文章编号:1672-3961(2013)02-0035-07 DOI:10.6040/j.issn.1672-3961.0.2012.289

谱聚类中特征向量的 Bagging 选取方法

王兴良,王立宏*,李海军
(烟台大学计算机学院,山东 烟台 264005)

摘要:谱聚类算法中用亲和矩阵特征值最大的 k 个特征向量并不总是能有效地发现数据集的结构。为了选取较好特征向量,提出了一种特征向量的 Bagging 选取算法。以成对约束计分方法为评价标准,对特征向量进行评价并选出较好的特征向量,将多次选择的特征向量进行 Bagging 集成(Bootstrap aggregating),得出 k 个特征向量的组合。该算法能够较好地选取出特征向量,根据 UCI 实验数据集的测试,证实该算法对测试数据集可以得出较好的预测结果。

关键词:特征向量选择;谱聚类;Bagging 方法;约束计分;拉普拉斯矩阵
中图分类号:TP301.6 **文献标志码**:A

Eigenvector selection in spectral clustering based on Bagging

WANG Xing-liang, WANG Li-hong*, LI Hai-jun
(School of Computer Science & Technology, Yantai University, Yantai 264005, China)

Abstract: For the spectral clustering algorithm, the largest k eigenvectors of the affinity matrix derived from the dataset were not always able to find the structure of dataset effectively. An eigenvector selection algorithm in spectral clustering based on Bagging was proposed in order to select better eigenvectors. The eigenvectors were evaluated by pairwise constraints score. First, some eigenvectors were ranked according to their constraint scores, and then the suitable eigenvectors were selected from the ranking list, finally the optimal combination of k eigenvectors was obtained by Bagging-based ensemble algorithm. The better eigenvectors could be achieved. Experimental results on UCI benchmark datasets showed that this algorithm could gain satisfactory prediction results.

Key words: eigenvector selection; spectral clustering; Bagging method; constraint score; Laplacian matrix

0 引言

聚类旨在发现数据集的内在结构,是机器学习与模式分析的重要方法。在各种聚类算法中,谱聚类算法得到广泛的关注和应用[1-5]。因为谱聚类算法在数据集的亲和矩阵的特征向量空间进行聚类,对于非线性可分的问题可以给出满意的聚类结果。通常的谱聚类算法在特征值最大的 k 个特征向量(以下简称为最大的 k 个特征向量)张成的空间内表示数据点并聚类成 k 个簇。文献[6]首次关注特征向量的选择问题,认为最大的 k 个特征向量并不总是能有效地发现数据的结构。文献[7]进一步地以合成数据和几个 UCI 数据集来证实确实存在有比最大的 k 个特征向量更好的特征向量组合,并且设计了基于熵的评价标准对特征向量进行选择,对较重要的特征向量进行组合,找出最优的特征向量组合。该组合可能并不是 k 个特征向量,或者多于 k 或者少于 k。文献[8]认为特征值之间的差可以用来确定多少个特征向量对聚类有意义,这样得出

收稿日期:2012-10-12 **网络出版时间**:2013-04-01 10:48
网络出版地址:http://www.cnki.net/kcms/detail/37.1391.T.20130401.1048.005.html
基金项目:国家自然科学基金资助项目(61170224);山东省计算机网络重点实验室开放课题基金资助项目(SDKLCN_2012_03)
作者简介:王兴良(1988-),男,山东临朐人,硕士研究生,主要研究方向为数据挖掘. E-mail:wangxingliang0911@163.com
*通讯作者:王立宏(1970-),女,吉林镇赉人,教授,博士,硕士生导师,主要研究方向为数据挖掘. E-mail:wanglh_000@163.com

的特征向量可能会多于 k 个，但聚类结果较好。本研究探讨如何找出恰好 k 个特征向量，使用它们张成的空间来表示数据和聚类数据的效果较通常的谱聚类算法要好。

1 谱聚类算法和相关问题

1.1 谱聚类算法

谱聚类算法通过构造各数据点的近邻图，得出相应的拉普拉斯矩阵 $\boldsymbol{L}$。求出 $\boldsymbol{L}$ 的 k 个最大的特征值对应的特征向量 $\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k$，以 $\boldsymbol{U}=[\boldsymbol{u}_1,\cdots,\boldsymbol{u}_k]$ 的每一行作为一个数据点在特征空间的表示，以常规的聚类算法如 k-means 完成聚类。由于谱聚类能完成非球形类的识别，因此得到广泛的应用。

