基于快速地标采样的大规模谱聚类算法

叶 茂* 刘文芬
(解放军信息工程大学 郑州 450002)
(数学工程与先进计算国家重点实验室 郑州 450002)

摘 要：为避免传统谱聚类算法高复杂度的应用局限，基于地标表示的谱聚类算法利用地标点与数据集各点间的相似度矩阵，有效降低了谱嵌入的计算复杂度。在大数据集情况下，现有的随机抽取地标点的方法会影响聚类结果的稳定性，k均值中心点方法面临收敛时间未知、反复读取数据的问题。该文将近似奇异值分解应用于基于地标点的谱聚类，设计了一种快速地标点采样算法。该算法利用由近似奇异向量矩阵行向量的长度计算的抽样概率来进行抽样，同随机抽样策略相比，保证了聚类结果的稳定性和精度，同 k均值中心点策略相比降低了算法复杂度。同时从理论上分析了抽样结果对原始数据的信息保持性，并对算法的性能进行了实验验证。

关键词：地标点采样；大数据；谱聚类；近似奇异值分解
中图分类号：TP181 **文献标识码**：A **文章编号**：1009-5896(2017)02-0278-07
DOI: 10.11999/JEIT160260

Large Scale Spectral Clustering Based on Fast Landmark Sampling

YE Mao LIU Wenfen
(*PLA Information Engineering University*, *Zhengzhou* 450002, *China*)
(*State Key Laboratory of Mathematical Engineering and Advanced Computing*, *Zhengzhou* 450002, *China*)

Abstract: The applicability of traditional spectral clustering is limited by its high complexity in large-scale data sets. Through construction of affinity matrix between landmark points and data points, the Landmark-based Spectral Clustering (LSC) algorithm can significantly reduce the computational complexity of spectral embedding. It is vital for clustering results to apply the suitable strategies of the generation of landmark points. While considering big data problems, the existing generation strategies of landmark points face some deficiencies: the unstable results of random sampling, along with the unknown convergence time and the repeatability of data reading in k-means centers method. In this paper, a rapid landmark-sampling spectral clustering algorithm based on the approximate singular value decomposition is designed, which makes the sampling probability of each landmark point decided by the row norm of the approximate singular vector matrix. Compared with LSC algorithm based on random sampling, the clustering result of new algorithm is more stable and accurate; compared with LSC algorithm based on k-means centers, the new algorithm reduces the computational complexity. Moreover, the preservation of information in original data is analyzed for the landmark-sampling results theoretically. At the same time, the performance of new approach is verified by the experiments in some public data sets.

Key words: Landmark sampling; Big data; Spectral clustering; Approximate singular value decomposition

1 引言

聚类分析可将数据集按照相似性分成子集，使得人们能根据分类结果找出数据的内在联系，是模式识别、数据挖掘的主要方法之一[1]。传统聚类算法(如 k 均值等)在非凸数据集上效果不佳，这使得适用于非凸数据集和能检测线性不可分簇的谱聚类算法[2,3]成为了聚类分析中的研究热点。但是，传统的谱聚类算法涉及构造相似度矩阵和对相应的拉普拉斯矩阵特征分解，需要$\mathrm{O}(n^2)$的空间复杂度和$\mathrm{O}(n^3)$的时间复杂度，这对于大规模数据集来说是难以承受的计算负担。

为提升谱聚类算法的扩展性，一个自然的想法就是设计可以减少特征分解复杂度的算法。2004 年，Fowlkes 等人[4]改进 Nyström 方法并将其用于谱聚

收稿日期：2016-03-21；改回日期：2016-07-18，网络出版：2016-09-30
*通信作者：叶茂 yemaoxxgc@163.com
基金项目：国家 973 计划(2012CB315905)， 国家自然科学基金(61502527, 61379150)
Foundation Items: The National 973 Program of China (2012CB315905), The National Natural Science Foundation of China (61502527, 61379150)

类，实现了快速近似特征分解。随后，Li 等人[5,6]又用近似奇异值分解(Singular Value Decomposition, SVD)方法提升了 Nyström 方法中特征分解的效率。而丁世飞等人[7]则设计了一种自适应采样的方法，改进了 Nyström 谱聚类的聚类效果。此外，Yan 等人[8]还提出了一个快速近似谱聚类的框架：先选择代表点，然后对代表点进行谱聚类，并将分类关系扩展到与代表点关联的其他点上。

2011 年，Chen 等人[9]提出了基于地标点的谱聚类(Landmark-based Spectral Clustering, LSC)算法，指出该方法适用于大规模数据集，并且性能要比 Nyström 方法和 Yan 的方法[8]好，并在随后给出了相关理论分析[10]。LSC 算法通过数据集点与地标点之间的相似度矩阵的乘积来近似得到整体的相似度矩阵，然后利用近似性质实现快速特征分解。但该方法用随机采样确定地标点，抽样结果不稳定，在大数据集时容易出现样本点集中于某一区域的情况。

当前，随机映射由于可在降低数据规模的同时保持大部分原始信息而被广泛用于聚类算法中[11−13]。本文利用随机映射得到近似 SVD 算法，然后由分解得到的近似奇异向量矩阵的行向量长度确定各点在数据集中的权重并计算抽样概率，以此得到快速抽样算法。通过理论分析，得出该抽样算法的抽样误差被限制在一个较小的界内，保证了抽样结果对原始数据的信息保持性。实验结果表明基于该抽样方法的LSC算法聚类结果要比基于随机抽样的算法稳定且聚类精度更高，比基于 k 均值中心点的方法运行速度快，从而验证了新方法的性能。

