一种基于加速迭代的大数据集谱聚类方法

陈丽敏[1,2] 杨 静[1] 张健沛[1]
(哈尔滨工程大学计算机科学与技术学院 哈尔滨 150001)[1]
(牡丹江师范学院计算机科学与技术系 牡丹江 157011)[2]

摘 要 传统谱聚类算法的诸多优点只适合小数据集。根据Laplacian矩阵的特点重新构造新的Gram矩阵,输入新构造矩阵的若干列,然后利用加速迭代法解决大数据集的谱聚类特征提取问题,使得在大数据集条件下,谱聚类算法只需要很小的空间复杂度就可达到非常快的计算速度。

关键词 聚类,谱聚类,大规模数据集,加速迭代法,Laplacian矩阵
中图法分类号 TP311 **文献标识码** A

Spectral Clustering Algorithm for Large Scale Data Set Based on Accelerating Iterative Method

CHEN Li-min[1,2] YANG Jing[1] ZHANG Jian-pei[1]
(Institute of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)[1]
(Institute of Computer Science and Technology, Mudanjiang Teachers College, Mudanjiang 157011, China)[2]

Abstract The advantage of the traditional spectral clustering algorithm is applicable in the small scale data set. A new method was proposed in the light of the laplacian matrix characteristics. First, a new Gram matrix was reconstructed and some lies of the new matrix were needed, then the eigen-decomposition based on accelerating iterative method was solved. The calculation speed of the proposed method is very fast and the space complexity is small for large scale data set.

Keywords Clustering, Spectral clustering, Large-scale data set, Accelerating iterative method, Laplacian matrix

1 引言

聚类分析是机器学习、数据挖掘领域中的重要理论,是发现和揭示数据内在结构的主流技术。处理庞大数据量时,能在聚类分析研究上取得好的成果,对于数据分析具有十分重要的意义。传统的聚类算法,如k-means算法、EM算法等都建立在凸球型的样本空间上,当样本空间不为凸时,多数情况下算法会陷入局部最优。

近年来提出的谱聚类算法是一种颇具竞争力的聚类算法[1,13]。谱聚类是根据顶点之间的权值划分相应的无向图,即对相应权值图的矩阵进行特征向量提取,得到新的数据特征,通过矩阵谱分析理论导出聚类对象的新特征,并利用新数据特征对原数据进行聚类[2,3]。谱聚类算法对数据的初始条件不做过多要求,它可以给出全局最优解,还可以使用经典的代数来解决聚类问题[4]。但谱聚类算法的诸多优点目前只是限于小数据集。由于谱聚类算法涉及求解特征向量问题,计算复杂度和存储量都很大,无法适用于大规模数据学习,因此,研究高效、快速、适合大规模数据集的谱聚类算法势在必行。

特征提取是谱聚类的关键问题之一。传统的特征分解方法需要非常大的内存,且计算量非常大。Weng等[5]提出了一种增量的协方差无关的方法,它不像传统方法一样特征分解协方差矩阵,而是循环地输入样本向量。算法的空间复杂度只有$O(n)$,计算复杂度也只有$O(pkn)$。其中n为全部样本数,p为迭代次数,k为聚类数。算法的收敛性也已经得到了证明[7]。依据文献[5]的思想,史卫亚等[6]利用Gram矩阵特点,提出了一种解决大规模数据集问题的核主成分分析算法,它通过迭代求出核主成分。这种迭代方法虽然空间复杂度非常小,但对于大数据集而言,要满足一定的精度要求,迭代次数p就会非常大,甚至远远超过样本数n,导致速度根本满足不了需要,收敛速度较慢,因此该方法有待提高。

本文根据正则Laplacian矩阵特点,重新构造了新的Gram矩阵,并使得新构造矩阵的特征值由大到小与原Laplacian矩阵的特征值由小到大所对应的特征向量相同。将一次选取的新构造矩阵的若干列,看成是迭代法[5]的输入样本,利用Aitken向量加速迭代法,提出一种快速解决大规模数据集的谱聚类特征向量提取问题。本算法使迭代计算速度显著加快,所需存储量也相对很小。

2 迭代算法分析

Weng等[5]利用统计学上的效能估计概念提出了一种增量的协方差无关的方法CCIPCA,算法要求所有样本序列u

收稿日期:2011-06-02 返修日期:2011-09-20 本文受国家自然科学基金(60873037,61073041,61073043)及牡丹江市科技攻关项目(G2009b328)资助。

陈丽敏 女,博士生,副教授,主要研究方向为数据挖掘、机器学习,E-mail:chenlimin_clm@126.com;杨 静 女,教授,博士生导师,主要研究方向为数据库与知识库、软件工程。