拉普拉斯矩阵 $\boldsymbol{L}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$，其中 $\boldsymbol{A}$ 是亲和矩阵，$\boldsymbol{D}$ 是对角矩阵且 $D_{ii}=\sum_{j=1}^{n}A_{ij}$。亲和矩阵 $\boldsymbol{A}$ 中的元素 A_{ij} 表示数据集中第 i 点和第 j 点的相似程度，计算 A_{ij} 一般采用高斯相似性函数[9]：$A_{ij}=\mathrm{e}^{-\|x_i-x_j\|^2/2\sigma^2}$。公式中参数 σ 对结果影响较大，且 σ 值不易选取。本研究采用尺度参数 σ_i[7]：

$$A_{ij}=\mathrm{e}^{-d^2(x_i,x_j)/\sigma_i\sigma_j}, \tag{1}$$

公式中 $\sigma_i=d(x_i,x_{il})$，x_{il} 是点 x_i 的第 l 个最近邻点，$d(x_i,x_j)$ 是点 x_i 和 x_j 的欧几里德距离。但是 l 的确定仍与结果密切相关，文献[7]将 l 设为 7，对此本研究进行了试验研究。

1.2 约束分值

特征选择（feature selection）是数据挖掘和机器学习的重要预处理步骤，特征选择能有效降维，提高学习精度，增强学习的可理解性。特征选择方法有很多种，按照是否使用训练数据的类别信息可以大体分为 3 类[10]：（1）无监督方法，如方差（variance）[11]、拉普拉斯分值（Laplacian Score）[12]等；（2）有监督方法，如 Fisher 分值（Fisher Score）[11]；（3）半监督方法，即通过数据的成对约束信息来选取特征，如约束分值（Constraint Score）[10]。

WAGSTAFF K 于 2000 年提出了两种实例间的成对约束 must-link（ML）和 cannot-link（CL）[13]，由于成对约束比类标号更易于获取，具有实用性，因此针对于这两种约束的半监督聚类研究得到广泛关注。如果两个数据点必须属于同一类，则它们之间存在 must-link 关系；如果两个点必须属于不同类，则它们之间是 cannot-link 关系。如果两个数据点之间存在 must-link 约束，那么好的特征应该使这两个点在该特征取值上相差不大；如果存在 connot-link 约束，则在该特征上两个点的取值相差较大。

基于上述思想的特征评估函数 Constraint Score[10]是：

$$C_r=\sum_{(x_i,x_j)\in\boldsymbol{M}}(f_{ri}-f_{rj})^2-\lambda\sum_{(x_i,x_j)\in\boldsymbol{C}}(f_{ri}-f_{rj})^2, \tag{2}$$

公式中，x_i 是数据点，f_{ri} 是 x_i 的第 r 个特征取值，$\boldsymbol{M}$ 和 $\boldsymbol{C}$ 分别是现有的成对约束条件，λ 的值一般取 0.1。由公式可知 C_r 值越小，说明第 r 个特征越好。

文献[10]提出该评估函数时是为了在数据空间选取特征（feature），本研究将 Constraint Score 用于评价谱空间内的特征向量（eigenvector），选取最优的特征向量组合，完成对测试数据集的预测。

2 特征向量的 Bagging 选取算法

Bagging（Bootstrap aggregating）策略是文献[14]提出的，旨在综合多种预测结果，给出最可信的预测。Bagging 策略有效的前提是各种预测结果之间的差异较大，随机性较强。在采用 Constraint Score 选取特征向量时，采用的成对约束条件不同，选出的特征向量也有较大差异，因此 Bagging 策略适用于特征向量的选择，可以消除成对约束条件不同对特征向量选取的影响。

设数据集 $X=\{x_1,x_2,\cdots,x_n\}$，每个 x_i 是 d 维空间内的点。计算相应的拉普拉斯矩阵 $\boldsymbol{L}$，对 $\boldsymbol{L}$ 进行特征值分解。根据文献[8]的结果，较大的特征值对应的特征向量往往是对聚类有帮助的，为简便起见，本研究只对 $\boldsymbol{L}$ 的 $2k$ 个最大的特征值对应的特征向量 $\boldsymbol{u}_1,\cdots,\boldsymbol{u}_{2k}$ 进行评价，从中找出较好的 k 个特征向量，评价的依据是训练数据集提供的部分成对约束信息，具体过程见图 1。

