一种基于矩阵低秩近似的聚类集成算法

徐 森[1],周 天[2],于化龙[3],李先锋[1]
(1.盐城工学院信息工程学院,江苏盐城 224051;2.哈尔滨工程大学水声技术重点实验室,黑龙江哈尔滨 150001;
3.江苏科技大学计算机科学与工程学院,江苏镇江 212003)

摘 要: 首先将聚类集成问题归结为直观的最佳子空间的求解问题;随后根据线性代数理论将该问题描述为带约束条件的优化问题,通过放松离散约束条件进一步约简为矩阵低秩近似问题;最后通过求解超图的加权邻接矩阵的奇异值分解问题获得最佳子空间的一组标准正交基.据此,设计了一个基于矩阵低秩近似的算法,该算法根据每个对象在低维空间下的坐标使用K均值算法进行聚类,从而得到最终的结果.在多组基准数据集上的实验结果表明:较之于传统的聚类集成算法,本文的算法获得了更好的聚类结果,且效率较高.

关键词: 无监督学习; 聚类分析; 聚类集成; 矩阵低秩近似
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2013) 06-1219-06
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.06.028

Matrix Low Rank Approximation-Based Cluster Ensemble Algorithm

XU Sen[1], ZHOU Tian[2], YU Hua-long[3], LI Xian-feng[1]
(1. *School of Information Engineering, Yancheng Institute of Technology, Yancheng, Jiangsu* 224051, *China*;
2. *Science and technology on Underwater Acoustic Laboratory, Harbin Engineering University, Harbin, Heilongjiang* 150001, *China*;
3. *School of Computer Science and Engineering, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu* 212003, *China*)

Abstract: As an important extension to conventional clustering algorithms, cluster ensemble techniques became a hotspot in machine learning area. In this paper, cluster ensemble problem was first viewed as a direct problem of seeking the best subspace. And then, we formally described the problem as an optimization problem with constraint according to linear algebra, and further transformed into a matrix low rank approximation problem by relaxing the discrete constraint. Lastly, a set of orthonormal basis of the best subspace was attained by solving the singular value decomposition problem of the hypergraph's weighted adjacent matrix. Hereby, a matrix low rank approximation-based algorithm was proposed, which called K-means algorithm to cluster objects according to their coordinates in the low dimensional space and obtained the final clustering result. Experiments on baseline datasets demonstrate the effectiveness of the proposed algorithm, and it outperforms other baseline algorithms.

Key words: unsupervised learning; clustering analysis; cluster ensemble; matrix low rank approximation

1 引言

聚类集成由 Strehl 和 Ghosh[1]正式提出,经过十年的发展,已经成为机器学习领域的研究热点之一.聚类属于无监督学习,对象的类别/簇标签未知.因此,聚类集成存在类别标签对应问题,而这正是聚类集成问题非常困难的原因[2].根据是否解决簇标签对应问题,聚类集成算法可分为两大类.第一类方法通过引入超图的邻接矩阵 $\boldsymbol{H}$ 将对象之间的两两关系表示出来,并构建新的相似度矩阵[1,3] $\boldsymbol{S}=\boldsymbol{H}\times\boldsymbol{H}^{\mathrm{T}}$(也称 Co-Association(CA)矩阵[4]),从而有效避免类别标签对应问题.该类方法将每个聚类成员提供的结果作为判断两个对象是否应该放在一个簇中的证据(evidence)[4],完全利用了聚类成员提供的信息,是解决聚类集成问题最常见的方法.Strehl 和 Ghosh[1]基于图划分算法提出了 CSPA(Cluster-based Similarity Partitioning Algorithm)、HGPA(HyperGraph Partitioning Algorithm)和 MCLA(Meta-CLustering Algorithm).Fred 和 Jain[4]使用层次聚类算法对 CA 矩阵聚类,提出 EA(Evidence Accumulation)方法.Fern 和 Brodley[5]基于二部图模型.提出 HBGF(Hybrid Bipartite Graph Formulation)算法.Li 等[6]将聚类集成问题约简为 SNMF(Symmetric Non-negative Matrix Factorization)问题.Iam-On

收稿日期:2012-05-03;修回日期:2012-10-21
基金项目:国家自然科学基金(No.60970542,No.41006057,No.6110507);国家863重点项目(No.2008A09701);国际科技合作聘专重点项目;江苏省高校"青蓝工程"资助项目;盐城工学院人才引进专项基金(No.XKR2011019)