(n)具有零均值，令 $A=E\{u(n)u^{\mathrm{T}}(n)\}$。该方法不是直接特征分解协方差矩阵$(u_1,\cdots,u_n)(u_1,\cdots,u_n)^{\mathrm{T}}$，而是循环地输入样本序列 $u(n)$。

设协方差矩阵的特征值为 λ，特征向量为 x，令 $v=\lambda x$，则 $x=v/\|v\|$。第 i 步令 $x(i)=v(i-1)/\|v(i-1)\|$，则得到文献[5]主成分迭代公式：

$$v_i(n)=\frac{n-1}{n}v_i(n-1)+\frac{1}{n}u_i(n)u_i^{\mathrm{T}}(n)\frac{v_i(n-1)}{\|v_i(n-1)\|} \quad (1)$$

$$u_{i+1}(n)=u_i(n)-u_i^{\mathrm{T}}(n)\frac{v_i(n)}{\|v_i(n)\|}\frac{v_i(n)}{\|v_i(n)\|} \quad (2)$$

令 $A_n=u(n)u^{\mathrm{T}}(n)$，则 $A=E\{A(n)\}$。令 $\prod_i(n)=\prod_{j=1}^{i-1}[I-\frac{v_j(n)v_j^{\mathrm{T}}(n)}{\|v_j(n)\|^2}]$，则得到文献[7]主成分迭代公式的统一表示：

$$v_i(n)=v_i(n-1)+\frac{1}{n}\left(\frac{\prod_i(n)A(n)\prod_i(n)}{\|v_i(n-1)\|}-I\right)v_i(n-1) \quad (3)$$

文献[7]已证明当 $n\to\infty$ 时，$v_i(n)\to\pm\lambda_i e_i$，其中 e_i 为相应特征值 λ_i 的单位特征向量。实验证明，式(3)迭代一定次数之后收敛很慢，速度远远不够，需要加速提高。分析式(3)可知，若在有限步尽快获得满足需要的近似解，对于误差矩阵 $A(\varepsilon_n)=A(n)-A$，若 $A(n)$ 趋于 A 的速度越快，式(3)就越快趋于准确解。若 $A(n)$ 趋于 A 较慢，则获得精度较高的近似解，迭代次数就会远远大于样本数 n。若 $A(n)=A$，式(3)迭代矩阵 $A(n)$ 与矩阵 A 无误差，收敛速度最快。若取 $A(n)$ 为 A，则 $A(n)=A$，迭代法收敛速度快，且可以使用加速方法加快迭代过程，但存储量过大。

3 Aitken 加速方法介绍

只要迭代算法的迭代过程收敛，迭代次数足够多，就可以使结果达到任意的精度。但是，若迭代过程收敛过于缓慢，计算量会变得非常大。因此，加速迭代过程是一个必要的选择。Aitken 加速方法是数值计算学科领域的经典加速方法[12]。对于收敛序列$\{x_k\}$，由微分中值定理可推出：

$$\bar{x}_{k+1}=x_k-\frac{(x_{k+1}-x_k)^2}{x_{k+2}-2x_{k+1}+x_k}=x_k-(\Delta x_k)^2/\Delta^2 x_k$$
$$(k=0,1,\cdots) \quad (4)$$

式(4)称为 Aitken Δ^2 加速方法。可以证明

$$\lim_{k\to\infty}\frac{\bar{x}_{k+1}-x^*}{x_k-x^*}=0$$

它表明序列$\{\bar{x}_k\}$的收敛速度比$\{x_k\}$的收敛速度快。将这种加速方法用于具体的迭代法上，可对原迭代法进行有效的加速，有时甚至能将发散的具体迭代格式通过这种加速后变成收敛。

对于向量序列，Aitken[11]又提出了一个加速向量收敛的方法，它也很有效。假定 u_s,u_{s+1},u_{s+2} 是 3 个连续的迭代量，$\bar{u}_i=u_i^{(s)}-\frac{(u_i^{(s)}-u_i^{(s+1)})^2}{u_i^{(s)}-2u_i^{(s+1)}+u_i^{(s+2)}}$，其中 $u_i^{(s)}$ 表示 $u^{(s)}$ 的第 i 个分量。当分母等于 0，该分量可以停止加速。Aitken 过程给出的向量比已导出的向量更加接近准确值[11]。

