一种基于 T-分布随机近邻嵌入的聚类集成方法

徐 森* 花小朋 徐 静 徐秀芳 皋 军 安 晶
(盐城工学院信息工程学院 盐城 224051)

摘 要：该文将 T-分布随机近邻嵌入(TSNE)引入到聚类集成问题中，提出一种基于 TSNE 的聚类集成方法。首先通过 TSNE 最小化超图邻接矩阵的行对应的高维数据点与低维映射点分布之间的 KL 散度，使得高维空间结构在低维空间得以保持，然后在低维空间运行层次聚类算法获得最终的聚类结果。在基准数据集上的实验结果表明：TSNE 能够提高层次聚类算法的聚类质量，该文方法获得了优于主流聚类集成方法的结果。

关键词：机器学习；聚类分析；聚类集成；层次聚类
中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2018)06-1316-07
DOI: 10.11999/JEIT170937

Cluster Ensemble Approach Based on T-distributed Stochastic Neighbor Embedding

XU Sen HUA Xiaopeng XU Jing XU Xiufang GAO Jun AN Jing
(*School of Information Engineering*, *Yancheng Institute of Technology*, *Yancheng* 224051, *China*)

Abstract: T-distributed Stochastic Neighbor Embedding (TSNE) is introduced into cluster ensemble problem and a cluster ensemble approach based on TSNE is proposed. First, TSNE is utilized to minimize Kullback-Leibler divergences between the high-dimensinal points corresponding to the rows of hypergraph's adjacent matrix and the low-dimensional mapping points, which preserves the structure of high-dimensional space in low-dimensional space. Then, a hierarchical clustering algorithm is carried out in the low-dimensional space to obtain the final clustering result. Experimental results on several baseline datasets indicate that TSNE can improve the cluster results of hierarchical clustering algorithm and the proposed cluster ensemble method via TSNE outperforms state-of-the-art methods.

Key words: Machine learning; Clustering analysis; Cluster ensemble; Hierarchical clustering

1 引言

聚类分析是模式识别、机器学习等方向的重要研究内容之一，在图像分割、复杂网络分析等领域获得了广泛应用，其目标是将数据集划分为若干簇，使得簇内对象相似度尽量高，簇间对象相似度尽量低[1–3]。在实际应用中，聚类算法通常需要采用相似性度量和聚类准则，其中潜含着对数据中包含的簇结构的某种假设。当这些假设与数据的真实分布不相符时，会产生错误或没有意义的结果。传统的聚类算法层出不穷，但没有一种算法能够有效识别所有类型的聚类结构[1,2]。面对诸多聚类算法，研究者们要做出明智的选择，需要完全理解特定的聚类算法，并了解数据的领域知识，显然这是不切实际的。

聚类集成通过组合多个不同的聚类结果(每个聚类结果称为聚类成员)能够获得比单一聚类算法更加优越的聚类结果[4]。目前，聚类集成已经发展成为传统聚类算法的重要扩展，吸引了国内外众多学者[4–12]，其关键问题在于如何将聚类成员组合为更加优越的聚类结果，一般称为共识函数设计问题/聚类集成问题。引起聚类集成问题困难的原因有两个：(1)聚类学习中对象是无标签的，所以不同聚类成员得到的簇标签没有显式的对应关系；(2)聚类成员可能包含不同的簇个数，这使得簇标签对应问题更加困难[4]。为使问题简化，许多研究者都将真实类别个数作为基聚类算法的输入，本文也沿用该方法。解决聚类集成问题最常见的方法是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，从而避免

收稿日期：2017-10-10；改回日期：2018-03-16；网络出版：2018-04-10
*通信作者：徐森 xusen@ycit.cn
基金项目：国家自然科学基金(61105057, 61375001)，江苏省自然科学基金(BK20151299)，江苏省产学研前瞻性联合研究项目(BY2016065-01)
Foundation Items: The National Natural Science Foundation of China (61105057, 61375001), The Natural Science Foundation of Jiangsu Province (BK20151299), The Industry-Education-Research Prospective Project of Jiangsu Province (BY2016065-01)

簇标签对应问题。根据处理的矩阵不同，可以分为：对 $\boldsymbol{H}$ 或其加权矩阵进行处理，包括超图划分算法(Hyper Graph Partitioning Algorithm, HGPA)[4]和元聚类算法(Meta-CLustering Algorithm, MCLA)[4]、基于 K 均值(K-Means, KM)的方法[7]等；对相似度矩阵 $\boldsymbol{S}=\boldsymbol{H}\times\boldsymbol{H}^{\mathrm{T}}/r$ 进行处理，包括基于簇的相似度划分算法(Cluster-based Similarity Partitioning Algorithm, CSPA)[4]、基于证据积累(Evidence Accumulation)的层次聚类方法[8]、基于谱聚类的方法[9]、基于近邻传播(Affinity Propagation)的方法[10]、基于密度峰值(Density Peaks)的方法[11]、加权共现矩阵(Weighted Co-association Matrices)[12]方法等。

与已有聚类集成方法不同，本文将 T-分布随机近邻嵌入(T-distributed Stochastic Neighbor Embedding, TSNE)[13−15]这一流行的降维技术引入到聚类集成问题中，首先通过 TSNE 最小化超图邻接矩阵的行对应的高维数据点与低维映射点分布之间的 KL 散度(Kullback-Leibler divergences)，使得高维空间结构在低维空间得以保持，然后再进行层次聚类，设计了一种基于 TSNE 的聚类集成方法。在多组基准数据集上的实验结果表明：TSNE 能够提高层次聚类算法的聚类质量，本文方法获得了优于主流聚类集成方法的结果。