等[7]提出一种新的基于链接的方法.第二类方法通过重新加标签(re-labeling)显式地解决类别标签对应问题.该类方法的关键思想在于根据一个预先指定的参考标签将聚类集体中的其它标签重新标记,使不同划分之间的标签能尽可能地一致.Topchy 等人[2]用一个多项式分布的混合模型构建共识函数,然后用 EM 算法求解最终的簇.唐伟和周志华教授[8]认为,有对应关系的聚类标签所覆盖的相同对象的个数应该是最大的,根据这一启发式方法对标记向量进行配准,提出选择性聚类集成方法.Sevillano 等[9]基于 Borda 投票机制提出了软聚类集成方法 BordaConsensus.Carpineto 等[10]将共识函数设计问题看作一个未知的目标划分与给定划分之间 PRI(Probabilistic Rand Index)的优化问题,并采用随机优化算法求解.

上述方法中,HGPA 和 MCLA 的时间复杂度为文本个数 n 的一次多项式,但聚类质量较差,其它方法计算复杂度都较高,为 n 的二次甚至三次多项式.本文将聚类集成问题描述为一个最佳 k 维子空间的求解问题,并形式化描述为带约束条件的优化问题;随后放松离散化约束条件,求解超图的加权邻接矩阵的奇异值分解(Singular Value Decomposition,SVD)获得问题的连续解;最后使用简单 K 均值算法进行聚类,获得最终的结果.本文的方法属于第一类方法,然而与其它方法不同的是,本文直接对超图的邻接矩阵进行分析,而不需要构建相似度矩阵,因此,从问题求解的规模上得到了有效降低.

2 聚类集成问题

2.1 问题描述与求解

设 $X=\{x_1,x_2,\cdots,x_n\}$ 为数据集,$P=\{P^{(1)},\cdots,P^{(r)}\}$ 为使用 K 均值算法对其划分 r 次得到的聚类集体.为了简化问题,研究者们通常将真实的类别个数作为 K 均值算法的输入,本文沿用该方法.聚类集成问题可以描述为如何根据 K 均值算法获得的 r 个聚类成员,组合为更好的聚类结果.

设 $\boldsymbol{H}=\boldsymbol{H}^{(1\cdots r)}=(\boldsymbol{H}^{(1)},\cdots,\boldsymbol{H}^{(r)})$ 为超图的 $n\times t$ 的邻接矩阵,该超图有 n 个顶点,$t=r\times k$ 条超边.设 $\boldsymbol{f}^l$ 为簇 C_l 的指示向量:若 $d_i\in C_l$,则 $\boldsymbol{f}_i^l=1$;否则 $\boldsymbol{f}_i^l=0$.考虑 $\boldsymbol{H}$,其中每个 $\boldsymbol{H}^{(i)}$ 都是一个 k 维子空间 $(\mathbf{R}^k)^{(i)}$ 对应的正交基构成的矩阵,显然,如果每个 $\boldsymbol{H}^{(i)}$ 都相同,那么每条超边对应的列向量对应于一个簇 C_l 的指示向量 $\boldsymbol{f}^l$.而实际情况是由于 K 均值算法陷入局部最优,使得每个 $\boldsymbol{H}^{(i)}$ 不完全相同.若将这 r 个 $\boldsymbol{H}^{(i)}$ 看成是由"潜在的"未知的 k 维子空间对应的正交基构成的矩阵受到扰动后得到的 r 个扰动矩阵(perturbed matrix),则聚类集成问题就归结为如何根据这 r 个 $\boldsymbol{H}^{(i)}$ 得到"最优"的 k 维子空间的一组标准正交基构成的矩阵.

定义 1 考虑 $\mathbf{R}^d$ 的两个子空间 $\boldsymbol{V}_1$ 和 $\boldsymbol{V}_2$,不妨设 $p=\dim\boldsymbol{V}_1\geqslant\dim\boldsymbol{V}_2=q\geqslant 1$.$q$ 个主角度(principal angle)$\Theta_1,\Theta_2,\cdots,\Theta_q\in[0,\pi/2]$ 可以用来度量子空间的相似度,主角度按顺序 $l=1,2,\cdots,q$ 定义为[11]:$\cos\Theta_l=\max\limits_{u\in V_1,v\in V_2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v}$,其中 $(\boldsymbol{u}_l,\boldsymbol{v}_l)=\arg\max\limits_{u\in V_1,v\in V_2}\boldsymbol{u}^{\mathrm{T}}\boldsymbol{v}$,且 $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}_i=0$,$\boldsymbol{v}^{\mathrm{T}}\boldsymbol{v}_i=0$,$i=1,2,\cdots,l-1$,$\|\boldsymbol{u}\|_2=\|\boldsymbol{v}\|_2=1$.向量 $\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_q$ 和 $\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_q$ 被称为主向量(principal vector),向量对 $(\boldsymbol{u}_i,\boldsymbol{v}_i)$ 被称为主向量对(principal pair),$\|\cdot\|_2$ 表示欧氏范数.

