文章编号: 1001-0920 (2012) 04-0501-06

基于模糊 K-调和均值的单词-文档谱聚类方法

刘 娜[1,2], 肖智博[1], 鲁明羽[1]

(1. 大连海事大学 信息科学技术学院，辽宁 大连 116026;
2. 大连工业大学 信息科学与工程学院，辽宁 大连 116034)

摘 要: 在分析单词-文档谱聚类方法的基本步骤, 找出其对初始值敏感的根本原因的基础上, 提出一种基于模糊 K-调和均值的单词-文档谱聚类方法. 首先从矩阵相似的角度对谱聚类中的Laplacian矩阵进行处理, 使其满足对初始值不敏感的条件; 然后通过加入模糊的概念, 用模糊 K-调和均值算法代替 K-均值算法, 使聚类结果对初始值不敏感. 实验结果表明, 所提出的方法不仅使聚类结果对初始值不敏感, 而且在一定程度上提高了数据的鲁棒性.

关键词: 谱聚类; K-均值; K-调和均值; 模糊K-调和均值

中图分类号: TP181 **文献标识码:** A

Spectral co-clustering documents and words based on fuzzy K-harmonic means

LIU Na[1,2], *XIAO Zhi-bo*[1], *LU Ming-yu*[1]

(1. College of Information Science & Technology，Dalian Maritime University，Dalian 116026，China；2. College of Information Science & Engineering，Dalian Polytechnic University，Dalian 116034，China. Correspondent：LU Ming-yu，E-mail：lumingyu@dlmu.edu.cn)

Abstract: Based on analysising the main step of spectral clustering and finding out its cause of sensitive to the initialization, a method of spectral co-clustering documents and words based on fuzzy K-harmonic means is proposed. Firstly, the matrix which is insensitive to the initialization is constructed. Then fuzzy K-harmonic means algorithm is used instead of K-means algorithm. The experiment result shows that the proposed method not only is initialization insensitive, but also can improve the accuracy and robustness of clustering results.

Key words: spectral clustering; K-means; K-harmonic means; fuzzy K-harmonic means

1 引 言

聚类分析是数据挖掘中一个非常活跃的研究领域, 而单词聚类和文档聚类是聚类分析中的热点问题. 近年来, 人们已经提出了许多聚类算法, 其中谱聚类方法具有不受簇空间形状限制、不会陷入局部最优解[1]及其有效性可从多个方面进行解释等特点[2-3], 越来越受到人们的关注.

谱聚类算法最初用于计算机视觉[4]和VLSI设计[5]等领域, 最近已开始用于机器学习[3], 并迅速成为国际上机器学习领域的研究热点. 典型的算法有: 2-way划分的Normailized-Cut (Ncut)算法[4], k-way划分的Normailized-Cut (Ncut)算法[2], 针对大规模数据集的谱聚类方法[6], 混合二部图模型[7], 自动确定聚类数目的谱聚类方法[8], 基于背景的相似性度量方法和尺度参数问题[9], Nystrom逼近方法(用以减少求解特征问题时的计算复杂度)[10], 利用谱聚类解决文本集成聚类问题[11]以及MDS方法[12]和用图理论对话语者进行识别的方法[13]等. 虽然这些谱聚类方法与其他聚类方法相比易于理解和实现, 具有识别非凸分布的能力等优势, 但仍存在对初始值敏感等缺陷.

针对以上谱聚类方法存在对初始值敏感、聚类效率不高等问题, 本文提出一种基于模糊 K-调和均值的单词-文档谱聚类方法. 首先对谱聚类方法中的Laplacian矩阵进行处理, 使其满足对初始值不敏感的条件, 并加以证明; 然后用模糊 K-调和均值聚类方法代替谱聚类中的 K-均值聚类方法, 并采用模糊加权

收稿日期: 2010-11-08; **修回日期:** 2011-01-14.

基金项目: 国家自然科学基金项目(61175053, 61073133, 60973067); 教育部创新团队及重点科研培育项目(2011ZD010).

作者简介: 刘娜(1978−), 女, 博士生, 从事数据挖掘、文本摘要等研究; 鲁明羽(1963−), 男, 教授, 博士生导师, 从事机器学习、数据挖掘等研究.