2　基础知识

本节先给出本文所用的一些矩阵相关符号，然后简述 LSC 算法和应用于快速采样的近似 SVD 算法。

2.1 矩阵的相关符号

记数据矩阵为 $\boldsymbol{A}\in\mathbb{R}^{n\times d}$，其中 n 是数据样本点个数，d 为数据特征个数，$\boldsymbol{A}^{\mathrm{T}}$ 为矩阵 $\boldsymbol{A}$ 的转置。记 $\boldsymbol{A}_{(i)}$ 为数据矩阵的第 i 行，$\boldsymbol{A}^{(j)}$ 为数据矩阵的第 j 列。分别记 $\|\boldsymbol{A}\|_2=\max_{\|\boldsymbol{x}\|=1}\|\boldsymbol{A}\boldsymbol{x}\|_2$ 和 $\|\boldsymbol{A}\|_F=\sqrt{\sum a_{ij}^2}$ 是矩阵 $\boldsymbol{A}$ 的谱范数和 Frobenius 范数。关于矩阵范数有如下性质。

性质 1　对于有恰当维数的矩阵 $\boldsymbol{A},\boldsymbol{B}$，两个矩阵乘积的范数满足：$\|\boldsymbol{A}\boldsymbol{B}\|_F\le\|\boldsymbol{A}\|_F\|\boldsymbol{B}\|_2$。

设矩阵 $\boldsymbol{A}$ 的秩为 $\rho\le\min\{n,d\}$，将矩阵的奇异值分解表示为 $\boldsymbol{A}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}$，其中 $\boldsymbol{U}\in\mathbb{R}^{n\times\rho}$ 和 $\boldsymbol{V}\in\mathbb{R}^{d\times\rho}$ 的列向量是正交的单位向量，$\boldsymbol{\Sigma}\in\mathbb{R}^{\rho\times\rho}$ 是对角线上数值按降序排列的对角矩阵。令 $\boldsymbol{U}_k,\boldsymbol{V}_k$ 分别是 $\boldsymbol{U},\boldsymbol{V}$ 的前 k 个列向量，对角阵 $\boldsymbol{\Sigma}_k\in\mathbb{R}^{k\times k}$ 是 $\boldsymbol{\Sigma}$ 前 k 个部分，则易知对于谱范数和 Frobenius 范数，矩阵 $\boldsymbol{A}_k=\boldsymbol{U}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k^{\mathrm{T}}$ 均是 $\boldsymbol{A}$ 的最优 k 秩近似，即 $\|\boldsymbol{A}-\boldsymbol{A}_k\|_\xi=\min_{\boldsymbol{X}\in\mathbb{R}^{n\times d},\,\mathrm{rank}(\boldsymbol{X})=k}\|\boldsymbol{A}-\boldsymbol{X}\|_\xi$，$\xi=F,2$。矩阵 $\boldsymbol{A}$ 的伪逆用 SVD 的形式表示为 $\boldsymbol{A}^\dagger=\boldsymbol{V}\boldsymbol{\Sigma}^{-1}\boldsymbol{U}^{\mathrm{T}}\in\mathbb{R}^{d\times n}$。因此，当 $\mathrm{rank}(\boldsymbol{A}\boldsymbol{A}^\dagger)=\rho=n$ 时，等式 $\boldsymbol{A}\boldsymbol{A}^\dagger=\boldsymbol{I}$。

记由矩阵 $\boldsymbol{V}\in\mathbb{R}^{d\times\rho}$ 的列向量形成的子空间为 $\mathrm{colspan}(\boldsymbol{V})$，记矩阵 $\boldsymbol{B}=\boldsymbol{A}^{\mathrm{T}}\in\mathbb{R}^{d\times n}$ 的列在该子空间上的投影为 $\pi_{\boldsymbol{V}}(\boldsymbol{B})$，记由 $\pi_{\boldsymbol{V}}(\boldsymbol{B})$ 列向量形成的 $\boldsymbol{B}$ 最优 k 秩近似为 $\pi_{\boldsymbol{V},k}(\boldsymbol{B})$，则有 $\pi_{\boldsymbol{V},k}(\boldsymbol{B})=\left(\pi_{\boldsymbol{V}}(\boldsymbol{B})\right)_k$。

2.2 基于地标表示的谱聚类算法

LSC 算法[9]主要思想是通过地标点来实现相似度矩阵的快速构造和特征分解。具体算法流程如表 1 的算法 1 所示。

表 1　LSC 算法

算法 1　基于地标的谱聚类(LSC)[9,10]

输入　n 个 d 维数据点形成的矩阵 $\boldsymbol{A}\in\mathbb{R}^{n\times d}$，簇个数 k，地标点个数 p；

输出　k 个簇；

(1) 产生 p 个地标点 $\boldsymbol{la}_1,\boldsymbol{la}_2,\cdots,\boldsymbol{la}_p$ (通常由均匀随机抽样产生)；

(2) 在地标点和数据点间构造稀疏的相似度矩阵 $\boldsymbol{Y}\in\mathbb{R}^{p\times n}$，使得

$$y_{ji}=\frac{K_h\left(\boldsymbol{a}_i,\boldsymbol{la}_j\right)}{\sum\limits_{j'\in\mathrm{LA}_{(i)}}K_h\left(\boldsymbol{a}_i,\boldsymbol{la}_{j'}\right)},\quad j\in\mathrm{LA}_{(i)}$$

其中，$\mathrm{LA}_{(i)}$ 是点 $\boldsymbol{a}_i$ 的近邻地标点指标集，$K_h(\cdot)$ 是带宽为 h 的核函数；

(3) 利用 SVD 算法得到矩阵 $\widehat{\boldsymbol{Y}}$ 的前 k 个右奇异向量 $\boldsymbol{Q}=[\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_k]\in\mathbb{R}^{n\times k}$，其中 $\widehat{\boldsymbol{Y}}=\widehat{\boldsymbol{D}}^{-1/2}\boldsymbol{Y}$，$\widehat{\boldsymbol{D}}\in\mathbb{R}^{p\times p}$ 是一个对角矩阵且 $\hat{d}_{ii}=\sum_j y_{ij}$；