设簇的指示向量构成的矩阵为 $\boldsymbol{F}\in\{0,1\}^{n\times k}$,不失一般性,可以将每个簇中的对象放到互相邻近的位置,使得 $\boldsymbol{F}$ 的每一列分别包含 $n_1,\cdots,n_k$ 个 1,n_i 表示簇 C_i 的大小,其余元素为 0.一个直观的想法是找到一个秩为 k 的矩阵 $\boldsymbol{F}^*\in\mathbf{R}^{n\times k}$,使得 r 个子空间 $(\mathbf{R}^k)^{(i)}$ 与 $\boldsymbol{F}^*$ 对应的子空间 $(\mathbf{R}^k)^*$ 的相似度之和最大化,即主角度的余弦和最大化.注意到,对于单位向量 $\boldsymbol{u}$、$\boldsymbol{v}$,不妨设其起点为原点,终点分别为 a、b,设 a、b 之间的欧氏距离为 d,$\boldsymbol{u}$、$\boldsymbol{v}$ 之间的夹角为 Θ,则:$\cos\Theta=(\|\boldsymbol{u}\|_2^2+\|\boldsymbol{v}\|_2^2-d^2)/2\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2=1-d^2/2$.

因此,主角度余弦和最大化问题可以形式化描述为如下的向量之间欧氏距离平方和最小化问题:

$$\boldsymbol{F}^*=\arg\min_{C}\sum_{l=1}^{r}\sum_{j=1}^{k}\sum_{i=1}^{k}\|\boldsymbol{u}_i-\boldsymbol{v}_{lj}\|_2^2$$

其中 $\boldsymbol{v}_{l1},\cdots,\boldsymbol{v}_{lk}$ 为 $(\mathbf{R}^k)^{(i)}$ 的标准正交基,$\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k$ 为 $\boldsymbol{C}$ 的标准正交基,且

$$\boldsymbol{u}_i=(\overbrace{0,\cdots,0}^{n_1},\cdots,\overbrace{1/n_i^{1/2},\cdots,1/n_i^{1/2}}^{n_i},\cdots,\overbrace{0,\cdots,0}^{n_k})^{\mathrm{T}} \tag{1}$$

定义 2 将 $\boldsymbol{Y}\in\mathbf{R}^{n\times k}$ 的列按自然序连接为列向量 $\mathrm{vec}\,\boldsymbol{Y}\in\mathbf{R}^{kn}$,该变换为线性双射,也称向量化(vectorization)[11].

于是有矩阵向量化内积等于矩阵内积:

$$<\boldsymbol{Y},\boldsymbol{Z}>\triangleq\mathrm{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Z})=(\mathrm{vec}\,\boldsymbol{Y})^{\mathrm{T}}\mathrm{vec}\,\boldsymbol{Z} \tag{2}$$

其中 $\mathrm{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Z})=\mathrm{tr}(\boldsymbol{Z}\boldsymbol{Y}^{\mathrm{T}})=\mathrm{tr}(\boldsymbol{Y}\boldsymbol{Z}^{\mathrm{T}})=\mathrm{tr}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y})=\mathbf{1}^{\mathrm{T}}(\boldsymbol{Y}\circ\boldsymbol{Z})\mathbf{1}$,$\mathrm{tr}(\cdot)$ 表示矩阵的迹,$\mathbf{1}$ 表示全 1 向量,"$\circ$"表示矩阵的阿达玛(Hadamard)积.