4 矩阵分析

4.1 基本知识

Laplacian 矩阵有一个完整的研究领域，称为谱图理论[8,9]。将数据集表示成各数据点间相似性的无向加权图 G，并用加权矩阵 W 表示，其中权值 $w_{ij}=w_{ji}\geqslant 0$。矩阵 W 的每行元素相加，得到该顶点的度。以所有度值为对角元素构成的对角矩阵即为度矩阵，用 D 表示。有如下定义。

正则 Laplacian 矩阵：

$L_{sym}=D^{-1/2}LD^{-1/2}=I-D^{-1/2}WD^{-1/2}$

$L_{rw}=D^{-1}L=I-D^{-1}W$

性质 1[2] Laplacian 矩阵 L_{sym} 和 L_{rw} 有 k 重特征值 0，当且仅当图 G 有 k 个连通分支；L_{rw} 的特征值 0 对应特征向量常向量 1，L_{sym} 的特征值 0 对应特征向量 $D^{1/2}1$。

Gershgorin 定理[10]：$n\times n$ 矩阵 A 的所有特征值包含在下面的圆盘的并集中：

$$K_i=\{\mu\in C\,|\,|\mu-a_{ii}|\leqslant\sum_{k=1,k\neq i}^{n}|a_{ik}|\}$$

圆盘 K_i 称为 Gershgorin 圆盘。

4.2 构造 Gram 矩阵

构造新的 Gram 矩阵，令：

$L=2I-L_{sym}=I+D^{-1/2}WD^{-1/2}$，$L_G=L^2/n$

定理 1 若 $L=2I-L_{sym}=I+D^{-1/2}WD^{-1/2}$，则

1)L 与 L_{sym} 有相同的特征向量；

2)L 的特征值为 $\lambda(L)=2-\lambda(L_{sym})$；

3)L 的特征值取值范围为 $\lambda(L)\in[0,2]$，L 为对称半正定矩阵。

证明：设 $\lambda(L_{sym})$、ω 为矩阵 L_{sym} 的特征值和特征向量，有：

$$L_{sym}\omega=\lambda(L_{sym})\omega$$
$$L\omega=(2I-L_{sym})\omega=2I\omega-L_{sym}\omega$$
$$=2\omega-\lambda(L_{sym})\omega=(2-\lambda(L_{sym}))\omega$$

故 1)、2)得证。

由 Gershgorin 圆盘定理知，L_{sym} 的特征值为 $0\leqslant\lambda_1(L_{sym})\leqslant\cdots\leqslant\lambda_n(L_{sym})\leqslant 2$，$L$ 的特征值为 $\lambda(L)=2-\lambda(L_{sym})$，则 $2\geqslant\lambda_1(L)\geqslant\cdots\geqslant\lambda_n(L)\geqslant 0$，故 L 的特征值取值范围为 $\lambda(L)\in[0,2]$，矩阵 L 对称显然。因为矩阵 L 的所有特征值均大于等于 0，故 L 为半正定，即为 Gram 矩阵。

由定理 1 可知，L 的最大特征值对应的特征向量即为 L_{sym} 的最小特征值对应的特征向量，L_{sym} 的前 k 个最小特征值对应的特征向量即为 L 的前 k 个最大特征值对应的特征向量。

定理 2 若 $L_G=L^2/n$，则 L_G 与 L 有相同的特征向量、不同的特征值 $\lambda(L_G)=\lambda(L)^2/n$，且 L_G 对称半正定。

定理 3 Gram 矩阵 L_G 与 L_{sym} 有相同的特征向量、不同的特征值 $\lambda(L_G)=\lambda(L)^2/n=(2-\lambda(L_{sym}))^2/n$。

因此，只要解出矩阵 L_G 的前 k 个最大特征值对应的特征向量，也即矩阵 L 的前 k 个最大特征值对应的特征向量，那么也就求解出了矩阵 L_{sym} 的前 k 个最小特征值对应的特征向量。

由矩阵 L_G 的定义可导出：

矩阵 L_G 的性质 1：

$$L_G=LL^{\mathrm{T}}/n=\frac{1}{n}\begin{pmatrix}l_{11}&\cdots&l_{1n}\\ \vdots&\ddots&\vdots\\ l_{n1}&\cdots&l_{nm}\end{pmatrix}\begin{pmatrix}l_{11}&\cdots&l_{1n}\\ \vdots&\ddots&\vdots\\ l_{n1}&\cdots&l_{nm}\end{pmatrix}^{\mathrm{T}}$$
$$=\frac{1}{n}(L(x_1),\cdots,L(x_n))(L(x_1),\cdots,L(x_n))^{\mathrm{T}}$$