谱聚类中特征向量的 Bagging 选取方法（Eigenvector selection algorithm in spectral clustering based on Bagging，ESSCB）如下：

输入 数据集 $X=\{x_1,x_2,\cdots,x_n\}$，$r\%$ 的数据作为训练集，$(1-r\%)$ 的数据作为测试集；类别数 k。

输出 测试集的聚类结果：

（1）依据公式（1）计算亲和矩阵 $\boldsymbol{A}\in\boldsymbol{R}^{n\times n}$，$A_{ii}=0,1\leqslant i\leqslant n$；

（2）计算度矩阵 $\boldsymbol{D}$ 和拉普拉斯矩阵 $\boldsymbol{L}=\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{A}\boldsymbol{D}^{-\frac{1}{2}}$；

（3）求出 $\boldsymbol{L}$ 的所有特征向量 $\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n\}$，按特征值降序排列；

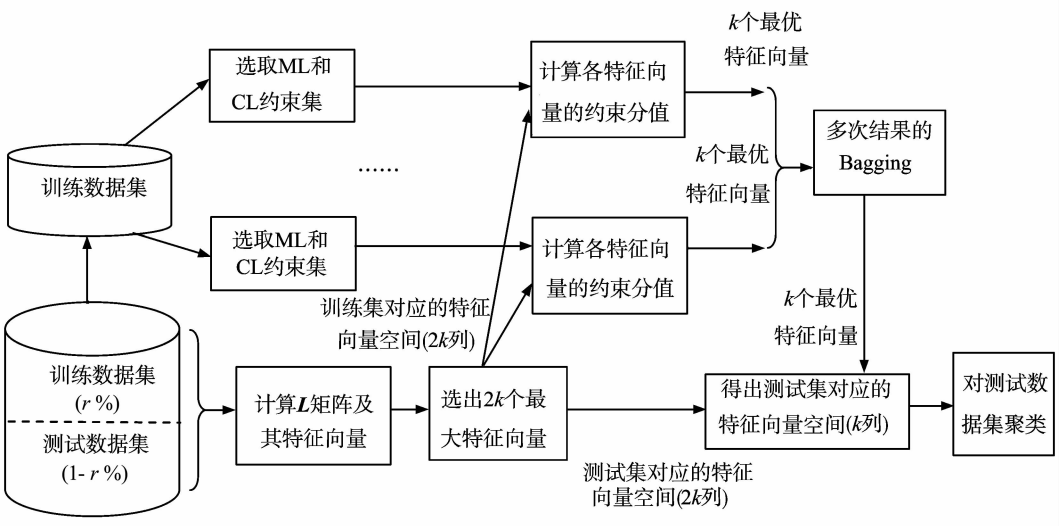

图 1　特征向量的 Bagging 选取方法

Fig. 1　Eigenvector selection based on Bagging

(4) 取出训练集的前 $2k$ 个特征向量组成特征向量矩阵 $\boldsymbol{U}=[v_1,v_2,\cdots,v_{2k}]$;

(5) 从训练集中随机选取 must-link 和 cannot-link 成对约束集 $\boldsymbol{M}$ 和 $\boldsymbol{C}$,两种约束的总数量是 cNum(如 10)个;

(6) 在 $\boldsymbol{M}$ 和 $\boldsymbol{C}$ 中成对约束下,使用特征向量评估函数 Constraint Score,对 $\boldsymbol{U}$ 中特征向量进行评估,选取使评估值 C_r 最小的前 k 个特征向量;

(7) 将(5)、(6)重复执行 m 次(如 200);

(8) 统计 m 次所选取的所有特征向量,记录出现次数最多的 k 个特征向量;

(9) 从测试集中取出这 k 个特征向量,组成矩阵 $\boldsymbol{V}$;

(10) 对每个矩阵 $\boldsymbol{V}$ 按行归一化(行中所有元素除以该行的模),$\boldsymbol{V}$ 中的每一行对应着测试集中的一个点;