的方法计算数据点与类别中心的距离. 该方法不仅能够解决单词-文档谱聚类方法中对初始值的敏感问题, 而且能够提高聚类效率, 增强数据的鲁棒性.

2 单词-文档谱聚类方法的敏感性分析

在谱聚类中, 人们采用图对集合中的对象以及它们之间的相互关系进行建模. 图中的顶点代表集合中的对象, 边表示对象间关系, 而关系的强弱则用边的权重表示. 由此, 图$G=(V,E)$的顶点集$V=\{v_1,v_2,\cdots,v_n\}$表示对象, 边集$E=\{\langle v_i,v_j\rangle\}$表示关系, 而边$\langle v_i,v_j\rangle$的权重$E_{ij}$则表示关系的强弱. 这样, 在图$G$中便可以将聚类问题转化为在图$G$上的图划分问题. 基于图论的最优划分准则是使得划分成的两个子图内部相似度最大, 子图之间的相似度最小[14]. 因此, 在谱聚类的图模型中, 聚类的任务体现在将原图切分成若干个子图, 使得每个子图内部边的权重较大, 而连接各子图的边的权重较小, 也就是使得切分子图时被切掉的边的权重之和尽量小.

谱聚类方法虽然可以在不同形状的样本空间中进行聚类, 并且聚类结果不收敛于局部最优解, 但其聚类结果对初始值敏感. 下面详细分析并证明单词-文档谱聚类方法对初始值敏感的真正原因.

目前, 大多数文档谱聚类方法通过向量空间模型将文档表示成单词-文档矩阵, 矩阵的行对应文档中的单词, 列对应文档. 比较典型的是将单词和文档建成二分图模型, 将聚类问题考虑成图形划分问题, 用第二大左右奇异向量产生单词和文档的聚类结果[15].

若已知集合

$$D=\{d_1,d_2,\cdots,d_m\},\ W=\{w_1,w_2,\cdots,w_n\}.$$

其中: D是文档的集合, W是文档中所包含的单词的集合. 则文档集D中包含m篇文档、n个单词, m篇文档可以被划分成K个类别. 文献[15]根据下式计算单词-文档相似矩阵:

$$A_{ij}=t_{ij}\log\frac{n}{|D_i|}. \tag{1}$$

其中: t_{ij}表示单词w_i在文档d_j中出现的次数, n表示文档总数, $|D_i|$表示包含单词w_i的文档数目.

本文在矩阵A的基础上构造矩阵

$$S=\begin{bmatrix}0 & A\\ A^{\mathrm{T}} & 0\end{bmatrix}. \tag{2}$$

对矩阵S进行归一化处理, 使

$$L=D^{-1}S, \tag{3}$$

其中D是对角阵, 即

$$D(i,i)=\sum_i S_{ij}. \tag{4}$$

下面分别分析证明矩阵S和矩阵L对单词-文档谱聚类结果敏感性的影响.

2.1 矩阵 S 对单词-文档谱聚类结果敏感性的影响

这里主要根据相似矩阵的相关知识, 分析证明同一文档集的不同矩阵S是相似的.

定义 1 如果两个n阶矩阵A与B相似, 则$A=P^{-1}BP$, P是n阶可逆矩阵, 且矩阵A与矩阵B的特征值相同, 特征向量$X_A=PX_B$.

已知集合$D=\{d_1,d_2,\cdots,d_m\}$, $W=\{w_1,w_2,\cdots,w_n\}$, D是文档的集合, W是文档中所包含的单词的集合. 若文档按$D=\{d_1,d_2,\cdots,d_m\}$, $W=\{w_1,w_2,\cdots,w_n\}$的顺序输入, 则根据式(1)和(2), 矩阵S的形式如下:

$$S=\begin{bmatrix}
0 & \cdots & 0 & 0 & A_{11} & A_{12} & \cdots & A_{1m}\\
0 & \cdots & 0 & 0 & A_{21} & A_{22} & \cdots & A_{2m}\\
\vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\
0 & \cdots & 0 & 0 & A_{n1} & A_{n2} & \cdots & A_{nm}\\
A_{11} & A_{21} & \cdots & A_{n1} & 0 & 0 & \cdots & 0\\
A_{12} & A_{22} & \cdots & A_{n2} & 0 & 0 & \cdots & 0\\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots\\
A_{1m} & A_{2m} & \cdots & A_{nm} & 0 & 0 & \cdots & 0
\end{bmatrix}.$$

