文章编号：1001-0920(2009)08-1277-04

文本聚类集成问题中的谱算法

徐 森[a]，卢志茂[b]，顾国昌[a]

（哈尔滨工程大学 a. 计算机科学与技术学院，b. 信息与通信工程学院，哈尔滨 150001）

摘 要：聚类集成中的关键问题是如何根据不同的聚类成员组合为更好的聚类结果. 引入谱聚类算法解决该问题，提出了基于相似度矩阵的谱算法(SMSA)，但该算法高昂的计算代价使其不适合大规模文本集. 进一步研究了谱聚类算法的特性，对超边的相似度矩阵进行谱分析，提出了基于超边相似度矩阵的元聚类算法(HSM-MCLA). 真实文本数据集的实验结果表明：SMSA 和 HSM-MCLA 比其他基于图划分的集成算法更优越；HSM-MCLA 可获得与 SMSA 相当的结果，而计算需求却明显低于 SMSA.

关键词：聚类分析；聚类集成；谱聚类；文本聚类；矩阵近似

中图分类号：TP391 **文献标识码**：A

Spectral algorithms in document cluster ensemble

XU Sen[a], *LU Zhi-mao*[b], *GU Guo-chang*[a]

(a. College of Computer Science and Technology, b. College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China. Correspondent: XU Sen, E-mail: xusen@hrbeu.edu.cn)

Abstract: A critical problem in cluster ensemble is how to combine multiple clusterers to yield a superior result. Spectral clustering is brought forth into solving this problem and similarity matrix spectral algorithm (SMSA) is proposed. Since the computational cost of SMSA is too high for large document datasets, the charactiristic of spectral clustering algorithm is further investigated. The hyperedges' similarity matrix are spectral analysed and hyperedges similarity matrix-based meta clustering algorithm (HSM-MCLA) is proposed. Experiments on real world document sets show that both SMSA and HSM-MCLA outperform other cluster ensemble techniques based on graph partitioning, and HSM-MCLA attains comparable results to SMSA with much lower computational cost than SMSA.

Key words: Clustering analysis; Cluster ensemble; Spectral clustering; Document clustering; Matrix approximation

1 引 言

聚类分析是极其困难而又非常重要的问题. 作为无监督学习的主要方法之一，其目标是发现对象/数据集中的“自然”分组(这样的“自然”分组称为“簇”)，使得簇内的对象彼此相似，不同簇内的对象不相似[1]. 目前已有上百种聚类算法，但还没有一种算法可以适用于不同大小、不同形状、不同密度甚至可能包含噪声的簇. 每种算法显式或隐式地对数据集强加某种结构，这又使得聚类有效性评估比较困难[1].

近年来，许多研究表明聚类集成技术能有效提高单聚类算法的精度和稳定性[2,3]. 其中关键问题在于：如何根据不同的聚类成员组合为更好的聚类结果. 比较常见的解决方法是根据聚类成员的标签得到相似度矩阵 $\boldsymbol{S}$ 后，使用层次聚类算法获得聚类结果[2]. 最常见的方法是 CSPA[3]，其中调用图划分算法 METIS[4] 对 $\boldsymbol{S}$ 进行聚类. Strehl 和 Ghosh 还提出了 HGPA 和 MCLA.

谱聚类算法对簇的形状不作强的假设，只需解决特征值问题，算法不存在局部最优解[5]. 另外，谱聚类可从不同的角度来解释其有效性，例如谱图理论[5]、矩阵扰动理论[5,6] 等.

鉴于谱聚类算法的上述优点，本文将其引入文本聚类集成问题，提出一种基于相似度矩阵的谱算法(SMSA). 然而该算法需要计算矩阵的特征值分解问题，时间复杂度为 $O(n^3)$，对于大规模文本集，其计算代价难以接受. 为此，本文进一步研究谱聚类算法的关键思想，即将原始高维空间中的数据映射到低维空间，使用 K 均值算法获得簇. 通过分析超边相似度矩阵的谱，间接得到超边的低维嵌入，并进

收稿日期：2008-08-25；**修回日期**：2008-12-22.

基金项目：国家自然科学基金项目(60603092)；高等学校博士学科点专项科研基金项目(20070217043).

作者简介：徐森(1983—)，男，江苏盐城人，博士生，从事人工智能、机器学习的研究；卢志茂(1972—)，男，哈尔滨人，教授，博士生导师，从事人工智能、智能信息处理等研究.