(4) 视矩阵 $\boldsymbol{Q}$ 的每行为一个数据点并对其进行 k 均值聚类，得到聚类结果。

从算法流程可以看出，步骤 3 实现了相似度矩阵 $\boldsymbol{W}$ 的近似构造 $\boldsymbol{W}\approx\widehat{\boldsymbol{Y}}^{\mathrm{T}}\widehat{\boldsymbol{Y}}$：对 $\widehat{\boldsymbol{Y}}$ 计算右奇异向量矩阵，此过程等价于对矩阵 $\widehat{\boldsymbol{Y}}^{\mathrm{T}}\widehat{\boldsymbol{Y}}$ 进行特征分解得到特征向量。由于 $\widehat{\boldsymbol{Y}}\in\mathbb{R}^{p\times n},p\ll n$，所以相似度矩阵分解的时间复杂度从 $\boldsymbol{W}$ 的特征分解时间 $O(n^3)$ 减少到了 $\widehat{\boldsymbol{Y}}$ 的 SVD 时间 $O(p^2n)$，空间复杂度从存储 $\boldsymbol{W}$ 所需的 $O(n^2)$ 减少到了存储 $\widehat{\boldsymbol{Y}}$ 所需的 $O(np)$，时间、空间复杂度均比原始谱聚类算法显著减少。

2.3 近似 SVD 算法

基于矩阵重构的采样始于 1988 年 Frieze 等人[14]的开创性成果：给定矩阵，通过与列向量欧几里得

长度平方成比例的概率抽样少的列，可快速得到原始矩阵的低秩近似。随后，文献[15,16]以与奇异向量矩阵的整行长度成比例的概率抽样矩阵的列，使得近似效果得到明显提升。2014 年，Boutsidis 等人[17]通过基于随机映射的近似 SVD 算法[18]来改进抽样算法效率，并得到了渐近最优抽样算法。本文所用的快速采样算法就是基于近似 SVD 算法得到的，SVD 具体流程如表 2 的算法 2 所示。

表 2 近似 SVD 算法

算法 2 相对误差近似 SVD 算法[18]

输入 数据矩阵 $\boldsymbol{A}\in\mathbb{R}^{n\times d}$，整数 $k<\mathrm{rank}(\boldsymbol{A})$，压缩后的数据规模 r；

输出 矩阵 $\boldsymbol{A}$ 的近似前 k 个左奇异向量 $\boldsymbol{Z}\in\mathbb{R}^{n\times k}$；

(1) 产生矩阵 $\boldsymbol{R}\in\mathbb{R}^{n\times r}$，其中每个矩阵元素均独立地以 0.5 概率取 ±1；

(2) 计算 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}\in\mathbb{R}^{d\times r}$；

(3) 对 $\boldsymbol{A}^{\mathrm{T}}\boldsymbol{R}$ 的列向量标准正交化，得到矩阵 $\boldsymbol{F}$；

(4) 对 $\boldsymbol{F}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\in\mathbb{R}^{r\times n}$ 进行 SVD 得到其前 k 个右奇异向量组成 $\boldsymbol{Z}$。

算法 2 的思想在于用随机映射对数据进行压缩，并使得在降低矩阵规模后仍保持原始矩阵的主要信息。Sarlos[19]指出，经过随机映射压缩数据，若压缩后数据规模满足特定参数，则该近似 SVD 算法所得到的近似奇异向量在最优低秩近似上能保持与精确的奇异向量接近的效果。

引理 1[19] 令 $\boldsymbol{A}\in\mathbb{R}^{n\times d}$，$\pi(\cdot)$ 是在 2.1 节定义的投影算子。如果 $0<\varepsilon\le1$，$\boldsymbol{R}'=(1/\sqrt{r})\boldsymbol{R}$，其中 $\boldsymbol{R}\in\mathbb{R}^{n\times r}$ 是满足算法 2 所要求的矩阵，且 $r=\Theta(k/\varepsilon+k\lg k)$，则至少以 $1/2$ 的概率，有

$$\left\|\boldsymbol{A}^{\mathrm{T}}-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}\right\|_F\le\left\|\boldsymbol{A}^{\mathrm{T}}-\pi_{\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{R}'),k}\left(\boldsymbol{A}^{\mathrm{T}}\right)\right\|_F\le(1+\varepsilon)\left\|\boldsymbol{A}^{\mathrm{T}}-\left(\boldsymbol{A}^{\mathrm{T}}\right)_k\right\|_F$$

成立。

3 基于快速地标采样的大规模谱聚类算法

相比于传统谱聚类算法，LSC 算法在时间和空间复杂度上均有很大优势，并且在聚类效果上也令人满意。作为算法的关键，地标点的选取在很大程度上影响了聚类效果。常用的方式是均匀随机采样，在大规模数据集上随机抽样的不稳定性很可能会导致所抽样本点集中于某一区域，这将使得算法聚类效果变差。

LSC 算法的思想是通过地标点的线性组合来实现所有数据点的表示，然后通过地标点与数据点的相似性度量来给出各个数据点之间的相似性度量，因此地标点的特征在于“代表性”。文献[15,16]提出了一种可抽取具有“代表性”数据点的方法：采用以与奇异向量矩阵整行长度平方成比例的概率抽样数据，使得较少数量的样本可以构造原始数据矩阵的一个低秩近似。在此基础上，本文采用近似 SVD 算法，在降低采样过程时间复杂度的同时，得到与精确 SVD 相近的采样结果，产生有“代表性”的点。本节首先给出基于近似 SVD 的快速采样算法，然后分析通过该算法得到的数据样本点在形成原始数据矩阵低秩近似时的误差，最后给出完整的基于快速地标点采样的谱聚类算法。

3.1 基于近似 SVD 的快速采样算法及误差分析

根据矩阵 SVD 分解结果进行抽样的相关理论分析已由文献[15]给出，而虽然文献[17,19]指出基于近似 SVD 分解结果进行抽样可使矩阵低秩近似的误差保持在小的界内，但并没有给出严格证明，本小节给出一个简洁的证明。