若逆序输入, 即当$D=\{d_m,\cdots,d_2,d_1\}$, $W=\{w_n,\cdots,w_2,w_1\}$时, 矩阵S'的形式如下:

$$S'=\begin{bmatrix}
0 & \cdots & 0 & 0 & A_{nm} & \cdots & A_{n2} & A_{n1}\\
\vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\
0 & \cdots & 0 & 0 & A_{2m} & \cdots & A_{22} & A_{21}\\
0 & \cdots & 0 & 0 & A_{1m} & \cdots & A_{12} & A_{11}\\
A_{nm} & \cdots & A_{2m} & A_{1m} & 0 & 0 & \cdots & 0\\
\vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\
A_{n2} & \cdots & A_{22} & A_{12} & 0 & 0 & \cdots & 0\\
A_{n1} & \cdots & A_{21} & A_{11} & 0 & 0 & \cdots & 0
\end{bmatrix}.$$

根据矩阵的初等变换知识可证明, 矩阵S经若干次初等变换后可得到矩阵S', 即存在可逆矩阵P, 使得$S=P^{-1}S'P$, 即矩阵S与矩阵S'相似.

数据逆序输入是比较特殊的情况, 事实上可以证明以任意顺序输入数据得到的矩阵都与矩阵S相似. 由此, 可得到如下定理:

定理 1 已知集合

$$D=\{d_1,d_2,\cdots,d_m\},\ W=\{w_1,w_2,\cdots,w_n\},$$

当以不同顺序输入时, 得到的矩阵S与S'相似, 即$S=P^{-1}S'P$, 并且矩阵S与S'的特征值相同, 特征向量$X_S=PX_{S'}$.

2.2 矩阵 L 对单词-文档谱聚类结果敏感性的影响

这里主要根据定理1以及相似矩阵的相关知识, 分析证明同一文档集的不同矩阵L是相似的.

定义 2 如果两个n阶矩阵A与B相似, 即$A=$

$P^{-1}BP$, P 是 n 阶可逆矩阵, 则由矩阵 A 和 B 得到的对角阵满足 $D_A = P^{-1}D_BP$.

下面根据定理 1 和定义 2 证明矩阵 $L = D^{-1}S$ 对单词-文档谱聚类结果敏感性的影响:

对于输入顺序不同的两组数据, 首先根据式 (1) 和 (2) 分别得到矩阵 S 和 S', 然后根据式 (3) 分别对矩阵 S 和 S' 进行归一化处理, 得到

$$L_S = D_S^{-1}S,\ L_{S'} = D_{S'}^{-1}S'.$$

由定理 1 可知, 矩阵 S 与矩阵 S' 相似, 即 $S = P^{-1}S'P$. 又由定义 2 可知, $D_S = P^{-1}D_{S'}P$. 因此

$$\begin{aligned}&L_S = D_S^{-1}S =\\&(P^{-1}D_{S'}P)^{-1}S =\\&(P^{-1}D_{S'}^{-1}P)S =\\&(P^{-1}D_{S'}^{-1}P)(P^{-1}S'P) =\\&P^{-1}D_{S'}^{-1}S'P =\\&P^{-1}L_{S'}P,\end{aligned}$$

即 L_S 与 $L_{S'}$ 相似.

3 基于模糊 K-调和均值的单词-文档谱聚类方法

单词-文档谱聚类方法实际上是由以下 2 个步骤组成: 第 1 步生成矩阵; 第 2 步用 K-均值算法进行聚类. 本文的第 2 节已阐述将矩阵相似化, 使其对输入数据不敏感. 因此, 最终的聚类结果是否对初始数据敏感便由第 2 步采用的 K-均值算法决定. K-均值算法虽然是比较经典的聚类算法, 但其本身存在很多缺点, 所以本文提出的基于模糊 K-调和均值的单词-文档谱聚类方法除对输入矩阵进行修改以外, 还对 K-均值算法加以修改, 用模糊 K-调和均值聚类算法代替 K-均值算法, 以降低聚类过程中对初始值的敏感程度, 并在计算类别中心的过程中加入模糊的概念, 以提高聚类的正确性.