(11) 对 $\boldsymbol{V}$ 进行 k-means 聚类,矩阵 $\boldsymbol{V}$ 的划分即是测试集的划分。

算法的时间复杂度分析如下:计算拉普拉斯矩阵 $\boldsymbol{L}$ 的复杂度是 $O(n^2)$,对该矩阵进行谱分解的复杂度是 $O(n^3)$,步骤(5)选取约束的时间复杂度是 $O(\mathrm{cNum})$,对 $2k$ 个特征向量的评价时间是 $O(\mathrm{cNum}*k)$,(5)-(6)循环 m 次的时间复杂度是 $O(\mathrm{cNum}*k*m)$,步骤(8)的时间复杂度是 $O(k*m)$,归一化数据集的时间复杂度是 $O(k*n)$,对 V 进行 k-means 聚类的时间复杂度是 $O(k*n*p)$,其中 p 是 k-means 的循环次数。综上,本研究的算法的瓶颈是拉普拉斯矩阵的特征向量分解,由于只需要前 $2k$ 个特征向量,因此可以考虑用 Power Method(幂方法)降低其时间复杂度至 $O(k*n^2)$[15-17]。试验中对较大数据集 Waveform,采用 Power Method 计算其前 $2k$ 个特征向量。

3　试验结果与分析

3.1　试验数据集

采用 UCI 基准数据集进行测试,数据集的基本信息见表 1。

表 1　数据集的基本信息

Table 1　Information of datasets

数据集名称	实例数量	属性数量	类别数
Iris	150	4	3
Wine	178	13	3
Ionosphere	351	34	2
Image Segmentation	2 310	19	7
Waveform	5 000	21	3

3.2　聚类评估指标

(1) 纯度

纯度定义为[18]

$$\mathrm{Purity}=\frac{1}{n}\sum_{r=1}^{k}\max_{i}(n_r^i),$$

其中 n 是测试数据数量,r 是簇号,n_r^i 是同时属于第 i 类和第 r 簇的数据数量。

纯度是最常用的聚类指标之一,纯度越大聚类结果越好。但是从其计算公式不难发现,纯度统计的是每个簇内占最多的类别的数据数量,而不考虑不同簇对应的类别是否相同,这样就会出现多个簇对应同一类别的情况,从而导致聚类结果退化。为此考虑精度指标。

(2) 精度

文献[7,19]计算精度为

$$\mathrm{Precision}=\frac{1}{n}\sum_{i=1}^{n}\delta(y_i,\ \mathrm{map}(c_i)),$$

公式中，n 是数据点的数量，y_i 和 c_i 分别是第 i 个数据点的原始类标号和聚类簇标号。当 $x=y$ 时 $\delta(x,y)=1$，否则为0。map()函数将簇标号映射为类标号，簇标号是类标号的排列，不会出现聚类结果退化。map()函数的选择应使精度达到最大。为此，采取以下方法[20]：

① 求出使 n_r^i 值最大的类号 i 和簇号 r，令 $\text{map}(r)=i$。类号 i 和簇号 r 不再参与计数比较。

② 回到舰，直到把所有簇中的样本点标上类号。

最后计算精度

$$\text{Precision}=\frac{m}{n},$$

其中，m 是标记的类号与原数据集中的类号相同的点数。

从精度的定义不难看出，精度要求簇号与类号一一对应，而纯度不做此要求，因此纯度为每个簇选择相应类别时可能找到更大的 n_r^i，导致纯度的计算值可能与精度相等，也可能比精度大。本研究中的实验也证实了这一点。

（3）熵

一个簇 S_r 的熵[18]：

$$E(S_r)=-\frac{1}{\log q}\sum_{i=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r},$$

其中 S_r 是第 r 簇，大小为 n_r，q 是数据集的类数量，n_r^i 定义同上。

各个簇熵的加权和即为聚类的整体熵

$$\text{Entropy}=\sum_{r=1}^{k}\frac{n_r}{n}E(S_r)。$$

由上述公式可知，熵越小，聚类效果越好。

（4）误差的平方和 SSE（Sum of Squared Error）

$$\text{SSE}=\sum_{i=1}^{k}\sum_{x\in C_i}\text{dist}^2(m_i,x),$$

其中，x 是簇 C_i 中的数据点，m_i 是簇 C_i 的中心点，$\text{dist}(x,y)$ 是点 x，y 的距离。由该公式可知，SSE 越小，簇内的点分布越紧凑，聚类效果越好。