$$=\frac{1}{n}\sum_{i=1}^{n}L(x_i)L(x_i)^{\mathrm{T}}$$

其中，$L(x_i)=(l_{i1},l_{i2},\cdots,l_{in})^{\mathrm{T}}$。

定理 4 设 $L(x_i)=(l_{i1},l_{i2},\cdots,l_{in})^{\mathrm{T}}$，$i=1,\cdots,n$ 是矩阵 L 的列向量，任取 m 个列向量，依次将该 m 个列向量表示成 $L'(x_i)=L(x_j)=(l_{j1},l_{j2},\cdots,l_{jn})^{\mathrm{T}}$，其中 i 在 $1\leqslant i\leqslant m$ 范围顺序取值，j 在 $1\leqslant j\leqslant n$ 范围任意取值，则矩阵 $\frac{1}{m}(L'(x_1)\cdots L'(x_m))(L'(x_1)\cdots L'(x_m))^{\mathrm{T}}$ 对称半正定，其中 $m\geqslant 1$。

证明：对称显然。因为

$$\frac{1}{m}(L'(x_1)\cdots L'(x_m))(L'(x_1)\cdots L'(x_m))^{\mathrm{T}}$$

$$=\frac{1}{m}\sum_{i=1}^{m}L'(x_i)L'(x_i)^{\mathrm{T}}$$

对于任意 $x=(x_1,x_2,\cdots,x_n)^{\mathrm{T}}$，有

$$x^{\mathrm{T}}\frac{1}{m}\sum_{i=1}^{m}L'(x_i)L'(x_i)^{\mathrm{T}}x=\frac{1}{m}\sum_{i=1}^{m}x^{\mathrm{T}}L'(x_i)L'(x_i)^{\mathrm{T}}x$$

$$=\frac{1}{m}\sum_{i=1}^{m}(x^{\mathrm{T}}L'(x_i))^2\geqslant 0$$

故矩阵 $\frac{1}{m}(L'(x_1)\cdots L'(x_m))(L'(x_1)\cdots L'(x_m))^{\mathrm{T}}$ 对称半正定。

假设 m 整除 n 为 k，由定理 4 可以导出矩阵 L_G 的另一性质。任取 m 列，且不重复取任一列，依次将该 m 个列向量顺序表示成 $L'(x_{s\times m+i})=L(x_j)=(l_{j1},l_{j2},\cdots,l_{jn})^{\mathrm{T}}$，其中 i 在 $1\leqslant i\leqslant m$ 范围顺序取值，s 在 $0\leqslant s\leqslant k-1$ 范围顺序取值，j 在 $1\leqslant j\leqslant n$ 范围任意取值，令 $LG^*=\frac{1}{m}(L'(x_1)\cdots L'(x_m))(L'(x_1)\cdots L'(x_m))^{\mathrm{T}}$，则有：

矩阵 L_G 的性质 2：

$$L_G=\frac{1}{k}\sum_{s=0}^{k-1}\frac{1}{m}(L'(x_{s\times m+1})\cdots L'(x_{s\times m+m}))(L'(x_{s\times m+1})\cdots L'(x_{s\times m+m}))^{\mathrm{T}}$$

分析 L_G 的性质 2 可知，当 $m=n$，有 $L_G=LG^*$。当 $m\neq n$，若每个式子 $\frac{1}{m}(L'(x_{s\times m+1})\cdots L'(x_{s\times m+m}))(L'(x_{s\times m+1})\cdots L'(x_{s\times m+m}))^{\mathrm{T}}$ 都相等，其中 $0\leqslant s\leqslant k-1$，则：

$$L_G=\frac{1}{m}(L'(x_{s\times m+1})\cdots L'(x_{s\times m+m}))(L'(x_{s\times m+1})\cdots L'(x_{s\times m+m}))^{\mathrm{T}}$$

当 $LG^*=\frac{1}{m}(L'(x_1)\cdots L'(x_m))(L'(x_1)\cdots L'(x_m))^{\mathrm{T}}$，$m\geqslant 1$，$m<<n$，其中 n 为样本数，m 为选取的样本数，要求所选取的 m 个样本点能够按比例均匀地分布于所要划分的 k 个类中，即选取的 m 个列向量 $L(x_i)$ 能够按比例均匀分布于矩阵 L 中，则 L_G 与 LG^* 误差非常小，$L_G\approx LG^*$。

5 本文方法

5.1 NJW 算法特征向量分析

NJW 算法是一种流行的谱聚类算法，是由 Ng 等人[4]提出的一个简单而有效的多类聚类方法。该算法需要计算矩阵 $\tilde{L}=I-L_{sym}=D^{-1/2}WD^{-1/2}$ 的前 k 个最大特征值对应的特征向量，也即计算矩阵 L_{sym} 的前 k 个最小特征值对应的特征向量。由定理 3 知，矩阵 L_{sym} 与 L_G 有相同的特征向量，且特征值满足 $\lambda(L_G)=(2-\lambda(L_{sym}))^2/n$，故计算 L_{sym} 的前 k 个最小特征值对应的特征向量，也即是求矩阵 L_G 的前 k 个最大特征值对应的特征向量。