K-均值算法是一种迭代算法, 其目的是使如下目标函数最小化:

$$\mathrm{KM}(X,C) = \sum_{i=1}^{K}\sum_{x\in C_i}\|x-m_i\|^2.$$

K-调和均值算法也是一种迭代算法, 但它与 K-均值算法不同的是它采用了调和平均值的概念, 将所有点到所有类别中心的均方距离的调和平均值之和作为算法的目标函数 $\mathrm{KHM}(X,C)$, 即

$$\mathrm{KHM}(X,C) = \sum_{i=1}^{N}\frac{K}{\sum\limits_{l=1}^{K}\dfrac{1}{\|x_i-m_l\|^p}}, \tag{5}$$

这使得该算法对初始点几乎不敏感[16-17].

当目标函数最小时, 类别中心的更新公式为

$$m_l = \frac{\sum\limits_{i=1}^{N}P(m_l|x_i)a(x_i)x_i}{\sum\limits_{i=1}^{N}P(m_l|x_i)a(x_i)},$$

$$a(x)>0,\ P(m_l|x_i)\geqslant 0,\ \sum_{l=1}^{K}P(m_l|x_i)=1. \tag{6}$$

其中

$$P(m_k|x_i) = \frac{\dfrac{1}{d_{i,k}^{p+2}}}{\sum\limits_{l=1}^{K}\dfrac{1}{d_{i,l}^{p+2}}},$$

$$a_p[x] = \frac{\sum\limits_{l=1}^{K}\dfrac{1}{\|x-m_i\|^{p+2}}}{\Big[\sum\limits_{l=1}^{K}\dfrac{1}{\|x-m_i\|}\Big]^2}.$$

即

$$m_k = \frac{\sum\limits_{i=1}^{N}\dfrac{1}{d_{i,k}^{p+2}\Big(\sum\limits_{l=1}^{K}\dfrac{1}{d_{i,l}^{p}}\Big)^2}x_i}{\sum\limits_{i=1}^{N}\dfrac{1}{d_{i,k}^{p+2}\Big(\sum\limits_{l=1}^{K}\dfrac{1}{d_{i,l}^{p}}\Big)^2}}.$$

K-均值和 K-调和均值对数据的划分都是一种硬划分, 划分界限非常明显. 换言之, 若数据 x 属于 A 类则不能属于 B 类, 但实际划分界限往往不那么明显, 数据 x 既可以属于 A 类也可以属于 B 类. 因此, 为了提高划分效率, 将模糊的概念引入 K-调和均值算法. 在计算数据点与类别中心的距离 $d_{ij} = \|x_i - m_j\|$ 时, 考虑到同一数据点对不同类别的隶属程度, 对距离公式进行模糊加权处理, 即用隶属度函数 $w_{ij}\Big(w_{ij}\in[0,1], \sum\limits_{j=1}^{k}w_{ij}=1\Big)$ 表示数据点 i 属于类别 j 的程度, w_{ij} 越大, 数据点 i 越属于类别 j. 因此, 加入模糊概念后的模糊 K-调和均值算法的目标函数为

$$\mathrm{FKHM}(X,C) = \sum_{i=1}^{N}\frac{K}{\sum\limits_{j=1}^{k}\dfrac{1}{w_{ij}^a\|x-m_i\|^p}}, \tag{7}$$

其中

$$w_{ij} = \frac{1}{\dfrac{(\|x_i-m_j\|^2)^{1/a+2}}{\sum\limits_{z=1}^{K}(\|x_i-m_z\|^2)^{1/a+2}}},\ a\geqslant 0.$$