3.3 实验方案与结果分析

本研究的算法 ESSCB 涉及的参数有3个：成对约束数量 cNum、局部尺度参数 σ 以及训练数据集的比例。训练数据集中每个数据点都有类标号，而两个数据点之间的成对约束信息可以通过比较类标号得出。成对约束的获取方法是：在训练数据集中随机选取2个数据点，比较它们的类标号。如果类标号相同，则这两个点之间是 must-link 约束，否则是 cannot－link 约束。按照这种方法逐个获取，就得到多个成对约束。将 ESSCB 算法与常规谱聚类算法进行对比，以证实能找到更好的特征向量组合。试验中进一步与文献[7]中提出的熵排序选择特征向量方法进行比较。试验中所需的参数 k 均设定为数据集的类别数。

试验方案1 测试 cNum 对聚类结果的影响

以 Image Segmentation 数据集为例，随机选取50%数据作为训练集，其余数据作为测试集。固定近邻点 $l=6$ 来计算局部尺度参数 $\sigma_i=d(x_i,x_{il})$（详见公式(1)），选取不同的约束对数 cNum，测试选出来的特征向量下测试数据集的聚类结果。如图2所示。图2列出了3种指标的变化情况，即 Purity，Precision 和 Entropy。

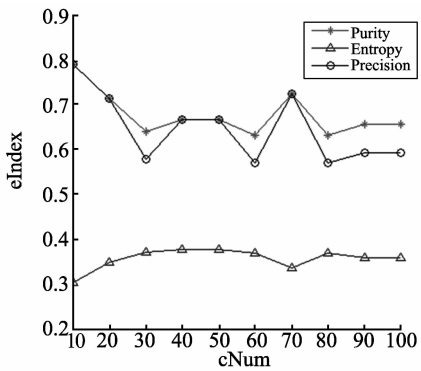

图2 约束数对聚类结果的影响

Fig. 2 Clustering result with different constraints

图2（cNum 是约束对数，eIndex 是评估指标）可以看出，在 cNum = 10 时，聚类效果较好，表现为熵较低，而纯度和精度较高，其中纯度高于精度的情况即是多个簇对应一个类别的情况。另外注意到还有次优解出现在成对约束数是70的位置，这个位置是不稳定的，当训练数据集改变时，有可能会改变。经过多次试验，cNum = 10 时聚类效果普遍好，因此选择 cNum = 10 作为合适的参数。分析其中的原因是：Bagging 策略有效的前提是各种预测结果之间的差异较大，随机性较强。在采用 Constraint Score 选取特征向量时，选择的成对约束数较小时，各组选择的成对约束重叠的可能性很小，采用的成对约束不同，选出的特征向量也有较大差异[19]，因此 Bagging 策略能得出较好的结果。

试验方案2 测试局部尺度参数对聚类结果的影响

仍以 Image Segmentation 数据集为例，随机选取50%数据作为训练集，其余数据作为测试集。在训练数据集中，随机选取 cNum = 10 对成对约束，测试不同的近邻点 l 确定的 σ 对聚类结果的影响。

从图 3 中(横轴 l 是近邻点数,纵轴 eIndex 是评估指标)可以看出,随着近邻点数 l 的增加,聚类结果大体出现下降趋势,$l=6$ 时聚类结果最好。当 l 变大时,$\sigma_i = d(x_i, x_{il})$ 变大,超过了其真实的局部尺度,度量有失偏颇。本研究中 $l=6$ 与文献[7]中取 $l=7$的设置基本一致。

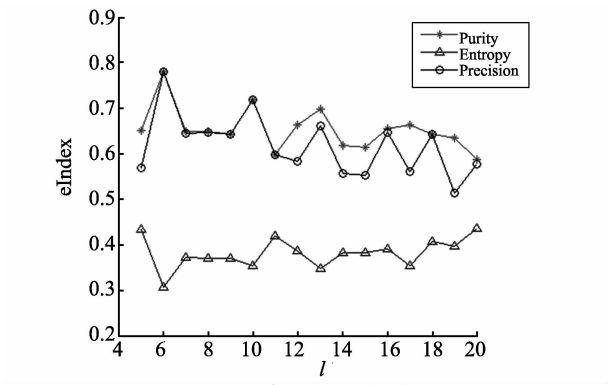

图 3　不同的近邻点对聚类结果的影响

Fig. 3　Clustering result with different nearest neighbor points

试验方案 3　训练数据集的比例对聚类结果的影响

仍以 Image Segmentation 数据集为例。在训练数据集中,随机选取 cNum = 10 对成对约束,固定近邻点 $l=6$ 来确定 σ。训练数据集所占比例从 10% ~90% 变化时,测试对其余数据聚类结果的影响。