5.2 本文算法推导

本文采用全连通图法构造相似矩阵 L_{sym}。根据正则 Laplacian 矩阵的性质，图 G 只有 1 个连通分支，因此矩阵 L_{sym} 的特征值 0 只有 1 重。L_{sym} 的最小特征值 0 对应的特征向量为 $D^{1/2}1$，因此矩阵 L 和 L_G 的最大特征值所对应的特征向量也为 $D^{1/2}1$。

设矩阵 L_G 的特征向量和特征值分别为 ω、$\lambda(L_G)$，令 $\psi=\lambda(L_G)\omega=L_G\omega$，则特征向量和特征值分别为 $\omega=\psi/\|\psi\|$，$\lambda(L_G)=\|\psi\|$。设 $\psi_i(p)$ 是第 i 个特征向量在第 p 次的迭代估计。根据文献[7]及前面的迭代法分析，为加快计算速度，在 p 次迭代时，第 i 阶特征向量的估计 $\psi_i(p)$ 计算可以表示为：

$$\psi_i(p)=\frac{p-1}{p}\psi_i(p-1)+\frac{1}{p}L_G\frac{\psi_i(p-1)}{\|\psi_i(p-1)\|}$$

$$=\frac{p-1}{p}\psi_i(p-1)+\frac{1}{p}(LL^{\mathrm{T}}/n)\frac{\psi_i(p-1)}{\|\psi_i(p-1)\|}$$

$$=\frac{p-1}{p}\psi_i(p-1)+\frac{1}{p}(\frac{1}{n}(L(x_1)\cdots L(x_n))(L(x_1)\cdots L(x_n))^{\mathrm{T}})\frac{\psi_i(p-1)}{\|\psi_i(p-1)\|} \tag{5}$$

但式(5)所需存储量过大，因为 $L_G\approx LG^*$。故只要所选取的 m 个列向量 $L(x_i)$ 能够按比例均匀分布于矩阵 L 中，L_G 与 LG^* 误差就会非常小。用 LG^* 近似代替 L_G，则 $\psi_i(p)$ 可近似表示为：

$$\psi_i(p)=\frac{p-1}{p}\psi_i(p-1)+\frac{1}{p}(\frac{1}{m}(L'(x_1)\cdots L'(x_m))(L'(x_1)\cdots L'(x_m))^{\mathrm{T}})\frac{\psi_i(p-1)}{\|\psi_i(p-1)\|} \tag{6}$$

式中，$m\geqslant 1$，$m<<n$。可对式(5)使用 Aitken 加速，以最少迭代次数得到满足要求的解。

令 $LL_1=(L'(x_1)\cdots L'(x_m))$，其他高阶特征向量组可用残留的数据向量计算。残留的数据向量是用最初数据减去其在低阶特征向量的投影：

$$LL_{i+1}=LL_i-(LL_i)^{\mathrm{T}}\frac{\psi_i(p)}{\|\psi_i(p)\|}\frac{\psi_i(p)}{\|\psi_i(p)\|} \tag{7}$$

本文采用 NJW 算法思想。为避免初始向量对迭代计算特征向量影响过大，初始向量选择常向量 1。算法如下：

Step1 计算 1 阶特征向量 $D^{1/2}1$，使用常向量 1 初始化前 2～k 阶特征向量；

Step2 选择 m，计算 LL_1，$m<<n$；

Step3 循环 2:k 进行下面计算；

Step4 利用式(7)计算 LL_k，输入向量 LL_k；

Step5 迭代 1:p 次进行下面计算；

Step6 计算式(6)；

Step7 利用 Aitken 方法加速式(6)；

Step8 转到 Step5；

Step9 转到 Step3；

Step10 用 k-means 对 k 个特征向量分类；

Step11 输出聚类集合。

若谱聚类的其他方法生成的相似矩阵也能够重构成新的 Gram 矩阵，使之与原相似矩阵有相同的特征向量，且能保证新构成的矩阵的前 k 个最大特征值对应的特征向量即是所需的特征向量，则也可用上述算法解决特征向量提取问题。

