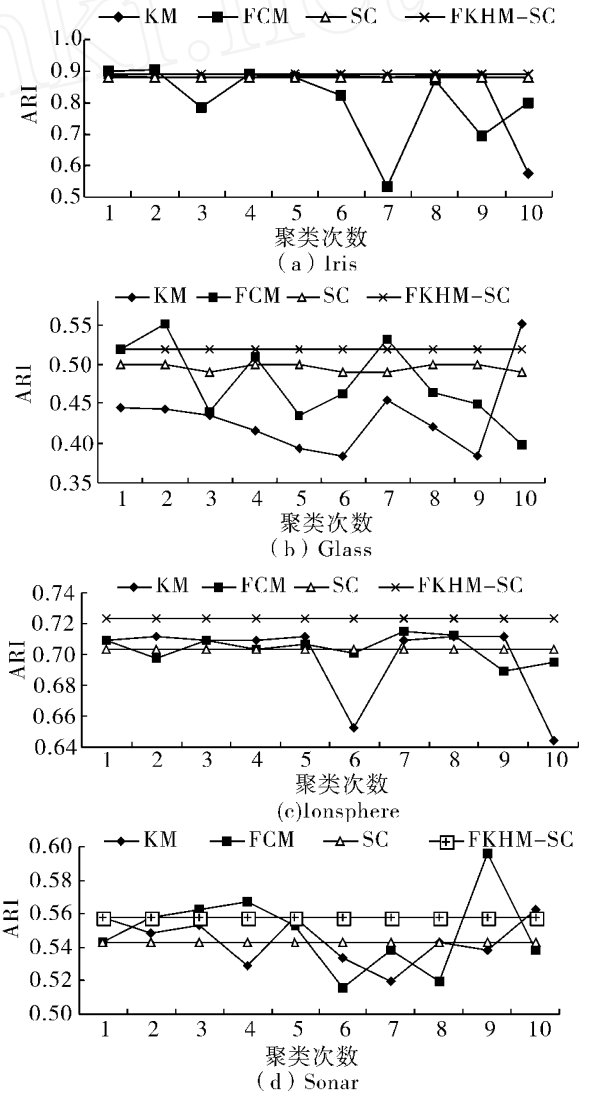


图 2 4种算法的对比实验结果

Fig 2 Comparisons of four algorithms

2)在 4个数据集上,FKHM-SC算法相对于 KM和 SC算法在聚类精确度上有显著的提高,且稳定性较高,即平均聚类性能优于 KM和 SC算法,而在 Glass和 Sonar数据集上,虽然 FCM算法的个别精确度高于 FKHM-SC算法,但是从图中可以看到,FCM算法波动性较大,总体平均性能仍然低于 FKHM-SC算法.这说明,FKHM-SC算法不仅稳定性较好,并且取得更高的聚类精确度.

4　结束语

通过分析传统谱聚类算法的对初值敏感的实质,提出一种基于 FKHM的谱聚类算法(FKHM-SC).实验表明,FKHM-SC算法在人工数据和真实数据上均取得较好的结果.相对于传统的 K-means、FCM和谱聚类算法,该算法不仅具有较高的稳定性,且获得的聚类中心与实际聚类中心更为接近,从而聚类性能有了显著的提高.下一步工作将算法应用到实际问题中.

参考文献:

[1]FIEDLER M. Algebraic connectivity of graphs[M]. Praha: Czechoslovak Mathematical Journal, 1973: 298-305.

[2]VERMA D, MEILA M. A comparison of spectral clustering algorithms[R]. University of Washington, 2003.

[3]FISCHER I, POLAND J. Amplifying the block matrix structure for spectral clustering[C]//Proceedings of the 14th Annual Machine Conference of Belgium and the Netherlands. Manno, Switzerland, 2005: 21-28.

[4]FOWLKES C, BELONGIE S, CHUNG F, et al. Spectral grouping using the Nyström method[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 26(2): 217-225.

[5]EKIN A, PANKANTI S, HAMPAPUR A. Initialization-independent spectral clustering with applications to automatic video analysis[C]//Proc of IEEE ICASSP. Montreal, Canada, 2004: 641-644.

[6]SHI J B, MALIK J. Normalized cuts and image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905.

[7]NG A Y, JORDAN M I, WEISS Y. On spectral clustering: analysis and an algorithm[C]//Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2002: 897-856.

[8]赵　恒,杨万海,张高煜.模糊 K-Harmonic Means聚类算法[J].西安电子科技大学学报, 2005, 32(4): 603-606.
ZHAO Heng, YANG Wanhai, ZHANG Gaoyu. Fuzzy K-harmonic means clustering algorithm[J]. Journal of XiDian University, 2005, 32(4): 603-606.

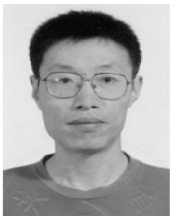
[9]ZHANG B, HSU M, DAYAL U. K-harmonic means—a data clustering algorithm[EB/OL]. [2006-01-12]. http://hpc.isti.cnr.it/~palmeri/datam/articles/HPL-1999-124.pdf.

[10]HANDL J, KNOWLES J. An evolutionary approach to multi-objective clustering[J]. IEEE Transactions on Evolutionary Computation, 2007, 11(1): 56-76.

作者简介:

汪　中,男,1984年生,硕士研究生,主要研究方向为数据挖掘、机器学习.

刘贵全,男,1970年生,副教授,博士,主要研究方向为数据挖掘、人工智能、网络安全等.2003年获安徽省科技成果三等奖.发表学术论文 50余篇.



陈恩红,男,1968年生,教授,博士生导师,主要研究方向为数据挖掘与机器学习、网络信息处理等.1995年获中国科学院院长奖学金优秀奖,1996年获中国科学技术大学惠普信息科学青年教师奖,2000年获王宽诚育才奖、安徽省科技进步二等奖,2004年获安徽省科技进步三等奖、中国科技大学优秀教学成果二等奖,2006年获王宽诚育才奖一等奖.发表学术论文 90余篇.