从图 4 中(ratio 是训练数据的比例,eIndex 是评估指标)可以看出,随着训练数据集的不同,测试数据集的聚类结果有一定变化,但是仍有大体持平的趋势。当成对约束数较小而成对约束的选择范围较大时,采用 Bagging 策略后得到的聚类结果并没有太大变化。因此,本研究中实验均以训练数据集占 50% 的比例进行的,可以看作是一般情况下的结果。

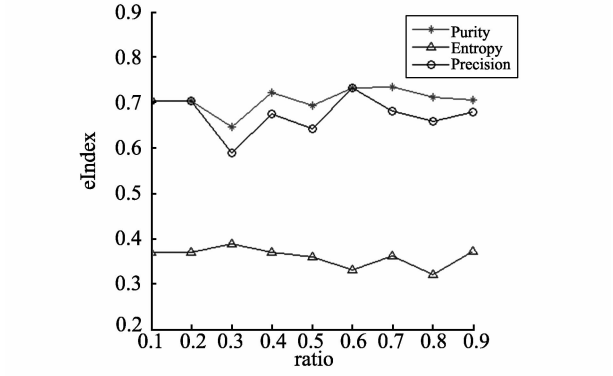

图 4　训练数据比例对聚类结果的影响

Fig. 4　Clustering result with different train datasets

试验方案 4　3 种算法对比

各种算法的对比见表 2。各数据集在 $l=6$ 近邻点,约束数 cNum = 10,训练数据集占 50% 的情况下,10 遍运行结果平均值。TopK 表示常规谱聚类算法,即选出的特征向量是特征值最大的 k 个特征向量。Ent 表示文献[7]中的基于熵的直接特征向量选择算法,即不进行特征向量的组合,而是直接选出熵最大的 k 个特征向量。计算熵的时候使用了训练数据集,而选出的特征向量用于测试数据集的聚类,不是全体数据的聚类。ES 表示各算法选出的特征向量最佳组合,[1,2,3]表示特征值第 1、2、3 大的特征值对应的特征向量,其余类似。

表 2　3 种算法的聚类结果对比

Table 2　Comparison of three algorithms

指标	算法	Iris	Wine	Ionosphere	Image Segmentation	Waveform
Purity	ESSCB	0. 933 3	0. 958 4	0. 696 6	0. 730 6	0. 751 3
	TopK	0. 922 7	0. 958 4	0. 740 9	0. 700 3	0. 523 2
	Ent	0. 806 7	0. 896 6	0. 690 3	0. 655 0	0. 556 1
Precision	ESSCB	0. 933 3	0. 958 4	0. 661 9	0. 719 9	0. 735 8
	TopK	0. 922 7	0. 958 4	0. 740 9	0. 690 2	0. 368 2
	Ent	0. 804 0	0. 888 8	0. 623 3	0. 640 5	0. 540 9
Entropy	ESSCB	0. 194 3	0. 143 3	0. 656 1	0. 335 3	0. 569 9
	TopK	0. 209 8	0. 143 3	0. 703 9	0. 349 6	0. 637 4
	Ent	0. 327 7	0. 233 8	0. 707 6	0. 399 6	0. 776 1
SSE	ESSCB	15. 559 0	6. 174 9	13. 174 8	201. 346 4	402. 067 0
	TopK	6. 251 0	6. 174 9	20. 488 0	114. 840 2	251. 061 5
	Ent	22. 446 7	33. 603 5	32. 027 6	209. 110 3	1 075. 726 3
ES	ESSCB	[1,2,3],[2,3,4]	[1,2,3]	[1,3]	[2,3,5,7,8,9,12]	[1,2,4]
	TopK	[1,2,3]	[1,2,3]	[1,2]	[1,2,3,4,5,6,7]	[1,2,3]
	Ent	[2,3,5]	[2,3,4]	[4,5]	[2,4,5,6,8,9,10]	[3,5,6]