一步获得文本的低维嵌入. 由此提出一种基于超边相似度矩阵的元聚类算法(HSM-MCLA).

2 文本聚类集成问题

聚类集成可分为两步:第 1 步把数据集作为输入,输出多个聚类结果;第 2 步把不同的聚类结果作为输入,对它们进行组合,输出最终的结果. 第 1 步称为聚类成员的生成阶段;第 2 步称为组合(或集成、融合)阶段.

设 $\boldsymbol{D}=\{d_1,d_2,\cdots,d_n\}$ 为文本集,$\boldsymbol{P}=\{\boldsymbol{P}^{(1)},\cdots,\boldsymbol{P}^{(r)}\}$ 为对其划分得到的集成成员或聚类成员[3]. $\boldsymbol{H}=\boldsymbol{H}^{(1,\cdots,r)}=(\boldsymbol{H}^{(1)},\cdots,\boldsymbol{H}^{(r)})$ 为超图的 $n\times t$ 的邻接矩阵. 该超图有 n 个顶点,t 条超边,t 等于集成成员包含的簇个数之和. 聚类学习中的对象是无标签的,不同聚类成员得到的簇标签没有显式的对应关系. 另外,聚类成员还可能包含不同的簇个数,这使簇标签对应的问题极其复杂. 为使问题简化,本文将真实文本类别个数 k_0 作为单聚类算法的输入. 文本聚类集成问题可描述为如何根据 r 个聚类成员产生的文本簇,得到更好的聚类结果.

层次聚类算法显式或隐式地给最终的簇强加了某种结构[1],如单连接算法仅考虑最近邻信息容易产生具有链式结构的簇;全连接算法仅考虑全局信息,而局部信息往往比全局信息更为重要. 另外,通过使用有序表或堆来存放簇之间的距离,可将层次聚类方法的时间复杂度降到 $O(n^2\log n)$[1],但要解决大规模文本聚类集成问题,仍需要耗费大量的 CPU 时间.

CSPA 算法具有二次的计算复杂度,但是它调用了高效的图划分算法 METIS. 文献[7]指出,在以往的聚类集成对比实验研究中,CSPA 的性能稳定,而且效果也很好,即 CSPA 算法可看成是解决聚类集成问题的基准算法.

HGPA 算法首先生成一个超图,其中顶点为数据点,每个聚类成员中的每个簇均为一条封闭的超边,包含了属于该簇的所有顶点;然后用超图划分算法 HMETIS[8] 进行聚类.

MCLA 算法首先使用二元 Jaccard 系数度量超边之间的相似度;然后使用 METIS 压缩超边集得到元簇,最后把对象指派到与其最相关的元簇中. MCLA 和 HGPA 的时间和空间复杂度都是 n 的一次多项式.

上面 3 种算法很容易受 METIS 或 HMETIS 算法参数设置的影响. 例如不平衡因子 UB 的选取,会对簇的大小产生很大的影响.

3 文本聚类集成问题中的谱算法

文献[6]基于矩阵扰动理论对相似度矩阵 $\boldsymbol{S}$ 进行谱分析,不但得到权矩阵的谱与簇个数之间的关系,而且得到了 $\boldsymbol{S}$ 的特征向量与簇之间的关系.

3.1 基于相似度矩阵的谱算法

为了解决文本聚类集成问题,本文引入文献[6]的思想,得到一种新的解决方案,称为 SMSA. SMSA 算法的主要步骤如下:

输入:$d\times n$ 的词-文本共现矩阵 $\boldsymbol{A}$,簇个数 k_0.

1) 运行 K 均值算法产生 r 个聚类成员,K 均值算法采用余弦相似度函数,每次选取不同的初始值.

2) 构建超图的邻接矩阵 $\boldsymbol{H}$,计算 $\boldsymbol{S}=\boldsymbol{H}\boldsymbol{H}^{\mathrm{T}}$.

3) 计算 $\boldsymbol{S}$ 前 k 个最大特征值 $\lambda_1,\cdots,\lambda_k$ 对应的特征向量 $v_1,\cdots,v_k$,构建矩阵 $\boldsymbol{V}=[v_1,\cdots,v_k]$. 其中 $k=\underset{i}{\operatorname{argmax}}\delta_i$,$\delta_i=|\lambda_i-\lambda_{i-1}|$ 表示特征间隙.

4) 设 $z_i\in\boldsymbol{R}^k$ 为对应于 $\boldsymbol{V}$ 的第 i 行的列向量,使用 K 均值算法把 $\mathscr{Z}=\{z_i\mid i=1,\cdots,n\}$ 聚为 k 个簇 $\boldsymbol{C}_1,\cdots,\boldsymbol{C}_k$.