令式(2)中 $\boldsymbol{Z}=\boldsymbol{Y}\in\mathbf{R}^{n\times k}$,则矩阵的 Frobenius 范数由向量内积得到:

$$\|\boldsymbol{Y}\|_{\mathrm{F}}^2=\|\mathrm{vec}\,\boldsymbol{Y}\|_2^2=<\boldsymbol{Y},\boldsymbol{Y}>=\mathrm{tr}(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y})=\sum_{i,j}\boldsymbol{Y}_{ij}^2=\sum_i\lambda(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y})_i=\sum_i\sigma(\boldsymbol{Y})_i^2 \tag{3}$$

其中 $\boldsymbol{Y}_{ij}$ 表示 $\boldsymbol{Y}$ 的第 i 行第 j 列对应的元素,$\lambda(\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y})_i$ 表示 $\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}$ 的第 i 个特征值,$\sigma(\boldsymbol{Y})_i$ 表示 $\boldsymbol{Y}$ 的第 i 个奇异值.

因为矩阵空间 $\mathbf{R}^{n\times k}$ 与向量空间 $\mathbf{R}^{kn}$ 在欧氏意义下

是等距同构的(isometrically isomorphic),且向量化是线性双射,所以 $\| \mathrm{vec}\boldsymbol{X} - \mathrm{vec}\boldsymbol{Y} \|_2 = \| \boldsymbol{X} - \boldsymbol{Y} \|_F$. 因此,问题(1)就是如下的等价问题(4)经过向量化处理后得到的优化问题:

$\boldsymbol{F}^* = \arg\min_{\boldsymbol{C}} \sum_{i=1}^{r} \| \boldsymbol{H}_n^{(i)} - \boldsymbol{C} \|_F^2$,约束条件为:$\mathrm{rank}(\boldsymbol{C}) = k$,且 $\boldsymbol{C}$ 的列向量的元素仅取问题(1)中的离散值　　(4)

$\boldsymbol{F}^* = \arg\min_{\mathrm{rank}(C)=k} \| \boldsymbol{H}_n - \boldsymbol{C} \|_F^2$,约束条件为 $\boldsymbol{C}$ 的每一列元素仅取如问题(1)中所示的离散值　　(5)

其中,$\boldsymbol{H}_n = \boldsymbol{H}\boldsymbol{D}^{-1/2}$,$\boldsymbol{D} = \mathrm{diag}(n^i)$,$n^i$ 为每条超边包含的对象个数,$i = 1, \cdots, t$.

问题(5)为 0 - 1 整数规划问题,它是 NP-hard 问题,对其直接求解很困难. 因此,本文借鉴正则化谱聚类算法思想(具体参见文献[12]),首先放松离散化约束条件,求得问题的连续解,然后采用 K 均值算法获得离散解(即由 0,1 构成的簇的指示向量). 首先放松对 $\boldsymbol{C}$ 的元素的离散化约束条件,得到如下的矩阵低秩近似问题:

$$\boldsymbol{F}_c^* = \arg \min_{\mathrm{rank}(C)=k} \| \boldsymbol{H}_n - \boldsymbol{C} \|_F^2 \tag{6}$$

该问题的解即为问题(5)的连续解.

定义 3　不妨设 $\boldsymbol{H}_n$ 的秩为 $\mathrm{rank}(\boldsymbol{H}_n) = p \leqslant t$,其 SVD 被定义为:

$$\boldsymbol{H}_n = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \tag{7}$$

其中 $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{I}_n$,($\boldsymbol{I}_n$ 为 n 阶单位阵),$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \cdots, \sigma_n)$,$\sigma_i$ 为 $\boldsymbol{H}_n$ 的奇异值,且有当 $1 \leqslant i \leqslant p$ 时,$\sigma_i > 0$,当 $i \geqslant p+1$ 时,$\sigma_i = 0$. $\boldsymbol{U}$ 和 $\boldsymbol{V}$ 的前 p 列分别称为 $\boldsymbol{H}_n$ 的左、右奇异向量,且 $\boldsymbol{H}_n$ 的左、右奇异向量分别等于 $\boldsymbol{H}_n\boldsymbol{H}_n^{\mathrm{T}}$ 和 $\boldsymbol{H}_n^{\mathrm{T}}\boldsymbol{H}_n$ 的 p 个非零特征值对应的特征向量,而奇异值为 $\boldsymbol{H}_n\boldsymbol{H}_n^{\mathrm{T}}$ 和 $\boldsymbol{H}_n^{\mathrm{T}}\boldsymbol{H}_n$ 的特征值的非负平方根[13].