首先给出基于近似 SVD 的抽样算法 3 如表 3 所示。

表 3 近似 SVD 的抽样算法

算法 3 基于近似 SVD 的抽样算法

输入 n 个 d 维数据点形成的数据矩阵 $\boldsymbol{A}\in\mathbb{R}^{n\times d}$，整数 k，抽样个数 p，随机映射压缩后的数据规模 r；

输出 抽样矩阵 $\boldsymbol{S}\in\mathbb{R}^{n\times p}$；

(1) 由算法 2(近似 SVD 算法)得到矩阵 $\boldsymbol{A}$ 前 k 个近似左奇异向量 $\boldsymbol{Z}\in\mathbb{R}^{n\times k}$；

(2) 计算概率：

$$\mathrm{pr}_i=\left\|\boldsymbol{Z}_{(i)}\right\|_2^2\Big/\left\|\boldsymbol{Z}\right\|_F^2,\quad i=1,2,\cdots,n;$$

(3) 令 $\boldsymbol{e}_i(i=1,2,\cdots,n)$ 是 $\mathbb{R}^n$ 的标准基，以下述概率产生抽样矩阵的列：

$$\Pr\left[\boldsymbol{S}^{(j)}=\boldsymbol{e}_i\cdot\frac{1}{\sqrt{p\cdot\mathrm{pr}_i}}\right]=\mathrm{pr}_i,\quad j=1,2,\cdots,p;\quad i=1,2,\cdots,n。$$

通过算法 3 产生抽样矩阵，然后由 $\boldsymbol{S}^{\mathrm{T}}\boldsymbol{A}\in\mathbb{R}^{p\times d}$ 可实现数据抽样。基于该抽样所得到的矩阵低秩近似的误差界由定理 1 给出。

定理 1 给定数据 $\boldsymbol{A}\in\mathbb{R}^{n\times d}$，令 $0<\varepsilon,\delta_1,\delta_2<1$，$k<\mathrm{rank}(\boldsymbol{A})$，$\pi(\cdot)$ 是投影算子。对于算法 3 产生的抽样矩阵 $\boldsymbol{S}$，若压缩后数据规模 $r=\Theta(k/\varepsilon+k\lg k)$，抽样个数 $p>4k\ln(2k/\delta_1)$，则至少以 $1/2-\delta_1-\delta_2$ 的概率，有

$$\left\|\boldsymbol{A}^{\mathrm{T}}-\pi_{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}}\left(\boldsymbol{A}^{\mathrm{T}}\right)\right\|_F^2\le\left\|\boldsymbol{A}^{\mathrm{T}}-\pi_{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S},k}\left(\boldsymbol{A}^{\mathrm{T}}\right)\right\|_F^2\le(1+\varepsilon)\left(1+\frac{1}{\delta_2}\cdot\frac{1}{1-\sqrt{\dfrac{4k\ln(2k/\delta_1)}{p}}}\right)\left\|\boldsymbol{A}^{\mathrm{T}}-\left(\boldsymbol{A}^{\mathrm{T}}\right)_k\right\|_F^2$$

成立。

从定理 1 可知，对于通过算法 3 所得到的矩阵 $\boldsymbol{A}$ 的行样本，其所能得到的最优原始矩阵近似误差与 $\|\boldsymbol{A}-\boldsymbol{A}_k\|_F^2$ 相差一个较小的因子，保持了原始矩阵的大部分信息。在证明定理前，先给出两个相关的引理：

引理 2[20]　令 $0<\delta<1$，$\boldsymbol{Z}\in\mathbb{R}^{n\times k}$ 且 $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}=\boldsymbol{I}_k$，$\boldsymbol{S}$ 是由算法 3 依据 $\boldsymbol{Z}$ 产生的抽样矩阵。若抽样个数 $4k\ln(2k/\delta)<p<n$，则对于 $i=1,2,\cdots,k$，至少以 $1-\delta$ 的概率，有

$$1-\sqrt{\frac{4k\ln(2k/\delta)}{p}}\le\sigma_i^2\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)\le 1+\sqrt{\frac{4k\ln(2k/\delta)}{p}}$$

成立，其中 $\sigma_i(\cdot)$ 是按降序排列第 i 个奇异值。

引理 2 指出，如果抽样规模足够，算法 3 通过列正交矩阵产生的抽样矩阵，作用于原矩阵后仍得到奇异值接近于 1 的矩阵，即将一个列正交矩阵抽样为一个近似的列正交矩阵。

引理 3[11]　令 $0<\delta<1$，任意 $\boldsymbol{Z}\in\mathbb{R}^{n\times k}$ 是算法 3 步骤 1 所需的矩阵，对于算法 3 所产生的抽样矩阵 $\boldsymbol{S}$，对于任意 $p>0$，任意 $\boldsymbol{B}\in\mathbb{R}^{d\times n}$，至少以 $1-\delta$ 的概率，有

$$\|\boldsymbol{BS}\|_F^2\le\frac{1}{\delta}\|\boldsymbol{B}\|_F^2$$

成立。

引理 3 说明根据算法 3 的抽样方法，对任意矩阵抽样并适当调整样本尺寸后，所产生的矩阵与原始矩阵在 Frobenius 范数平方上接近，即该抽样算法对矩阵的 Frobenius 范数没有产生太大的影响。

利用上述引理，给出定理 1 的证明：

证明　由 $\pi_{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S},k}(\boldsymbol{A}^{\mathrm{T}})$ 的定义可知，对任意矩阵 $\boldsymbol{X}\in\mathbb{R}^{p\times n}$ 且 $\mathrm{rank}(\boldsymbol{X})\le k$，有

$$\left\|\boldsymbol{A}^{\mathrm{T}}-\pi_{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S},k}\left(\boldsymbol{A}^{\mathrm{T}}\right)\right\|_F^2\le\left\|\boldsymbol{A}^{\mathrm{T}}-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{SX}\right\|_F^2 \tag{1}$$