从表2可以看出,对于数据集Iris,本研究中算法ESSCB选出的特征向量组合[1,2,3]、[2,3,4]在10遍试验结果中出现次数相同,因此表中结果是两种特征向量组合的平均值。Wine数据集上的试验发现,ESSCB和TopK所有指标(包括ES)的值都相同,这是比较少见的现象。ESSCB算法对其余数据集的计算结果与TopK不同,除Ionosphere外,ESSCB的聚类纯度、精度、熵指标都等于或好于TopK,这也说明ESSCB是有效的。而基于熵的算法Ent聚类结果各项指标除个别情况外都不好,这也是为什么文献[7]在该算法基础上进一步采用间接方法,对选出的特征向量进行组合的原因。

4 讨论

本研究中算法虽然采用了训练数据集中的点来构造成对约束,但并没有把成对约束信息反映到亲和矩阵中,因此并不是半监督的谱聚类算法。文献[10]采用约束计分方法来选择特征(feature),而本研究将数据集转换到特征向量空间,用约束计分的方法来选择特征向量(eigenvector),并且在多次选出的特征向量组合之间采用Bagging策略选出最终的特征向量组合。文献[19]也采用了Bagging策略将每次聚类的簇标号结果进行集成,而本研究的目的在于选出较好的一组特征向量,这不同于对多组特征向量的预测结果的集成。与文献[7-8]的另一个差别在于,本研究恰好选择 k 个特征向量,而不再尝试少于 k 或多于 k 的各种组合情况。这样做的原因是:一般情况下随着特征向量的增多,聚类精度会变大[8]。特征向量数量不同时进行聚类精度的比较,失去了一个公平比较的基准,因此本研究选择恰好 k 个特征向量。

5 结论

本研究主要关注谱聚类中特征向量的选择问题,借助成对约束计分的方法,对前 $2k$ 特征向量的重要性进行评价,并对选出的 k 个较好特征向量进行Bagging集成,得到 k 个最优特征向量。最后采用纯度(purity)、精度(precision)、熵(entropy)等指标对试验结果进行评价,试验证实该方法能取得较好结果。对于较大的数据集,采用了Power Method算法计算,降低了算法的时间复杂度。对于多次选出的特征向量组合如何进行更好的集成,比如对每个组合先进行评价,然后选出较好的组合再集成等选择集成方法是下一步工作。

参考文献:

[1] 张向荣,骞晓雪,焦李成. 基于免疫谱聚类的图像分割[J]. 软件学报, 2010, 21(9):2196-2205.
ZHANG Xiangrong, QIAN Xiaoxue, JIAO Licheng. Immune spectral clustering algorithm for image segmentation [J]. Journal of Software, 2010, 21(9):2196-2205.

[2] 杨宁,唐常杰,王悦,等. 基于谱聚类的多数据流演化事件挖掘[J]. 软件学报, 2010, 21(10):2395-2409.
YANG Ning, TANG Changjie, WANG Yue, et al. Mining evolutionary events from multi-Streams based on spectral clustering[J]. Journal of Software, 2010, 21(10): 2395-2409.

[3] NIE Feiping, ZENG Zinan, TSANG I W, et al. Spectral embedded clustering: a framework for in-sample and out-of-sample spectral clustering [J]. IEEE Transactions on Neural Networks, 2011, 22(11):1796-1808.

[4] SIDI O, KAICK O V, KLEINANAN Y, et al. Unsupervised co-segmentation of a set of shapes via descriptor-space spectral clustering [J]. ACM Transacti-ons on Graphics, 2011, 30(6):126-135.

[5] 姜大庆,夏士雄,周勇. 基于半监督自动谱聚类算法的网络故障检测[J]. 计算机工程与应用,2012,48(30):89-94.
JIANG Daqing, XIA Shixiong, ZHOU Yong. Network fault detection based on semi-supervised automatic spectral clustering algorithm[J]. Computer Engineering and Applications, 2012, 48(30):89-94.

[6] XIANG Tao, GONG Shaogang. Spectral clustering with eigenvector selection[J]. Pattern Recognition, 2008, 1: 1012-1029.

[7] ZHAO Feng, JIAO Licheng, LIU Hanqiang, et al. Spectral clustering with eigenvector selection based on entropy ranking[J]. Neurocomputing, 2010, 73:1704-1717.

[8] RELAGLIATI N, VERRI A. Spectral clustering with more than K eigenvectors[J]. Neurocomputing, 2011, 74:1391-1401.

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

[10] ZHANG Daoqiang, CHEN Songcan, ZHOU Zhihua. Constraint score: a new filter method for feature selection with pairwise constraints[J]. Pattern Recognition, 2008, 41:1440-1451.