输出:文本簇 $\boldsymbol{D}_1,\cdots,\boldsymbol{D}_k$,其中 $\boldsymbol{D}_i=\{d_j\mid z_j\in\boldsymbol{C}_i,d_j\in\boldsymbol{D}\}$,$1\leqslant i\leqslant k$.

本文从矩阵低秩近似的角度解释算法 SMSA 的有效性. 当 $\boldsymbol{S}$ 的前 k 个特征值很大而余下的 $n-k$ 个特征值很小时,用一个秩为 k 的矩阵 $\boldsymbol{V}$ 来近似 $\boldsymbol{S}$. 在矩阵 $\boldsymbol{S}$ 中,使用 n 维表示文本数据之间的关系,当前 k 个特征值比其余特征值较大时,可在 k 维空间中表示这些数据. 设 $\boldsymbol{\Lambda}=\operatorname{diag}(\lambda_1,\cdots,\lambda_k)$,则 $\boldsymbol{S}^*=\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}}\approx\boldsymbol{S}$,并且特征间隙[5] $\delta=|\lambda_k-\lambda_{k+1}|$ 越大,$\boldsymbol{S}^*-\boldsymbol{S}$ 的 Frobenius 范数越小,因此这些特征向量是把数据嵌入到 R^k 空间中的最佳估计. 一旦用 k 维特征空间来表示数据,只需根据由 k 个特征向量定义的 k 维坐标,使用 K 均值算法进行聚类[5] 即可.

3.2 基于超边相似度矩阵的元聚类算法

SMSA 算法非常简单,只需解决特征值分解问题. 但对于大规模数据集,对 n 阶方阵 $\boldsymbol{S}$ 的特征值分解所需的高昂计算量难以接受. 注意到谱聚类算法是将原始高维数据嵌入到 k 维子空间中,再使用 K 均值算法获得聚类结果.

设 f^l 为簇 $\boldsymbol{C}_l$ 的指示向量,若 $d_i\in C_i$,则 $f_i^l=1$;否则,$f_i^l=0$. 考虑超边的邻接矩阵 $\boldsymbol{H}$,其中每个 $\boldsymbol{H}^{(i)}$ 都是一个 k_0 维子空间对应的正交基构成的矩阵. 显然,如果每个 $\boldsymbol{H}^{(i)}$ 都相同,则每条超边对应的列向量对应于一个簇 $\boldsymbol{C}_l$ 的指示向量 f^l. 实际情况是 K 均值算法易于陷入局部最优,使得每个 $\boldsymbol{H}^{(i)}$ 不完全相同. 现在的问题是如何根据这 r 个 $\boldsymbol{H}^{(i)}$ 得到一个最优的 k 维子空间.

本文将每条超边看成一个对象,而每个文本看成描述对象的一个属性,使用下面的方法对 $\boldsymbol{H}$ 的列进行加权:$\boldsymbol{H}_n=\boldsymbol{H}\boldsymbol{D}^{-1/2}$,其中 $\boldsymbol{D}=\operatorname{diag}(n^i)$,$n^i$ 为每

条超边包含的文本数，$i=1,\cdots,t$. 这样，$\boldsymbol{S}_t=\boldsymbol{H}_n{}^{\mathrm{T}}\boldsymbol{H}_n$ 即为 t 条超边的相似度矩阵.

本文把对 $\boldsymbol{S}_t$ 进行谱分析的算法称为 HSM-MCLA. 它与 MCLA 不同之处在于：MCLA 使用二元 Jaccard 系数度量超边之间的相似度，而 HSM-MCLA 则用超边对应的向量夹角的余弦值度量超边之间的相似度；MCLA 调用 METIS 划分超边集得到元簇，将文本划分到与其最相关的元簇中；HSM-MCLA 使用谱聚类的思想得到超边的低维嵌入，却不根据超边的低维嵌入使用 K 均值算法得到元簇，而是进一步得到文本的嵌入，再根据它们在该空间的坐标，使用 K 均值算法获得文本簇.

HSM-MCLA 算法的主要步骤如下：

输入：$d\times n$ 的词-文本共现矩阵 $\boldsymbol{A}$，簇个数 k_0.

1) 运行 K 均值算法产生 r 个聚类成员，K 均值算法采用余弦相似度函数，每次选取不同的初始值.