定理 1　设 $\boldsymbol{H}_n$ 的 SVD 由式(7)给出,若定义 $\boldsymbol{H}_k = \sum_{i=1}^{k} \boldsymbol{u}_i \cdot \sigma_i \cdot \boldsymbol{v}_i^{\mathrm{T}} = \boldsymbol{U}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k^{\mathrm{T}}$,且 $k < p$,其中 $\boldsymbol{U}_k = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_k]$,$\boldsymbol{\Sigma}_k = \mathrm{diag}(\sigma_1, \cdots, \sigma_k)$,$\boldsymbol{V}_k = [\boldsymbol{v}_1 \cdots \boldsymbol{v}_k]$,$\sigma_1 \geqslant \cdots \geqslant \sigma_k \geqslant \cdots \geqslant \sigma_p > \sigma_{p+1} = \cdots = \sigma_t = 0$,则下式成立:

$$\min_{\mathrm{rank}(\boldsymbol{C})=k} \| \boldsymbol{H}_n - \boldsymbol{C} \|_F^2 = \| \boldsymbol{H}_n - \boldsymbol{H}_k \|_F^2 = \sigma_{k+1}^2 + \cdots + \sigma_p^2 \tag{8}$$

根据定理 1,问题(5)的连续解为 $\boldsymbol{F}_c^* = \boldsymbol{H}_k$. 由此,我们不仅获得了 n 个对象的 k 维嵌入,即 $\boldsymbol{U}_k$,还得到了 t 条超边的 k 维嵌入,即 $\boldsymbol{V}_k$. 需要指出的是,虽然由 $\boldsymbol{H}_k$ 的列向量生成的 k 维子空间是唯一的,但是 $\boldsymbol{H}_k$ 并不唯一,这是因为对于任意的正交变换 $\boldsymbol{P}$,$\boldsymbol{H}_k\boldsymbol{P}$ 同样能使式(8)成立. 考虑到 K 均值算法的目标函数具有旋转不变性,而正交变换相当于旋转操作,因此,为了获得离散的指示矩阵 $\boldsymbol{F}^*$,本文使用 K 均值算法对 $\boldsymbol{U}_k$ 的行进行聚类.

2.2　算法描述

将本文的算法称为基于矩阵低秩近似的算法(Matrix Low Rank Approximation-based Algorithm,简记为 MLRAA),根据式(7),等式两边同时右乘 $\boldsymbol{V}\boldsymbol{\Sigma}^{\dagger}$,其中 $\boldsymbol{\Sigma}^{\dagger}$ 为 $\boldsymbol{\Sigma}$ 的广义逆(generalized inverse),经过简单整理得到 $\boldsymbol{U} = \boldsymbol{H}_n\boldsymbol{V}\boldsymbol{\Sigma}^{\dagger}$,故可以通过求解 t 阶方阵 $\boldsymbol{H}_n^{\mathrm{T}}\boldsymbol{H}_n$ 的特征值和特征向量间接求得 $\boldsymbol{U}$.

MLRAA 算法的主要流程如下所示,因为只需求解前 k 个最大奇异值 $\sigma_1, \cdots, \sigma_k$ 对应的左奇异向量,而 $\sigma_1, \cdots, \sigma_k$ 大于 0,所以第 3 步中只需求解 $\boldsymbol{\Sigma}_k = \mathrm{diag}(\sigma_1, \cdots, \sigma_k)$的逆矩阵,即 $\Lambda_k^{-1/2}$.

输入:数据集 $X = \{x_1, x_2, \cdots, x_n\} \in \mathbf{R}^{d \times n}$($d$ 为特征个数),簇个数 k.

(1)运行 K 均值算法 r 次,每次随机选取初始质心.

(2)构建超图的邻接矩阵 $\boldsymbol{H}$,计算 $\boldsymbol{H}_n = \boldsymbol{H}\boldsymbol{D}^{-1/2}$,其中 $\boldsymbol{D} = \mathrm{diag}(n^i)$,$n^i$ 为每条超边包含的对象个数,$i = 1, \cdots, t$.

(3)计算矩阵 $\boldsymbol{H}_n^{\mathrm{T}}\boldsymbol{H}_n$ 的前 k 个最大特征值 $\lambda_1, \cdots, \lambda_k$ 及其对应的特征向量 $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_k$,构建矩阵 $\boldsymbol{V}_k = [v_1 \cdots v_k]$,$\boldsymbol{\Lambda}_k = \mathrm{diag}(\lambda_1, \cdots, \lambda_k)$,求 $\boldsymbol{U}_k = \boldsymbol{H}_n\boldsymbol{V}_k\boldsymbol{\Lambda}_k^{-1/2}$.