成立。因为 $\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)^{\dagger}\boldsymbol{Z}^{\mathrm{T}}\in\mathbb{R}^{p\times n}$ 的秩小于等于 k，所以 $\left\|\boldsymbol{A}^{\mathrm{T}}-\pi_{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S},k}\left(\boldsymbol{A}^{\mathrm{T}}\right)\right\|_F^2\le\left\|\boldsymbol{A}^{\mathrm{T}}-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)^{\dagger}\boldsymbol{Z}^{\mathrm{T}}\right\|_F^2$。

利用矩阵 $\boldsymbol{A}$ 的近似奇异向量矩阵 $\boldsymbol{Z}$，将其分解为：$\boldsymbol{A}^{\mathrm{T}}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{ZZ}^{\mathrm{T}}+\boldsymbol{E}$，则有

$$\begin{aligned}\left\|\boldsymbol{A}^{\mathrm{T}}-\pi_{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S},k}\left(\boldsymbol{A}^{\mathrm{T}}\right)\right\|_F^2&\le\left\|\boldsymbol{A}^{\mathrm{T}}-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)^{\dagger}\boldsymbol{Z}^{\mathrm{T}}\right\|_F^2\\&=\left\|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{ZZ}^{\mathrm{T}}+\boldsymbol{E}-\left(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{ZZ}^{\mathrm{T}}+\boldsymbol{E}\right)\boldsymbol{S}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)^{\dagger}\boldsymbol{Z}^{\mathrm{T}}\right\|_F^2\\&=\left\|\boldsymbol{A}^{\mathrm{T}}\boldsymbol{ZZ}^{\mathrm{T}}-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{ZZ}^{\mathrm{T}}\boldsymbol{S}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)^{\dagger}\boldsymbol{Z}^{\mathrm{T}}\right.\\&\quad\left.+\boldsymbol{E}-\boldsymbol{ES}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)^{\dagger}\boldsymbol{Z}^{\mathrm{T}}\right\|_F^2\end{aligned} \tag{2}$$

成立。根据引理 2 可知，至少以 $1-\delta_1$ 的概率成立 $\mathrm{rank}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)=k$。而 $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)^{\dagger}\in\mathbb{R}^{k\times k}$，因此至少以 $1-\delta_1$ 的概率有 $\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)^{\dagger}=\boldsymbol{I}_k$。于是式(2)可化为

$$\left\|\boldsymbol{E}-\boldsymbol{ES}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)^{\dagger}\boldsymbol{Z}^{\mathrm{T}}\right\|_F^2 \tag{3}$$

因为 $\boldsymbol{EZ}=\left(\boldsymbol{A}^{\mathrm{T}}-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{ZZ}^{\mathrm{T}}\right)\boldsymbol{Z}=\boldsymbol{0}$，由矩阵形式的毕达哥拉斯定理可知上式能化为

$$\begin{aligned}&\left\|\boldsymbol{E}-\boldsymbol{ES}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)^{\dagger}\boldsymbol{Z}^{\mathrm{T}}\right\|_F^2\\&\quad=\|\boldsymbol{E}\|_F^2+\left\|\boldsymbol{ES}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)^{\dagger}\boldsymbol{Z}^{\mathrm{T}}\right\|_F^2\end{aligned} \tag{4}$$

而由性质 1 可知

$$\begin{aligned}\left\|\boldsymbol{ES}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)^{\dagger}\boldsymbol{Z}^{\mathrm{T}}\right\|_F^2&\le\|\boldsymbol{ES}\|_F^2\left\|\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)^{\dagger}\boldsymbol{Z}^{\mathrm{T}}\right\|_2^2\\&=\|\boldsymbol{ES}\|_F^2\left\|\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)^{\dagger}\right\|_2^2\end{aligned} \tag{5}$$

其中，“=”由 $\boldsymbol{Z}$ 的列正交性保证。因为 $\sigma_i\left(\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)^{\dagger}\right)=\sigma_{k+1-i}\left(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{S}\right)^{-1}$，所以根据引理 2 和引理 3，可知至少以 $1-\delta_1-\delta_2$ 的概率，有

$$\begin{aligned}\left\|\boldsymbol{A}^{\mathrm{T}}-\pi_{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S}}\left(\boldsymbol{A}^{\mathrm{T}}\right)\right\|_F^2&\le\left\|\boldsymbol{A}^{\mathrm{T}}-\pi_{\boldsymbol{A}^{\mathrm{T}}\boldsymbol{S},k}\left(\boldsymbol{A}^{\mathrm{T}}\right)\right\|_F^2\\&\le\left(1+\frac{1}{\delta_2}\cdot\frac{1}{1-\sqrt{\dfrac{4k\ln(2k/\delta_1)}{p}}}\right)\|\boldsymbol{E}\|_F^2\end{aligned} \tag{6}$$

成立。最后，由引理 1 中 $\|\boldsymbol{E}\|_F=\left\|\boldsymbol{A}^{\mathrm{T}}-\boldsymbol{A}^{\mathrm{T}}\boldsymbol{ZZ}^{\mathrm{T}}\right\|_F\le(1+\varepsilon)\left\|\boldsymbol{A}^{\mathrm{T}}-\left(\boldsymbol{A}^{\mathrm{T}}\right)_k\right\|_F$ 即可证明结论。　　证毕

定理 1 表明，采用基于近似 SVD 的抽样可以保证抽样误差在特定的界内，这使得采样的样本具有较好的代表性。因此，利用该方法所得到的采样样本，其与数据点形成的相似度矩阵能较好地描述数据之间的关系。