根据式 (6), 得到类中心的更新公式为

$$m_k=\frac{\displaystyle\sum_{i=1}^{N}\frac{w_{ik}^a}{d_{i,k}^{p+2}\Big(\displaystyle\sum_{l=1}^{K}\frac{w_{i,l}^a}{d_{i,l}^p}\Big)^2}x_i}{\displaystyle\sum_{i=1}^{N}\frac{w_{ik}^a}{d_{i,k}^{p+2}\Big(\displaystyle\sum_{l=1}^{K}\frac{w_{i,l}^a}{d_{i,l}^p}\Big)^2}}. \tag{8}$$

本文提出的基于模糊K-调和均值的单词-文档谱聚类算法具体步骤如下:

Step 1: 根据式(1)～(3)构造矩阵S和矩阵L.

Step 2: 计算矩阵L的前q个特征值$\lambda_1,\lambda_2,\cdots,\lambda_q$对应的特征向量$u_1,u_2,\cdots,u_q$, 并构造矩阵

$$Z=[\,D^{-1/2}U\,]. \tag{9}$$

其中: $U=\{u_1,u_2,\cdots,u_q\}$, $q=\arg\max x_i$, $x_i=\|\lambda_i-\lambda_{i-1}\|$.

Step 3: 初始化K个类别中心$c_1,c_2,\cdots,c_K$.

Step 4: 将矩阵Z的每一行看成一个数据对象x_i, $i=1,2,\cdots,m+n$, 根据类别中心对数据进行划分.

Step 5: 根据式(8)更新类别中心$c'_1,c'_2,\cdots,c'_K$.

Step 6: 返回Step 4, 直到类中心的距离不再改变, 即

$$\|[c'_1,c'_2,\cdots,c'_k]-[c_1,c_2,\cdots,c_k]\|\leqslant m,\ m=0.01. \tag{10}$$

4 实验结果与分析

为验证算法的有效性, 本文以ftp://ftp.cs.cornell.edu/pub/smart中的MEDLINE, CISI和CRANFIELD数据集为样本数据集, 将两个或多个样本数据集混合构成测试数据集, 比如, CISIMED-300测试数据集由CISI和MEDLINE两个样本集构成, 共300篇文档, 分别包含CISI样本集中的200篇文档, MEDLINE样本集中的100篇文档. 除使用以上提到的3个标准数据样本集外, 本文还使用大连海事大学智能技术研究中心分类样本集中的数据构造了1个具有5个类别的测试数据集Class-5. 测试数据集如表1所示, 前2个测试集是按单词数量的53%和64%取特征词, 后4个测试集是按单词数量的80%左右取特征词, 特征词的数量会影响最终的聚类结果.

表 1 测试数据集

数据集名称	文档数目	单词数目	特征词数目	文档描述
CISIMED-300	300	11 230	6 000	CISI: 200 篇; MED: 100 篇
CISIMEDCRAN-300	300	10 150	6 500	CISI: 100 篇; MED: 100 篇; CRAN: 100 篇
MEDCRAN-600	600	16 280	13 000	MED: 300 篇; CRAN: 300 篇
MEDCRANCISI-1 000	1 000	23 130	18 000	MED: 400 篇; CRAN: 300 篇; CISI: 300篇
MEDCRANCISI-1 500	1 500	27 670	20 000	MED: 500 篇; CRAN: 500 篇; CISI: 500篇
Class-5	500	42 380	34 000	政治、经济、文教、体育、国际类文档各 100 篇

4.1 基于模糊K-调和均值的单词-文档谱聚类方法实验结果

采用本文提出的方法对数据集进行测试, 测试结果用Purity和Entropy两个指标评价[15]. Purity指标表示某一个类别中占主导地位的类别数量与该类别数量的比值. 显然, 该比值越大, 说明某类中占主导地位的类别的“纯度”越高. Entropy值在0～1之间, 越靠近0, 越说明该类的成员是由同一个类组成; 越靠近1, 越说明该类的成员是由不同的类组成.