2) 构建超图的邻接矩阵 $\boldsymbol{H}$，计算对角度矩阵 $\boldsymbol{D}$，其对角元素 $d_j=\sum_{i=1}^{n}\boldsymbol{H}_{ij}$，计算 $\boldsymbol{H}_n=\boldsymbol{H}\boldsymbol{D}^{-1/2}$，计算 $\boldsymbol{S}_t=\boldsymbol{H}_n{}^{\mathrm{T}}\times\boldsymbol{H}_n$.

3) 计算 $\boldsymbol{S}_t$ 前 k 个最大特征值 $\lambda_1,\cdots,\lambda_k$ 对应的特征向量 $v_1,\cdots,v_k$，构建矩阵 $\boldsymbol{V}=[v_1,\cdots,v_k]$. 其中 $k=\underset{i}{\operatorname{argmax}}\delta_i,\delta_i=|\lambda_i-\lambda_{i-l}|$.

4) 计算 $\boldsymbol{Q}=\boldsymbol{H}_n\boldsymbol{V}_k$，设 $z_i\in\boldsymbol{R}^k$ 为对应于 $\boldsymbol{Q}$ 的第 i 行的列向量，使用 K 均值算法把 $\mathscr{Z}=\{z_i\mid i=1,\cdots,n\}$ 聚为 k 个簇 $\boldsymbol{C}_1,\cdots,\boldsymbol{C}_k$.

输出：文本簇 $\boldsymbol{D}_1,\cdots,\boldsymbol{D}_k$，其中 $\boldsymbol{D}_i=\{d_i\mid z_i\in\boldsymbol{C}_i,d_i\in\boldsymbol{D}\}$，$1\leqslant i\leqslant k$.

算法 HSM-MCLA 需要计算 $t\times n$ 的矩阵和 $n\times t$ 的矩阵之积，而 SMSA 需要计算 $n\times t$ 的矩阵和 $t\times n$ 的矩阵之积，它们都可通过高效的数值计算软件实现(如 Matlab). k 和 r 通常都远小于 n，所以 t 比 n 小得多，t 阶方阵 $\boldsymbol{S}_t$ 的特征值分解所需的时间明显少于 n 阶方阵 $\boldsymbol{S}$ 的特征值分解. 另外，SMSA 的空间复杂度为 $O(n^2)$，而 HSM-MCLA 仅为 $O(n)$. 因此从总体上看，HSM-MCLA 的计算代价明显小于 SMSA.

4　实验设计及结果分析

4.1　数据集

实验使用 6 个不同的数据集，表 1 给出了这些数据集的具体描述. 对于每个数据集，使用停用词表移去停用词，并且去掉出现在少于两个文本中的词.

数据集 hitech 和 reviews 取自 San Jose Mercury 报纸，它们是 TREC[9] 文本集的一部分，hitech 包含了关于计算机、电子、健康、医疗和科技方面的文章，reviews 包含了关于食物、电影、音乐、广播和饭店方面的文章，所有文本的类别标签唯一. 数据集 tr31 和 tr41 取自 TREC-6[9] 和 TREC-7[9] 文本集，这些数据的类别对应于某个特殊类别的查询. 数据集 re0 和 re1 取自 Reuters-21578 文本分类测试集[10]. 可将标签分为两部分，对于每个数据集，选择有唯一类别标签的文本.

表 1　实验数据集描述

数据集	文本个数	词 数	类别个数
hitech	2301	13170	6
reviews	4069	23220	5
tr31	927	10128	7
tr41	878	7454	10
re0	1504	2886	13
re1	1657	3758	25

4.2　实验结果与分析

由于文本的类别标签已知，可采用源自信息论的标准化互信息(NMI)[3] 来量化聚类结果和已知类别标签的匹配程度. 与纯度和熵等准则相比，NMI 值对 k 值的选取无偏好，因此是近年来比较流行的评价指标. 当两个类标签一一对应时，NMI 值达到最大值 1. 本文采用平均标准互信息(ANMI)来度量最终的结果和 r 个聚类标签之间的平均标准互信息. ANMI 值越大，表明聚类集成算法发现聚类成员之间一致性的能力越强.

将本文设计的两种算法 SMSA 和 HSM-MCLA 与 CSPA，HGPA，MCLA 进行对比实验，根据实验结果分析算法的性能.