3.2 基于快速地标点采样的谱聚类算法

3.1 节从矩阵低秩近似误差的角度在理论上分析了基于近似 SVD 的抽样样本的代表性，根据 3.1 节结论，我们提出了使用基于近似 SVD 的抽样方法来采样地标点的 LSC 算法，称为基于快速地标点采样的谱聚类算法 (Landmark-based Spectral Clustering with Fast Sampling, LSC-FS)，该算法的具体流程如表 4 所示。

LSC-FS 算法主要分为地标点采样和基于地标点的谱聚类两部分，因为基于地标点的谱聚类算法复杂度已经在 2.2 节给出，所以我们主要对地标点采样部分进行算法复杂度分析。

表 4 基于快速地标点采样的谱聚类算法(LSC-FS)

算法 4 基于快速地标点采样的谱聚类算法 (LSC-FS)

输入 n 个 d 维数据点形成的数据矩阵 $\boldsymbol{A}\in\mathbb{R}^{n\times d}$，簇个数 k，抽样个数 p；

输出 k 个簇；

(1) 令经过随机映射压缩后的数据规模 $r=\Theta(k/\varepsilon+k\lg k)$，抽样个数 $p>4k\ln(2k/\delta_1)$，利用算法 3 进行地标点采样；

(2) 根据得到的地标点，利用算法 1 进行基于地标点的谱聚类，将数据点分为 k 个簇。

对于抽样过程，第 1 步是计算矩阵的近似奇异向量。根据算法 2 的计算流程，计算近似奇异向量的时间为 $O(ndr)+O(dr^2)+O(ndr)+O(nr^2)$，其中算法 2 步骤 2 矩阵乘积需 $O(ndr)$ 的时间，步骤 3 列标准正交化需 $O(dr^2)$，步骤 4 矩阵乘积和 SVD 需 $O(ndr)+O(nr^2)$。抽样算法剩余步骤为确定抽样概率并进行抽样，计算复杂度为 $O(nk)$。由于在实际中常出现 $r\ll\min\{n,d\}$ 且 $k<r$，所以本文所设计的新算法在采样阶段的计算复杂度为 $O(ndr)$。

地标点的生成方法常见的是随机采样，而另外一种地标点的生成方法是用 k 均值的中心点代替。如果用 k 均值的中心点作为地标点，其生成过程计算复杂度为 $O(ndkl)$，其中 l 为迭代次数。由定理 1 的要求可知，$r=\Theta(k/\varepsilon+k\lg k)$，所以新算法的采样过程计算量通常要比基于 k 均值的采样过程要小(当 $l>(1/\varepsilon+\lg k)$ 时)。并且当数据规模极大，超出系统的内存时，k 均值聚类算法需要不断地执行数据读取操作，而新算法的抽样过程对数据的读取次数至多需要 3 次，更高效。

4 实验结果与分析

本节进行实验分析，对算法的有效性和运行时间两类指标进行评估。

我们对两个较大数据集进行实验。第 1 个被称为 MNIST，是一个手写数字的数据集[1)]。该数据集共有 70000 个对象，每个对象是 28×28 像素的属于数字 0 到 9 的图像，其中每个像素是从 $[0,255]$ 中取出的整数。实验时，我们将每个对象视为 784 维的向量。第 2 个被称为 RCV1，是路透社的新闻文档集。为方便实验对比，我们采用与文献[21]一致的处理方式，对其中的 103 类共计 193844 个有 47236 个特征的文档进行聚类分析。实验算法在英特尔 Core i7-4790 @ 3.60 GHz CPU, 16 GB 内存的计算机上运行，实验代码在 MATLAB 环境下编写。

为验证新算法的聚类有效性和效率，本文对原始谱聚类(记为 SC), Nyström 近似谱聚类(记为 Nyström)、基于随机采样的地标点谱聚类(记为 LSC-R)和基于 k 均值中心的地标点谱聚类(记为 LSC-K)与本文算法(记为 LSC-FS)进行实验比较。对于 Nyström 算法，我们采用文献[21]给出的带正交化的 MATLAB 代码，对于 LSC 算法，我们采用文献[9]给出的实现代码。

在实验过程中，通过改变抽样个数来比较不同个数的采样点对实验结果的影响。为避免算法中随机化过程对实验结果的影响，对每一个采样点数，各个算法都独立进行 20 次并取平均值作为算法结果；为比较的公平性，所有相似度矩阵构造过程中的近邻个数都选为 5。

4.1 评价指标

算法有效性描述的是聚类算法对数据进行划分的正确程度，通过对算法聚类结果 $\{\text{cluster}_i\mid i\in[1,c]\}$ 和预定义的类标签 $\{\text{label}_j\mid j\in[1,l]\}$ 进行相似性比对得出。本文用聚类精确性(Cluster Accuracy, CA)[22] 和标准化互信息 (Normalized Mutual Information, NMI)[23]两种指标。

CA 度量了聚类结果中被正确划分到预定义类标签的数据点的比例，按式(7)计算：

$$\text{CA}=\sum_{i=1}^{c}\frac{\max(\text{cluster}_i\mid\text{label})}{n}\tag{7}$$

其中，c 是聚类结果中簇的个数，n 是数据量，cluster_i 是第 i 个簇，$\max(\text{cluster}_i\mid\text{label})$ 表示聚类结果中第 i 个簇 cluster_i 中标签 $\text{label}_j(j\in[1,l])$ 所对应的样本点个数的最大值。从 CA 定义可知 CA 越大，聚类效果越好，CA 最大值为 1。

NMI 也评估了聚类算法的划分质量。将各个簇所占数据总量的比率视为随机变量取值该簇标签的概率，那么可得到两个概率分布，NMI 度量的是两个概率分布之间共享的信息量。将两个随机变量分别记为 C 和 L，按式(8)来计算 NMI：

$$\text{NMI}=\frac{MI(C,L)}{\sqrt{H(C)\cdot H(L)}}=\frac{\sum\limits_{c,l}n_{c,l}\lg\left(\dfrac{n\cdot n_{c,l}}{n_c\cdot n_l}\right)}{\sqrt{\left(\sum\limits_{c}n_c\lg\left(\dfrac{n_c}{n}\right)\right)\left(\sum\limits_{l}n_l\lg\left(\dfrac{n_l}{n}\right)\right)}}\tag{8}$$