(4)设 $z_i \in \mathbf{R}^k$ 为对应于 $\boldsymbol{U}_k$ 的第 i 行的列向量,使用 K 均值算法把 $\boldsymbol{Z} = \{z_i \mid i = 1, \cdots, n\}$ 聚为 k 个簇 $C_1, \cdots, C_k$.

输出:划分结果 $\pi = \{D_1, \cdots, D_k$,其中 $D_i = \{x_j \mid z_j \in C_i, x_j \in X, i = 1, \cdots, k\}\}$.

3　实验

3.1　实验数据集

实验采用 11 个基准数据集,分别来自 TREC、Reuters 和 WebACE,从而保证了数据集的多样性,其具体描述如表 1 所示. 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 classic 由用于评估信息检索系统的四种摘要构成,每个摘要集构成单独的一类. 数据集 fbis 来自 TREC-5[14],它由 Foreign Broadcast Information Service 数据集构成,每一类对应于数据集中的一个类别. 数据集 hitech,reviews 和 sports 取自 San Jose Mercury 报纸上的文章,它们是 TREC[14] 文本集的一部分,每个数据集都由包含不同文章的多个主题所构成. 数据集 la12 来自 TREC-5[14],它由 Los Angles Times 上的文章构成. 数据集 tr31 和 tr41 取自 TREC-6[14]和 TREC-7[14]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[15],其标签被分为两个子集,对于每个数据集,选择有唯一类别标签的

文本.数据集 k1a,k1b 来自于 WebACE project[16],每个文本对应于 Yahoo!(http://www.yahoo.com)主题层次下的一个网页.

表 1 实验数据集描述

数据集	文本个数	特征词个数	类数
classic	7094	12009	4
fbis	2463	12674	17
hitech	2301	13170	6
reviews	4069	23220	5
la12	6279	21604	6
tr31	927	10128	7
tr41	878	7454	10
re0	1504	2886	13
re1	1657	3758	25
k1a	2340	13879	20
k1b	2340	13879	6

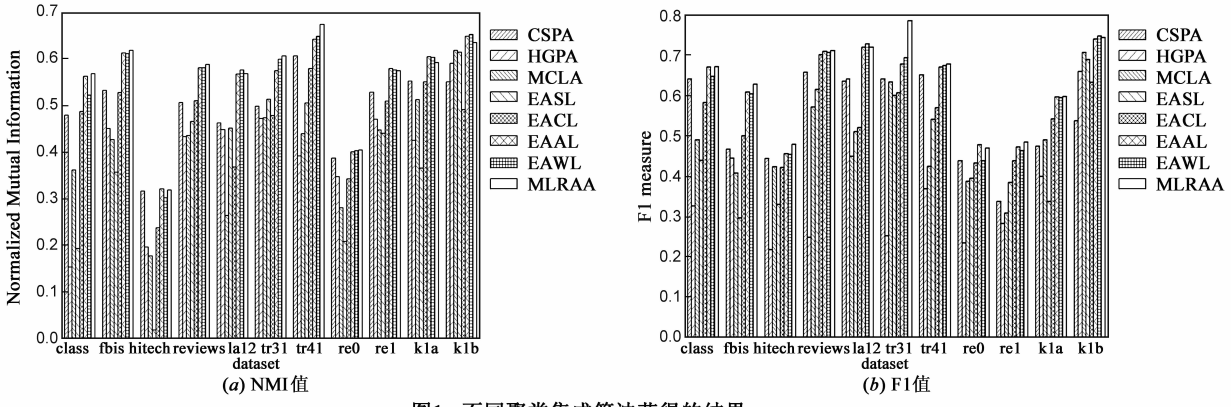
3.2 实验结果与分析

本文采用源于信息论的 NMI 值(Normalized Mutual Information)[1]来量化聚类结果和已知类别标签的匹配程度.当两个类别标签一一对应时,NMI 值达到最大值 1.另外,本文还采用了在信息检索与自然语言处理领域常用的 F1 值(F1-measure)来综合评价文本聚类质量.F1 值越大,聚类质量越高.

将本文的 MLRAA 算法与以下 7 个算法相比较,它们是:文献[1]提出的 CSPA、HGPA、MCLA;文献[4]提出的 4 个基于单连接(Single-Linkage,SL)、全连接(Complete-Linkage,CL)、组平均(Average-Linkage,AL)和 Ward 的 EA 算法,为方便起见,分别简记为 EASL、EACL、EAAL 和 EAWL.