表2～表7是对不同数据集的测试结果. 其中, 表3是对CISIMEDCRAN-300数据集进行测试的实验结果. 表3中的数据表示CISI的100篇文档全被划分到C_0类; MEDLINE的100篇文档有10篇被划分到C_0类, 80篇被划分到C_1类, 10篇被划分到C_2类; CRANFIELD的100篇文档全被划分到C_2类. 从划分结果看, 共有20篇文档被划分到错误的类别中. 因此, 第C_0, 第C_1和第C_2类的Purity指标分别是0.909, 1和0.909; Entropy指标分别为0.277, 0和0.277. 显然, Purity指标比较接近1, Entropy指标比较接近0, 说明聚类的效果很好.

表 2 CISIMED-300数据集的聚类结果

类别	CISI	MEDLINE	Purity	Entropy
C_0	198	0	1	0
C_1	2	100	0.98	0.139

表 3 CISIMEDCRAN-300数据集的聚类结果

类别	CISI	MEDLINE	CRANFIELD	Purity	Entropy
C_0	100	10	0	0.909	0.277
C_1	0	80	0	1	0
C_2	0	10	100	0.909	0.277

表 4 MEDCRAN-600数据集的聚类结果

类别	MEDLINE	CISI	Purity	Entropy
C_0	220	10	0.957	0.258
C_1	80	290	0.784	0.753

表 5 MEDCRANCISI-1 000数据集的聚类结果

类别	MEDLINE	CRANFIELD	CISI	Purity	Entropy
C_0	250	130	10	0.641	0.678
C_1	40	270	10	0.844	0.466
C_2	10	0	280	0.966	0.137

表 6 MEDCRANCISI-1 500数据集的聚类结果

类别	MEDLINE	CRANFIELD	CISI	Purity	Entropy
C_0	410	80	90	0.707	0.735
C_1	70	410	20	0.82	0.516
C_2	20	10	390	0.929	0.276

表 7　Class-5 数据集的聚类结果

类别	政治	经济	文教	体育	国际	Purity	Entropy
C_0	60	0	20	0	20	0.600	0.590
C_1	0	70	10	10	40	0.538	0.571
C_2	20	10	40	20	10	0.400	0.914
C_3	20	0	10	70	10	0.636	0.642
C_4	0	20	20	0	20	0.333	0.687

从表 2～表 7 的数据中还可以看出, 文档数相同的情况下, 类别数越小, 划分效果越好; 类别数相同的情况下, 数据量越少, 划分效果越好. 因此, 本文提出的方法对类别相对较少、数据量不大的集合的划分效果要好于类别数较多、数据量较多的集合的划分效果.

4.2　与 K-均值、K-调和均值相比较的实验结果

本文提出的基于模糊 K-调和均值的单词-文档聚类方法(以下简称FKH单词-文档方法)首先使Laplacian矩阵对初始值不敏感, 然后利用模糊 K-调和均值聚类方法对初始值不敏感的特点, 在计算数据点与类别中心的距离过程中进行模糊加权处理, 使聚类结果不仅对初始值不敏感而且效率有所提高, 其有效性可参见上面第 4.1 节的分析. 本实验着重对所提出的模糊 K-调和均值聚类方法进行验证, 以说明用模糊 K-调和均值代替 K-均值方法的有效性. 实验过程是在矩阵不变的前提下分别对单词-文档谱聚类中的 K-均值、K-调和均值和模糊 K-调和均值方法进行比较. 这 3 者进行比较的意义在于, K-均值方法是单词-文档谱聚类中最经典、最常用的聚类方法, 该聚类方法的缺点是对初始值比较敏感, 克服该缺点的一个有效方法是采用对初始值几乎不敏感的 K-调和均值; 而模糊 K-调和均值方法是在 K-调和均值方法的基础上加入模糊的概念, 使聚类的效果更准确. 因此, 将这 3 者进行比较可以说明模糊 K-调和均值方法的敏感性和准确率.

因为文档的类别是已知的, 所以本实验使用簇与类别标志之间的标准化互信息 (NMI)[18]来分析聚类的结果. NMI 的值直接度量了聚类结果的鲁棒性, 当 NMI 的值越接近 1 时, 说明簇标记越接近原有的类别标记; 当 NMI 的值接近 0 时, 说明聚类过程只进行了一个随机划分. 图 1～图 4 是在不同迭代次数下, 采用上述 3 种方法在 MEDCRAN-600, MEDCRANCISI-1 000, MEDCRANCISI-1 500 和 Class-5 共 4 个数据集上得到的测试结果.