其中，$MI(C,L)$ 表示随机变量 C 和 L 的互信息，$H(\cdot)$ 表示随机变量的熵，n 是点的总个数，$n_{c,l}$ 表示既在簇 c 中又在类 l 中的点的个数，n_c 表示簇 c 中点的个数，n_l 表示类 l 中点的个数。NMI 取值范围也是 $[0,1]$，值越大，聚类效果越好；当取 1 时，表示聚类结果完全正确。

[1)] MNIST 数据可从 http://yann.lecun.com/exdb/mnist/上下载

4.2 实验结果

4 种快速谱聚类算法加上原始谱聚类算法在两个数据集上的性能表现如表 5 所示，从左往右依次从运行时间(s), CA(%), NMI(%)3 个方面进行对比。为便于比较，4 种快速算法的抽样个数均设为 1000。需要指出的是，由于原始谱聚类算法在 RCV1 上运行时间太久，所以只运行了两次，不求方差。

从表 5 可以看出，在聚类有效性方面，本文算法的聚类精度要比 LSC-R 算法和 Nyström 近似算法要高，比 LSC-K 算法低；在算法效率方面，新算法比 LSC-K 算法运行时间明显少，并且随着数据集及数据维数的增大，运行时间并没有比 LSC-R 算法差别很大。对算法有效性和效率综合考虑，虽然 LSC-K 算法在聚类效果上来讲表现很好，但随着数据集及其维数的增大，该算法将会越来越慢；从表中的方差项中可以看出新算法通常比 LSC-R 算法要更为稳定。因此，从聚类效果、算法效率及稳定性方面均衡考虑，本文算法有优势。从算法流程可以看出，算法还可以较好地实现并行化处理，这使得新算法更有吸引力。

为了研究抽样个数对各个快速谱聚类方法的影响，我们在 MNIST 数据集上固定其他参数，令抽样个数从 100 到 1100 每隔 100 进行变化，实验结果如图 1–图 3 所示。

从图 1，图 2 中可以看出，不同于 Nyström 近似算法有效性指标变化不大的情况，本文算法的聚类效果随着抽样个数的增多而变好。这说明地标点个数也是新算法的重要参数之一，地标点个数越多，本文算法能获得更多的数据点间的关系信息，聚类效果越好。再结合图 3 可知，在保持运行时间相差不大的情况下，本文算法比 LSC-R 算法的聚类效果要好；虽然聚类效果没有 LSC-K 算法好，但新算法的运行时间要短得多。因此新算法在效率和聚类效果上取得了较好的平衡。

表 5 不同聚类算法的性能对比

数据集	MNIST(s)		RCV1(s)		MNIST (CA, %)		RCV1 (CA, %)		MNIST (NMI, %)		RCV1 (NMI, %)	
	均值	方差	均值	方差	均值	方差	均值	方差	均值	方差	均值	方差
SC	229.4263	0.2813	4170.1800	–	72.0364	1.2661	18.8831	–	76.8007	0.0297	28.6100	–
Nyström	12.5323	0.0296	79.8958	93.0927	54.1959	11.6039	18.1698	0.7788	46.8375	2.1256	25.1684	0.1514
LSC-R	8.4028	1.3248	220.2876	665.0509	64.0393	22.3497	14.9734	0.7705	62.4723	1.9947	22.0985	0.2806
LSC-K	17.3978	0.9337	1387.3599	91108.6600	72.5109	10.6272	16.9626	0.9527	74.6424	3.5439	26.0988	0.7673
LSC-FS	8.2314	0.4918	230.8977	730.1973	66.6129	5.5174	15.1609	0.4707	64.2162	1.0114	22.4991	0.0829

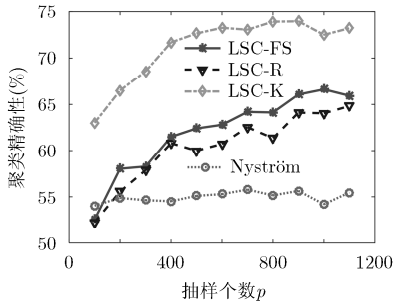

图 1 聚类精确性与抽样个数的关系图

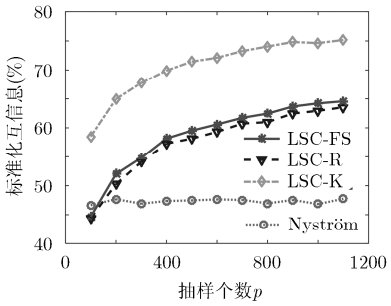

图 2 标准化互信息与抽样个数的关系图

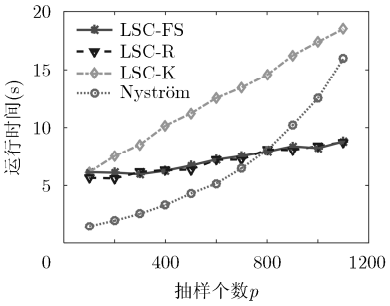

图 3 运行时间与抽样个数的关系图

5 结束语

基于地标表示的谱聚类算法可通过地标点快速实现相似度矩阵的构造和相应拉普拉斯矩阵的分解，是一种适用于大数据集的谱聚类算法。针对随机抽样地标点效果不稳定，用 k 均值中心作为地标点运行时间长的问题，本文设计了一种快速地标点采样算法。本文算法基于近似奇异值分解，可使每个地标点的抽样概率对应于其在数据集中的权重。本文不仅从理论上分析了该抽样算法结果对原始信息的保持性，还从公开数据集上验证了新算法在效率和有效性上的优势。

参 考 文 献

[1] 何清, 李宁, 罗文娟, 等. 大数据下的机器学习算法综述[J]. 模式识别与人工智能, 2014, 27(4): 327–336.
HE Qing, LI Ning, LUO Wenjuan, *et al*. A survey of machine learning algorithms for big data[J]. *Pattern Recognition and*

Artificial Intelligence, 2014, 27(4): 327–336.