5.3 算法分析

输入 L 矩阵的 m 列，迭代过程 Step2－Step9 的空间复杂度为 $O(mn)$，时间复杂度为 $O(mkpn)$，其中 p 为迭代次数，n 为总样本数，m 为选取矩阵 L 的列数，k 为特征向量个数，也即聚类数。因为使用了 Aitken 加速法，迭代次数 p 很小，$m \ll n$，计算速度明显提高。算法可根据存储空间大小、收敛速度等要求，适当选取 m 值。

6 实验

为验证本文算法的有效性，分别对小数据集、中等数据集及大数据集的相似矩阵 L_{sym} 用标准方法计算前 k 个特征向量，然后用本文算法计算前 k 个特征向量，并进行验证、比较。实验中选择高斯函数 $k(x,y)=\exp(-\|x-y\|/2\sigma^2)$ 生成全连通图，构造相似矩阵 L_{sym}。

实验 1　二维人工数据集 $g1$ 样本分别由 3 个正态分布函数生成，3 个聚类样本数分别是 30、20、10，均值分别为[1，4]、4.5、[5，3]，标准方差分别为[0.2，0；0，0.5]、0.1、[0.3，0；0，0.1]，$\sigma=1$。数据集 $g1$ 的样本分布如图 1 所示。

本实验首先比较了当 $m=60$、第 2 个特征向量加速与不加速时特征向量内积收敛速度的情况。如图 3 所示，其中最上方蓝色圆圈是用标准方法得到的第 2 特征向量对应的特征值，最下方绿色圆点是没有使用 Aitken 加速第 2 特征向量内积收敛的情况，中间一条红色圆点是使用了 Aitken 加速第 2 特征向量内积收敛的情况。可以看出，采用 Aitken 加速法，计算速度明显提高。然后比较了当 $m=20$，$m=30$，$m=40$，迭代次数 $p=30$ 时，第 2、3 特征向量的内积，如表 1 所列。其中第 1 特征值可通过给定矩阵直接计算。当 m 分别取 20、30 和 40 时，聚类结果均正确；当 $m=20$ 时误差最大，$m=40$ 时误差最小。虽然 $m=20$ 误差比较大，但仍能保证聚类结果的正确性。

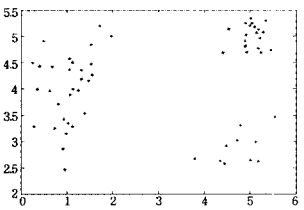

图 1　数据集 g1 数据分布

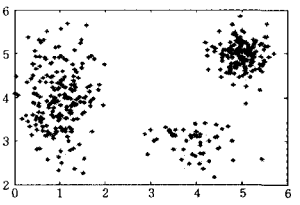

图 2　数据集 g2 数据分布

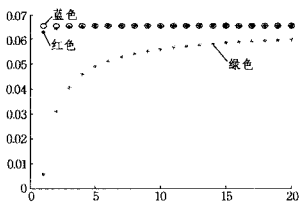

图 3　数据集 g1 第 2 特征向量加速与未加速内积收敛情况

表 1　数据集 g1 特征值比较

	前 3 个特征值		
	1(标准值)	2	3
$\lambda(L_{sym})$(标准)	0	0.0205	0.2462
$\lambda(L)$(标准)	2	1.9795	1.7538
$\lambda(L_G)$(标准)	0.0667	0.0653	0.0513
$\lambda(LG^*)$(m=20)	0.0667	0.0584	0.0412
$\lambda(LG^*)$(m=30)	0.667	0.0614	0.0468
$\lambda(LG^*)$(m=40)	0.667	0.0631	0.0501

实验 2　二维人工数据集 $g2$ 样本分别由 3 个正态分布函数生成，3 个聚类样本数分别是 200、150、50，均值分别为[1，4]、5、[4，3]，标准方差分别为[0.2，0；0，0.55]、0.1、[0.35，0；0，0.1]，$\sigma=1$。数据集 $g2$ 的样本分布如图 2 所示。

本实验分别比较了当 $m=20$，$m=40$，$m=60$，迭代次数 $p=30$ 时，第 2、3 特征向量的内积，如表 2 所列。其中第 1 特征值可通过给定矩阵直接计算。当 m 分别取 40 和 60 时，聚类结果均正确；当 $m=20$ 时误差最大，$m=60$ 时误差最小；$m=20$ 时也基本能够保证聚类结果的正确性。

表 2　数据集 g2 特征值比较

	前 3 个特征值		
	1(标准值)	2	3
$\lambda(L_{sym})$(标准)	0	0.0233	0.2379
$\lambda(L)$(标准)	2	1.9767	1.7621
$\lambda(L_G)$(标准)	0.0100	0.0098	0.0078
$\lambda(LG^*)$(m=20)	0.0100	0.0071	0.0049
$\lambda(LG^*)$(m=40)	0.0100	0.0083	0.0057
$\lambda(LG^*)$(m=60)	0.0100	0.0089	0.0068