从图 1～图 4 中的曲线可以看出:

1) 虽然本文提出的方法在 4 个测试数据集上得到的 NMI 值高低不同, 但总体而言, 该方法对应的 NMI 值明显高于 K-均值方法和 K-调和均值方法的 NMI 值, 说明该方法的聚类效果较好.

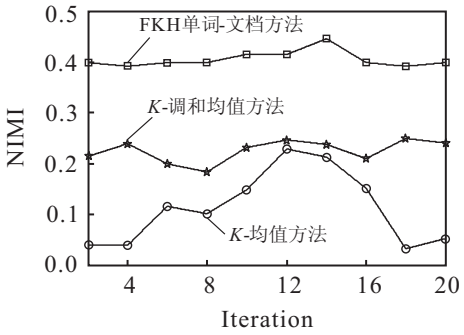

图 1　3 种方法在 MEDCRAN-600 数据集上的聚类结果

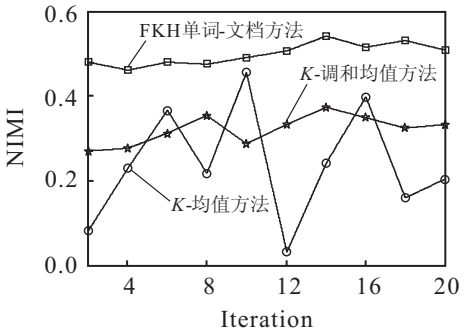

图 2　3 种方法在 MEDCRANCISI-1 000 数据集上的聚类结果

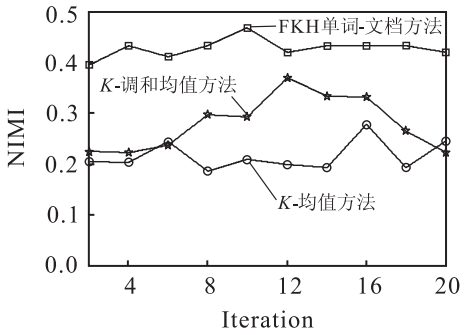

图 3　3 种方法在 MEDCRANCISI-1 500 数据集上的聚类结果

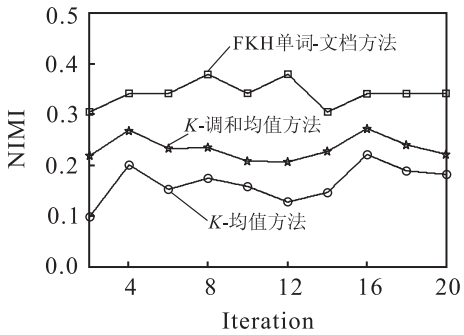

图 4　3 种方法在 Class-5 数据集上的聚类结果

2) 当迭代次数不同时, 本文方法的曲线波动较小, 说明算法的稳定性较好; K-调和均值方法的曲线也相对比较平滑; 相比而言, K-均值算法的曲线波动较大.

3) 对于不同的数据集而言, 本文方法所得到的 NMI 值不尽相同, 这与数据集本身的结构、内容以及特征词选择、阈值设定、循环次数等条件有关. 本文

下一步将重点验证特征词对聚类结果的影响, 以提高聚类效果.

4) 本文提出的方法虽然在聚类精度上取得了较好的效果, 但建立的矩阵是$m+n$阶的, 所以空间复杂度是$O((m+n)^2)$, $m+n$阶方阵的特征值求解的时间复杂度是$O((m+n)^3)$[19]. 因此, 如何降低该方法的时间、空间复杂度是今后研究工作的重点.