将 8 个聚类集成算法分别在不同数据集上进行聚类,获得的 NMI 值和 F1 值见图 1.对于每个数据集,在聚类成员生产阶段,首先将每个文本根据 TF-IDF(Term Frequency-Inverse Document Frequency)加权,将其归一化;随后使用 K 均值算法产生 5 个聚类成员,K 均值算法每次随机选取初值,并且采用余弦函数计算文本之间的相似度.在聚类集成阶段,将真实类别个数提供给 8 个聚类集成算法.HGPA 算法获得的聚类结果不稳定,我们运行 10 次取平均值;其他算法都获得了稳定的结果.

由图 1 易见:(1)比较 CSPA、HGPA 和 MCLA,除了在数据集 k1b 上获得的 NMI 值低于 MCLA,CSPA 都获得了最好的结果.(2)比较 EASL、EACL、EAAL 和 EAWL,EAAL 和 EAWL 的聚类效果明显优于 EASL 和 EACL.(3)与基于图划分的算法相比,MLRAA 算法在所有其它数据集上获得的 NMI 值和 F1 值都比 CSPA、HGPA 和 MCLA 高.(4)与基于层次聚类的算法相比,MLRAA 在数据集 hitech、re1 和 k1a 上的 NMI 值略低于 EAAL,在 k1b 上的 NMI 值略低于 EAWL,而在其它 7 个数据集上都获得了最高的 NMI 值;MLRAA 在 re0 上获得的 F1 值比 EAAL 略低,在 la12 和 k1b 上的 F1 值略低于 EAWL,在其它 8 个数据集上都获得了最高的 F1 值.总体来看,与 EAAL 和 EAWL 相比,MLRAA 的聚类质量毫不逊色.



(*a*) NMI值 (*b*) F1值

图1 不同聚类集成算法获得的结果

在实际应用中,数据集的规模通常是极其庞大的,因此聚类算法的运行效率是评价算法扩展性的一个关键因素.表 2 列出了 8 个聚类集成算法在集成阶段的运行时间,表中最后一行 Total 表示每个算法在 11 组数据集上运行的总时间.根据表 2,可以看出,算法的总体运行时间满足下面的次序:HGPA < MCLA < MLRAA < CSPA < EASL < EAAL < EAWL < EACL,即 HGPA 和 MCLA 算法的运行效率最高,MLRAA 算法次之,基于层次聚类的算法运行效率最低.特别地,在 classic(n = 7094)这个规模较大的数据集上,MLRAA 的运行时间仅为 1.25s,而层次聚类算法的运行时间则高达几千甚至几万秒.显然,当进行海量数据挖掘时,层次聚类算法高昂的计算成本很难让人接受,而 MLRAA 算法较高的运行效率则使其易于扩展到大规模应用领域.

综合前面的实验结果,可以发现,与层次聚类算法相比,本文算法在聚类精度上毫不逊色,而在运行效率上则高效许多;与图划分算法相比,本文的算法在聚类精度上明显高出许多,而在运行效率上则相差无几.

表 1　聚类集成算法在集成阶段的运行时间(单位:s)

dataset	CSPA	HGPA	MCLA	EASL	EACL	EAAL	EAWL	MLRAA
classic	23.09	0.38	0.81	7778.15	24218.43	24483.70	24651.65	1.25
fbis	1.45	0.29	0.80	48.66	144.65	136.43	147.33	0.89
hitech	3.98	0.11	0.30	89	199	198	193	0.67
reviews	9.45	0.13	0.38	861	3375	2127	2982	0.90
la12	14.71	0.31	0.82	2301.03	7090.75	7101.17	7328.53	1.38
tr31	0.54	0.08	0.18	2.77	9.35	7.99	9.67	0.09
tr41	0.42	0.09	0.22	1.44	4.68	4.11	5.28	0.12
re0	0.94	0.16	0.38	11.64	36.46	33.87	37.82	0.29
re1	0.88	1.19	0.73	5.64	19.13	19.63	23.23	0.72
kla	1.40	1.18	0.86	25.58	75.39	70.86	80.73	0.89
klb	2.80	0.12	0.31	86.69	267.24	210.79	242.39	0.79
Total	59.66	4.04	5.79	11211.6	35440.08	34393.55	11201.63	7.99

4　结论

本文首先从最佳低维子空间的求解出发,将聚类集成问题形式化描述为带约束条件的优化问题;随后放松离散化约束条件,通过求解超图的加权邻接矩阵的奇异值分解获得了问题的连续解;最后使用简单 K 均值算法进行聚类,获得了最终的结果.对比实验结果表明,本文设计的 MLRAA 算法获得了优越的聚类结果,且算法运行效率较高.本文的方法为解决聚类集成问题提供了一种新思路.