[11] BISHOP C M. Neural networks for pattern recognition [M]. Oxford: Oxford University Press, 1995:41-42, 105-112.

[12] HE X, CAI D, NIYOGI P. Laplacian score for feature selection[C]//Advances in Neural Information Process-

ing Systems. Cambridge: MIT Press, 2005, 18: 507-514.

[13] WAGSTAFF K, CARDIE C. Clustering with Instance-level constraints [C]//Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000). San Francisco: Morgan Kaufmann Publishers Inc, 2000: 1103-1110.

[14] BREIMAN L. Bagging predictors [J]. Machine Learning, 1996, 24 (2): 123-140.

[15] MAVROEIDIS D, Accelerating spectral clustering with partial supervision [J]. Data Mining and Knowledge Discovery, 2010, 21: 241-258.

[16] JOURNEE M, NESTEROV Y, RICHTARIK P, et al. Generalized power method for sparse principal component analysis [J]. Journal of Machine Learning Research, 2010, 11: 517-553.

[17] GOLUB G H, LOAN C F V. Matrix computations [M]. Baltimore, USA: John Hopkins University Press, 1996.

[18] ZHAO Ying, KARYPIS G. Criterion functions for document clustering: experiments and analysis [R]. Minneapolis, USA: Computer Science Department, University of Minnesota: 2001.

[19] SUN Dan, ZHANG Daoqiang. Bagging constraint score for feature selection with pairwise constraints [J]. Pattern Recognition, 2010, 43: 2106-2118.

[20] 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 Zhihua. Bagging-based selective clusterer ensemble [J]. Journal of Software, 2005, 16 (4): 496-502.

(编辑:陈燕)

(上接第 34 页)

[15] DASGUPTA S, LITTMAN L M, MCALLESTER A D. PAC generalization bounds for co-training [C]//Advances in Neural Information Processing Systems 14. Cambridge, MA: MIT Press, 2002: 375-382.

[16] PIERCE D, CARDIE C. Limitations of co-training for natural language learning from large datasets [C]//Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2001: 1-9.

[17] BARTLETT S M, MOVELLAN R J, SEJNOWSKI J T. Face recognition by independent component analysis [J]. IEEE Transactions on Neural Networks, 2002, 13 (6): 1450-1464.

[18] 董治强, 刘琚, 邹欣, 等. 基于 ICA 的语音信号表征和特征提取方法[J]. 山东大学学报: 工学版, 2010, 40(4): 19-22.
DONG Zhiqiang, LIU Ju, ZOU Xin, et al. Speech signal representation and feature extraction based on ICA [J]. Journal of Shandong University: Engineering Science, 2010, 40(4): 19-22.

[19] OJA E. Original contribution: principal components, minor components, and linear neural networks [J]. Neural Networks, 1992, 5(6): 927-935.

[20] OJA E. The nonlinear PCA learning rule in independent component analysis [J]. Neurocomputing, 1997, 17 (1): 25-45.

[21] LEE T W. Independent component analysis: theory and applications [M]. Boston: Kluwer Academic Publishers, 1998.

[22] 孙浩军, 杜育林, 姜大志. 基于信息熵的高维分类型数据子空间聚类算法[J]. 山东大学学报: 工学版, 2011, 41(5): 37-45.
SUN Haojun, DU Yulin, JIANG Dazhi. ESCHCD: entropy-based algorithm for subspace clustering with high dimensional categorical datasets [J]. Journal of Shandong University: Engineering Science, 2011, 41 (5): 37-45.

[23] 杨竹青, 李勇, 胡德文. 独立成分分析方法综述 [J]. 自动化学报, 2002, 28(5): 762-772.
YANG Zhuqing, LI Yong, HU Dewen. Independent component analysis: a survey[J]. Acta Automatica Sinica, 2002, 28(5): 762-772.

[24] HYVÄRINEN A, OJA E. A fast fixed-point algorithm for independent component analysis [J]. Neural Computation, 1997, 9(7): 1483-1492.

[25] HYVÄRINEN A. Fast and robust fixed-point algorithms for independent component analysis [J]. IEEE Transactions on Neural Networks, 1999, 10(3): 626-634.

[26] ZHANG Minling, ZHOU Zhihua. CoTrade: confident co-training with data editing [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 2011, 41(6): 1612-1626.

(编辑:胡春霞)