实验 3　选取 UCI 数据集 waveform，其样本容量为 5000，属性 40 个。本实验选取其中 300、500 和 1000 个样本，$m=60$，$p=40$。本文方法所得第 2、3 特征向量相对标准求解特征向量的误差率如表 3 所列。实验表明，本文算法聚类结果是有效的。当数据样本取 5000，$m=80$ 时，仍可以计算。

表 3　waveform 误差比较

相对误差	样本数		
	300	500	1000
第 2 特征向量误差	3.6%	4.3%	5.7%
第 3 特征向量误差	4.9%	5.8%	7.4%

结束语　数据集本身的聚类数很少，即 k 值比较小，且聚类分布明显时，即使迭代次数较少，误差相对较大一些，聚类效果也比较好。若矩阵自身数据分布比较均匀，迭代矩阵误差很小，则迭代收敛速度相对较快，精度较高。当迭代次数一定时，m 值越大，精度越高，但存储空间变大。实验表明，m 取 40～60 效果就足够好。当样本数目很大时，传统方法无法提取相似矩阵的特征向量，但本文方法仍然能够提取谱聚类相似矩阵的特征向量。下一步，在大规模数据集情况下，在提高速度的同时，如何提高精度，仍是一个挑战。

参 考 文 献

[1] Von Luxburg U. A tutorial on spectral clustering[R]. TR-149. Max Planck Institute for Biological Cybernetics, 2006

[2] Shi J, Malik J. Normalized cuts and image segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888-905

[3] Ng A, Jordan M, Weiss Y. On spectral clustering: Analysis and an algorithm[C]// Advances in Neural Information Processing Systems. MIT Press, 2002: 849-856

[4] Von Luxburg U. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17(4): 395-416

[5] Weng J, Zhang Y, Huang W S. Candid covariance-free incremental principal component analysis[J]. IEEE Trans. on Pattern Analysis Machine Intelligence, 2003, 25(8): 1034-1040

[6] 史卫亚，郭跃飞，薛向阳. 一种解决大规模数据集问题的核主成分分析算法[J]. 软件学报，2009，20(8)：2153-2159

[7] Zhang Y, Weng J. Convergence analysis of complementary can-

did incremental principal component analysis[R]. MSU-CSE-01-23. East Lansing: Michigan State University, 2001

[8] Chung F. Spectral graph theory(Vol. 92 of the CBMS Regional Conference Series in Mathematics)[C]// Conference Board of the Mathematical Sciences. Washington, 1997

[9] Fiedler M. Algebraic connectivity of graphs[J]. Czechoslovak Mathematical Journal, 1973, 23(98): 298-305

[10] 胡茂林. 矩阵计算与应用[M]. 北京: 科学出版社, 2008: 219-220

[11] Aitken A C. The evaluation of the latent roota and latent vectors of a matrix[C]// Proc. roy. Soc. Eedinb. 1937, 57: 269-304

[12] Wilkinson J H. 代数特征值问题[M]. 石钟慈, 邓健新, 译. 北京: 科学出版社, 2003: 588-597

[13] 蔡晓妍, 戴冠中, 杨黎斌. 谱聚类算法综述[J]. 计算机科学, 2008, 35(7)

(上接第 164 页)

表 2 中, precision(准确率)=正确抽取事件数目/抽取事件数目, recall(召回率)=正确抽取事件数目/实际事件数目, *F*-Measure(*F* 度量)=(2×recall×precision)/(recall+precision)。对表 2 及实验过程进行分析可知:

1)受命事件的抽取效率远高于其他事件,这主要是因为触发此类事件的动词只有"调任"一个,大大降低了后期处理难度。因此,在保证完整性的前提下尽可能缩减触发动词数目,是提高此类方法效率的重要环节。

2)除受命事件外,调遣事件的抽取效率要高于其他事件。这是因为和其他动词相比,调遣动词的 ASP 更加清晰规范。因此,进行全面深入的句法语义分析,寻找更加可靠的匹配规则,仍是提高此类方法效率的重中之重。

3)如果放弃人工干预,进行全开放实验,结果(平均准确率约为 55.2%、平均召回率约为 53.9%、平均 *F* 度量值为 54.5%)和现有结果比起来存在较大差距,这说明今后很长一段时间内,半自动化抽取方法仍将占据主流。