5 结 论

单词聚类和文档聚类是聚类分析中的热点问题. 针对现有谱聚类方法存在对初始值敏感、聚类效率不高等问题, 本文提出了一种基于模糊K-调和均值的单词-文档谱聚类方法. 该方法从矩阵相似的角度对谱聚类中的单词-文档Laplacian矩阵进行处理, 使其满足对初始值不敏感的条件. 因为最终影响谱聚类结果的因素除矩阵以外, 还有第2阶段采用的聚类算法, 所以本文用模糊K-调和均值算法代替K-均值算法, 使第2步的聚类结果不仅对初始值不敏感, 而且通过加入模糊的概念使聚类效果有所提高. 但是, 如何通过样本集本身的特点自动确定聚类的数目, 如何设定优化聚类过程中的初始类别中心以及如何降低算法的复杂度等问题尚有待进一步研究.

参考文献(References)

[1] Luxburg U V. A tutorial on spect ral clustering[J]. Statistics and Computing, 2007, 17(4): 395-416.

[2] Ng A Y, Jordan M L, Weiss Y. On spectral clustering: Analysis and an algorithm[C]. Advances in Neural Information Processing Systems 14. Columbia, 2001: 849-856.

[3] Tian Z, Li X B, Ju Y W. Spectral clustering based on matrix perturbation theory[J]. Science in China Series F: Information Sciences, 2007, 50(1): 63-81.

[4] Malik J, Belongie S, Leung T, et al. Contour and texture analysis for image segmentation[J]. Int J of Computer Vision, 2000, 43(1): 7-27.

[5] Weiss Y. Segmentation using eigenvectors: A unified view[C]. Proc IEEE Int Conf on Computer Vision. Corfu, 1999: 975-982.

[6] Prieto R, Jiang J, Choi C H. A new spectral clustering algorithm for large training sets[C]. Int Conf on Machine Learning and Cybernetics. Xi'an, 2003, 1: 147-152.

[7] Fern X Z, Brodley C E. Solving cluster ensemble problems by bipartite graph partitioning[C]. Proc of the 21st Int Conf on Machine Learning. New York: ACM, 2004: 281-288.

[8] Sanguinetti G, Laidler J, Lawrence N. Automatic determination of the number of clusters using spectral algorithms[C]. Proc of IEEE Machine Learning for Signal Processing. Connecticut, 2005: 28-30.

[9] Fischer I, Poland J. Amplifying the blockmatrix structure for spectral clustering[C]. Proc of the 14th Annual Machine Conf of Belgium and the Netherlands. Enschede, 2005: 21-28.

[10] Fowlkes C, Belongie S, Chung F. Spectral grouping using the Nystrom method[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 26(2) : 217-225.

[11] 徐森, 卢志茂, 顾国昌. 解决文本聚类集成问题的两个谱算法[J]. 自动化学报, 2009, 35(7): 997-1002.
(Xu S, Lu Z M, Gu G C. Two spectral algorithms for ensembling document clusters[J]. Acta Automatica Sinica, 2009, 35(7): 997-1002.)

[12] John Nerbonne. Data-driven dialectology[J]. Language and Linguistics Compass, 2009, 3(1): 175-198.

[13] Wieling M, Nerbonne J. Bipartite spectral graph partitioning to co-cluster varieties and sound correspondences in dialectology[C]. Proc of the 2009 Workshop on Graph-based Methods for Natural Language Processing. Singapore, 2009: 14-22.

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

[15] Dhillon I. Co-clustering documents and words using bipartite spectral graph partitioning[C]. Proc of the 7th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. San Francisco, 2001: 269-274.

[16] Zhang B, Hsu M, Dayal U. K-harmonic means-a data clustering algorithm[J/OL]. (1999-6-28)[2010-10-10]. http://www.hpl.hp.com/techreports/1999/HPL-1999-124.pdf.

[17] Zhang B. Generalized K-harmonic means-boosting in unsupervised learning[J/OL]. (2000-10-12)[2010-10-10]. http: // www.hpl.hp.com / techreports/2000/HPL-2000-137.html.

[18] Strehl A, Ghosh J. Cluster ensembles-a knowledge reuse framework for combining partitionings[J]. The J of Machine Learning Research, 2002, 3(1): 583-617.

[19] Berry M, Do T, O'Brien G, et al. SVDPACKC (version 1.0) user's guide[EB/OL]. (2007-11-22)[2010-10-10]. http://citeseer.ist.psu.edu/9643.html.