在 6 个数据集上分别对 5 种集成算法进行聚类，获得的 NMI 值和 ANMI 值见表 2 和表 3，每组实验获得的最高 NMI 和 ANMI 值以黑体标出. 对于每个数据集，使用 K 均值算法获得 5 个聚类成员. 因为 HGPA 调用了 HMETIS 算法(不平衡因子 UB = 0.05)，而 HMETIS 得到了局部最优解，所以 HGPA 算法获得的聚类结果不稳定，运行 10 次取平均值；调用图划分算法 METIS 的 CSPA 和 MCLA 获得的聚类结果比较稳定；本文的两种谱算法同样能得到稳定的结果，这也是谱聚类算法的一个主要优点，即算法不存在局部最优[5].

根据表 2 和表 3 可作出不同的比较. 首先，SMSA 和 HSM-MCLA 除了在 hitech 上获得比 CSPA 略低的 NMI 值外，在所有其他数据集上获得的 NMI 和 ANMI 值都比 CSPA，HGPA 和 MCLA 高. 其次，比较 CSPA，HGPA 和 MCLA 的 NMI 值，CSPA 得到最好的聚类结果，而 HGPA 和 MCLA 的

表 2 不同聚类集成算法的 NMI 值

数据集	CSPA	HGPA	MCLA	SMSA	HSM-MCLA
hitech	**0.316**	0.196	0.177	0.302	0.302
reviews	0.506	0.433	0.435	**0.582**	**0.582**
tr31	0.498	0.472	0.473	0.596	**0.601**
tr41	0.606	0.391	0.439	**0.669**	0.651
re0	0.387	0.347	0.280	**0.400**	0.390
re1	0.528	0.470	0.447	0.571	**0.574**

表 3 不同聚类集成算法的 ANMI 值

数据集	CSPA	HGPA	MCLA	SMSA	HSM-MCLA
hitech	0.717	0.522	0.444	0.747	**0.751**
reviews	0.826	0.838	0.844	**0.967**	**0.967**
tr31	0.724	0.737	0.754	**0.866**	**0.866**
tr41	0.763	0.472	0.623	**0.848**	0.846
re0	0.780	0.725	0.622	0.838	**0.839**
re1	0.771	0.684	0.642	0.811	**0.827**

NMI 值互有高低. 这一实验结果与文献[3] 中的结果一致. 再次,比较 SMSA 和 HSM-MCLA,二者获得的性能非常接近. 最后,从总体上看,不同算法获得的NMI和ANMI值有很大的联系,即高的ANMI值能得到高的 NMI 值. 但也有例外情况, 例如 CSPA 在数据集 tr31 上获得的NMI值比 MCLA 高,而得到的 ANMI 值却低于 MCLA.

5 结 论

本文将谱聚类算法引入文本聚类集成问题,设计并实现了两种算法 SMSA 和 HSM-MCLA. 实验结果表明,两种基于谱聚类的集成算法可获得稳定的结果,它们比基于图划分的集成算法更优越. 另外,HSM-MCLA 的计算复杂度仅为 n 的一次多项式,可有效解决大规模文本聚类集成问题.

参考文献(References)

[1] Tan P N, Steinbach M, Kumar V. Introduction to data mining [M]. Boston: Addison-Wesley Longman Publishing Co, 2005.

[2] Fred A L, Jain A K. Combining multiple clusterings using evidence accumulation[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(6): 835-850.

[3] Strehl A, Ghosh J. Cluster ensembles: A knowledge reuse framework for combining partitionings [C]. AAAI'02. Edmonton: AAAI/MIT Press, 2002: 93-98.

[4] Karypis G, Kumar V. A fast and high quality multilevel scheme for partitioning irregular graphs[J]. SIAM J on Scientific Computing, 1998, 20(1): 359-392.

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

[6] 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.

[7] 罗会兰, 孔繁胜, 李一啸. 聚类集成中的差异性度量研究[J]. 计算机学报, 2007, 30(8): 1315-1324.
(Luo H L, Kong F S, Li Y X. An analysis of diversity measures in clustering ensembles [J]. Chinese J of Computers, 2007, 30(8): 1315-1324.)

[8] Karypis G, Aggarwal R, Kumar V, et al. Multilevel hypergraph partitioning: Applications in VLSI domain [C]. Proc of the Design and Automation Conf. New York, 1997: 526-529.

[9] TREC. Text retrieval conference[DB/OL]. [2007-11-22]. http:// trec. nist. gov.

[10] Lewis D D. Reuters-21578 Text Categorization Test Collection Distribution 1. 0[DB/OL]. [2007-11-22]. http://www. research. att. com/~lewis.

下 期 要 目