4)此外,实验过程中出现的词法分析错误、短语识别错误和句型匹配错误对实验结果也造成了不同程度的影响。

为评判本文方法的优劣,课题组基于相同语料和任务,选取任职事件测定了另外两种代表性事件抽取方法的效率,这两种方法分别是事件框架的方法和 IEPAM 方法,其方法过程可参看文献[11,12]。之所以选择上述两者作为比较对象,一方面因为它们的效率得到了已有成果的肯定;另一方面因为它们分属模式匹配方法和机器学习方法,而本文方法属于模式匹配方法,因此,用本文方法与前者比较可以判定其在同类方法中的优劣,与后者比较则可评判其在异类方法中的优劣。实验结果如表 3 所列。

表 3 动词论元结构方法与其他方法的对比分析

	动词论元结构方法		事件框架方法		IEPAM 方法	
	任命	担任	任命	担任	任命	担任
precision	77.8%	74.2%	73.2%	70.6%	59.1%	57.3%
Recall	74.1%	71.3%	70.3%	67.7%	56.8%	54.9%
F-Measure	75.9%	72.7%	71.9%	68.4%	57.4%	55.5%

从表 3 可以看出:1)以动词论元结构为核心来完成事件分类和定位要优于抽象的事件框架机制,由此形成的模式匹配方法更具应用潜质。2)与模式匹配方法相比,机器学习方法在效率上仍存在较大差距,虽然不具有灵活的可移植性,但模式匹配方法在一定时期内仍将是主流。

综合表 2 和表 3 的实验结果可知,基于动词论元结构的事件抽取方法已具备基本应用价值,值得深入研究。

结束语 本文所提方法属于动词驱动的信息抽取方法,触发动词在其中一直起着驱动和引导作用,和事件驱动的信息抽取方法不同,此类方法主要依靠动词聚合来实现跨事件融合[13]。

利用动词的 ASP 来确定其句法成分(如主语、宾语等)和语义角色(如施事、受事等)之间的映射规则,是本文所提方法中最关键的技术。此技术虽然只能视为浅层的文本理解技术,但从应用层面看,由于事件抽取只关注有限的基本信息,不深究文本意义和写作意图等深层理解问题,因此,此技术完全能够满足中文事件抽取的应用要求。

为适应计算机系统,本课题组在实现上述技术时尽可能穷举了表征特定事件的动词及其 ASP。在此过程中,本课题组借鉴了大量汉语言本体研究成果,并从中找出了一些具备可计算性的句法语义规则。

当然,仅靠本文提及的动词论元结构知识来实现全面的事件信息抽取显然不切实际,进一步考虑利用基于论元结构的篇章和逻辑知识来确定事件信息类型和特征分布将是本课题组下一步的努力方向。

参 考 文 献

[1] 赵妍妍, 秦兵, 车万翔, 等. 中文事件抽取技术研究[J]. 中文信息学报, 2008, 22(1): 3-8

[2] 谭红叶. 中文事件抽取关键技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2008

[3] 邓擘, 郑彦宁, 樊孝忠. 汉语信息抽取中事件的定位与分类[J]. 情报理论与实践, 2009, 28(10): 104-107

[4] 顾阳. 论元结构介绍[J]. 国外语言学, 1994(1): 1-11

[5] 袁毓林. 用动词的论元结构跟事件模板相匹配——种由动词驱动的信息抽取方法[J]. 中文信息学报, 2002, 19(5): 37-43

[6] 袁毓林. 一套动词的论元角色的语法指标[J]. 世界汉语教学, 2003(3): 24-36

[7] Yangarber R, Grishman R, Tapanainen P, et al. Automatic Acquisition of Domain Knowledge for Information Extraction[C]// Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000). Saarbriicken, Germany, 2000: 412-416

[8] 姜吉发. 自由文本的信息抽取模式获取的研究[D]. 北京: 中国科学院, 2004

[9] Kim J, Moldovan D. Acquisition of Linguistic Patterns for Knowledge-based Information Extraction[J]. IEEE Transactions on Knowledge and Data Engineering, 1995, 7(5): 713-724

[10] 许荣华, 吴刚, 李培峰, 等. 基于指代消解的中文事件融合方法[J]. 计算机应用, 2009, 29(8): 2264-2267

[11] 梁晗, 陈群秀, 吴平博. 基于事件框架的信息抽取系统[J]. 中文信息学报, 2006, 20(2): 40-46

[12] 姜吉发. 一种事件信息抽取模式获取方法[J]. 计算机工程, 2005, 31(15): 27-29

[13] 许荣华, 吴刚, 李培峰, 等. 基于事件框架的主题事件融合研究[J]. 计算机应用研究, 2009, 26(12): 4542-4545