致谢　感谢审稿专家给本文提出了宝贵的参考意见.

参考文献

[1] STREHL A, GHOSH J. Cluster ensembles——a knowledge reuse framework for combining partitionings[J]. The Journal of Machine Learning Research, 2002, 3(12): 583 – 617.

[2] TOPCHY A, JAIN A K, PUNCH W. A mixture model for clustering ensembles[A]. Michael W B, et al. Proceedings of the 4th SIAM International Conference on Data Mining[C]. Florida: Society for Industrial and Applied Mathematics, 2004. 379 – 390.

[3] 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 Sen, LU Zhi-mao, GU Guo-chang. Two spectral algorithms for ensembling document clusters[J]. Acta Automatica Sinica, 2009, 35(7): 997 – 1002. (in Chinese)

[4] FRED A, LOURENGO A. Supervised and Unsupervised Ensemble Methods and Their Applications[M]. Berlin: Springer, 2008. 3 – 30.

[5] FERN X Z, BRODLEY C E. Solving cluster ensemble problems by bipartite graph partitioning[A]. Russ G, Dale S. Proceedings of 21st International Conference on Machine Learning [C]. New York: ACM, 2004. 36 – 43.

[6] LI T, DING C, JORDAN M I. Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization[A]. Naren R, Osmar Z. Proceedings of the 7th IEEE International Conference on Data Mining[C]. Washington: IEEE Computer Society, 2007. 577 – 582.

[7] IAM On N, BOONGEON T, GARRETT S, PRICE C. A link-based cluster ensemble approach for categorical data clustering [J]. IEEE Transactions on Knowledge and Data Engineering, 2012, 24(3): 413 – 425.

[8] 唐伟,周志华.基于 Bagging 的选择性聚类集成[J].软件学报,2005,16(4):496 – 502.
TANG Wei, ZHOU Zhi-hua. Bagging-based selective clusterer ensemble[J]. Journal of Software, 2005, 16(4): 496-502. (in Chinese)

[9] SEVILLANO X, ALIAS F, SOCORO J C. BordaConsensus: a new consensus function for soft cluster ensembles[A]. Wessel K, et al. Proceedings of the 30th Annual International ACM SIGIR[C]. New York: ACM, 2007. 743 – 744.

[10] CARPINETO C, ROMANO G. Consensus clustering based on a new probabilistic rand index with application to subtopic retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(12): 2315 – 2326.

[11] DATTORRO J. Convex Optimization & Euclidean Distance

Geometry[M]. USA: Meboo Publishing, 2005. 8 – 45.
[12] LUXBURG U V. A tutorial on spectral clustering[J]. Statistics and Computing, 2007, 17(4): 395 – 416.
[13] BERRY M W. Large-scale sparse singular value computations [J]. The International Journal of Supercomputer Applications, 1992, 6(1): 13 – 49.
[14] TREC. Text Retrieval Conference [OL]. http://trec.nist.gov, 2011-09-12.
[15] LEWIS D D. Reuters-21578 Text Categorization Test Collection Distribution 1.0[OL]. http://www.daviddlewis.com/resources/testcollections/reuters21578/, 2011-09-12.
[16] BOLEY D, et al. Document categorization and query generation on the world wide web using WebACE[J]. Artificial Intelligence Review, 1999, 13(5 – 6): 365 – 391.
[17] 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 Hui-lan, KONG Fan-sheng, LI Yi-xiao. An analysis of diversity measures in clustering ensembles[J]. Chinese Journal of Computers, 2007, 30(8): 1315 – 1324. (in Chinese)

作者简介

徐 森(通信作者) 男,1983 年 1 月出生于江苏省滨海县.现为盐城工学院副教授,从事模式识别与智能信息处理方面的研究工作.
E-mail: xusen@ycit.cn

周 天 男,1979 年 7 月出生于江苏省响水县.现为哈尔滨工程大学副教授,硕士生导师,从事智能信息处理和目标探测与识别方面的研究工作.
E-mail: zhoutian@hrbeu.edu.cn