[2] DING S, JIA H, ZHANG L, *et al*. Research of semi-supervised spectral clustering algorithm based on pairwise constraints[J]. *Neural Computing and Applications*, 2014, 24(1): 211–219. doi: 10.1007/s00521-012-1207-8.

[3] NG A Y, JORDAN M I, and WEISS Y. On spectral clustering: Analysis and an algorithm[C]. Neural Information Processing Systems: Natural and Synthetic, Vancouver, Canada, 2001: 849–856.

[4] FOWLKES C, BELONGIE S, CHUNG F, *et al*. Spectral grouping using the Nystrom method[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, 26(2): 214–225. doi: 10.1109/TPAMI.2004.1262185.

[5] LI M, KWOK J T, and LU B L. Making large-scale Nyström approximation possible[C]. Proceedings of the 27th International Conference on Machine Learning, Haifa, Israel, 2010: 631–638.

[6] LI M, BI W, KWORK J T, *et al*. Large-scale Nyström kernel matrix approximation using randomized SVD[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, 26(1): 152–164. doi: 10.1109/TNNLS.2014.2359798.

[7] 丁世飞, 贾洪杰, 史忠植. 基于自适应 Nyström 采样的大数据谱聚类算法[J]. 软件学报, 2014, 25(9): 2037–2049. doi: 10.13328/j.cnki.jos.004643.
DING Shifei, JIA Hongjie, and SHI Zhongzhi. Spectral clustering algorithm based on adaptive Nyström sampling for big data analysis[J]. *Journal of Software*, 2014, 25(9): 2037–2049. doi: 10.13328/j.cnki.jos.004643.

[8] YAN D, HUANG L, and JORDAN M I. Fast approximate spectral clustering[C]. Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 2009: 907–916. doi: 10.1145/1557019.1557118.

[9] CHEN X and CAI D. Large scale spectral clustering with landmark-based representation[C]. Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, San Francisco, California, USA, 2011: 313–318.

[10] CAI D and CHEN X. Large scale spectral clustering via landmark-based sparse representation[J]. *IEEE Transactions on Cybernetics*, 2015, 45(8): 1669–1680. doi: 10.1109/TCYB.2014.2358564.

[11] BOUTSIDIS C, ZOUZIAS A, MAHONEY M W, *et al*. Randomized dimensionality reduction for-means clustering[J]. *IEEE Transactions on Information Theory*, 2015, 61(2): 1045–1062. doi: 10.1109/TIT.2014.2375327.

[12] COHEN M, ELDER S, MUSCO C, *et al*. Dimensionality reduction for k-means clustering and low rank approximation[C]. Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, Portland, OR, USA, 2015: 163–172. doi: 10.1145/2746539.2746569.

[13] KHOA N L D and CHAWLA S. A scalable approach to spectral clustering with SDD solvers[J]. *Journal of Intelligent Information Systems*, 2015, 44(2): 289–308. doi: 10.1007/s10844-013-0285-0.

[14] FRIEZE A, KANNAN R, and VEMPALA S. Fast Monte-Carlo algorithms for finding low-rank approximations[C]. Proceedings of the 39th Annual Symposium on Foundations of Computer Science, Palo Alto, California, USA, 1998: 370–378. doi: 10.1109/SFCS.1998.743487.

[15] DRINEAS P, MAHONEY M W, and MUTHUKRISHNAN S. Sampling algorithms for l2 regression and applications[C]. Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm, Miami, Florida, USA, 2006: 1127–1136.

[16] DRINES P, MAHONEY M W, and MUTHUKRISHNAN S. Subspace sampling and relative-error matrix approximation: Column-based methods[C]. 9th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems and 10th International Workshop on Randomization and Computation, Barcelona, Spain, 2006: 316–326. doi: 10.1007/11830924_30.

[17] BOUTSIDIS C, DRINEAS P, and MAGDON-ISMAIL M. Near-optimal column-based matrix reconstruction [J]. *SIAM Journal on Computing*, 2014, 43(2): 687–717. doi: 10.1137/12086755X.

[18] HALKO N, MARTINSSON P G, and TROPP J A. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions[J]. *SIAM Review*, 2011, 53(2): 217–288. doi: 10.1137/090771806.

[19] SARLOIS T. Improved approximation algorithms for large matrices via random projections[C]. Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science, Berkeley, California, USA, 2006: 143–152. doi: 10.1109/FOCS.2006.37.

[20] MAGDON-ISMAIL M. Row sampling for matrix algorithms via a non-commutative Bernstein bound[OL]. http://arxiv.org/ abs/1008.0587, 2015.10.

[21] CHEN W Y, SONG Y, BAI H, *et al*. Parallel spectral clustering in distributed systems[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(3): 568–586. doi: 10.1109/TPAMI.2010.88.

[22] AFAHAD A, ALSHATRI N, TARI Z, *et al*. A survey of clustering algorithms for big data: Taxonomy and empirical analysis[J]. *IEEE Transactions on Emerging Topics in Computing*, 2014, 2(3): 267–279. doi: 10.1109/TETC.2014.2330519.

[23] STREHL A and GHOSH J. Cluster ensembles — A knowledge reuse framework for combining multiple partitions[J]. *The Journal of Machine Learning Research*, 2003, 3: 583–617. doi: 10.1162/153244303321897735.

叶 茂： 男，1988 年生，博士生，研究方向为数据挖掘.

刘文芬： 女，1965 年生，教授，博士生导师，研究方向包括概率统计、网络通信、信息安全.