2　本文方法

2.1 问题求解

设 $\boldsymbol{D}=\{\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_n\}$ 为数据集，假设 r 个聚类成员构成聚类集体 $\boldsymbol{M}=\left\{\boldsymbol{M}^{(1)},\boldsymbol{M}^{(2)},\cdots,\boldsymbol{M}^{(r)}\right\}$，聚类成员 $\boldsymbol{M}^{(i)}$ 包含 k 个簇 $\boldsymbol{C}^i=\left\{\boldsymbol{C}_1^i,\boldsymbol{C}_2^i,\cdots,\boldsymbol{C}_k^i\right\}$，$i=1,2,\cdots,r$。设 $\boldsymbol{H}=\left\{\boldsymbol{H}^{(1)},\boldsymbol{H}^{(2)},\cdots,\boldsymbol{H}^{(r)}\right\}$ 为超图的邻接矩阵，该超图有 n 个顶点，$s=r\times k$ 条超边，$\boldsymbol{H}^{(i)}$ 为聚类成员 $\boldsymbol{M}^{(i)}$ 对应的超图的邻接矩阵，$\boldsymbol{H}^{(i)}$ 的每一列分别包含 $n_1^i,n_2^i,\cdots,n_k^i$ 个 1，n_j^i 表示簇 $\boldsymbol{C}_j^i$ 的大小，其余元素为 0。若将 $\boldsymbol{H}$ 的每一列看成是一个特征，则可根据行向量之间的欧氏距离计算对象之间的相似度，行向量之间的距离越远则所对应的数据点之间的相似度越低，反之相似度越高。设 $\boldsymbol{H}$ 的行对应的高维数据点为 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，本文引入 TSNE[13]将高维数据点在低维空间用映射点表示，使得高维空间结构在低维空间得以保持，即距离较大的高维数据点在低维空间下的映射点距离也较大，反之较小。设低维映射点为 $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}$，则 $\boldsymbol{x}_i,\boldsymbol{x}_j$ 对应的低维映射点的联合概率为

$$P_{ij}=\begin{cases}\dfrac{\exp\left(-\delta_{ij}^2/\sigma\right)}{\sum\limits_k\sum\limits_{i\neq k}\exp\left(-\delta_{ij}^2/\sigma\right)}, & i\neq j\\ 0, & i=j\end{cases}\tag{1}$$

其中，$\delta_{ij}^2=\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|^2$ 表示 $\boldsymbol{x}_i,\boldsymbol{x}_j$ 之间的距离。联合概率可通过式(2)求得：

$$P_{ij}=\left(P_{j|i}+P_{i|j}\right)\Big/(2n)\tag{2}$$

其中，$P_{j|i}$ 表示条件概率。

$$P_{j|i}=\begin{cases}\dfrac{\exp\left(-\left\|\boldsymbol{x}_i-\boldsymbol{x}_j\right\|^2\Big/2\sigma_i^2\right)}{\sum\limits_{k\neq i}\exp\left(-\left\|\boldsymbol{x}_i-\boldsymbol{x}_k\right\|^2\Big/2\sigma_i^2\right)}, & i\neq j\\ 0, & i=j\end{cases}\tag{3}$$

其中，σ_i 为中心在 $\boldsymbol{x}_i$ 的高斯分布的方差。通过预先设定的复杂度因子(perplexity)执行二元搜索，可求得能生成 P_i 的 σ_i，复杂度因子 $\mathrm{Perp}\left(P_i\right)=2^{H(P_i)}$，其中 $H\left(P_i\right)$ 为分布 P_i 的熵：

$$H\left(P_i\right)=-\sum_j P_{j|i}\log_2 P_{j|i}$$

通过 1 个自由度的 T-分布计算低维映射点 $\boldsymbol{y}_i$，$\boldsymbol{y}_j$ 之间的相似度，使相似度较低(即距离较大)的高维数据点在低维映射空间下的距离较大。

$$Q_{ij}=\begin{cases}\dfrac{\left(1+\left\|\boldsymbol{y}_i-\boldsymbol{y}_j\right\|^2\right)^{-1}}{\sum\limits_k\sum\limits_{l\neq k}\left(1+\left\|\boldsymbol{y}_k-\boldsymbol{y}_l\right\|^2\right)^{-1}}, & i\neq j\\ 0, & i=j\end{cases}\tag{4}$$

在理想情况下，当低维映射点 $\boldsymbol{y}_i$ 和 $\boldsymbol{y}_j$ 正确建模高维数据点 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的相似度时，联合概率 $Q_{ij}=P_{ij}$。在一般情况下，联合概率 Q_{ij} 与 P_{ij} 之间存在误差，为了最小化联合概率 Q_{ij} 到 P_{ij} 的差异，通过 KL 散度(Kullback-Leibler divergences)衡量两个分布 $\boldsymbol{P}$，$\boldsymbol{Q}$ 的差异，得到式(5)的代价函数：

$$\begin{aligned}C(\boldsymbol{Y})=\mathrm{KL}(\boldsymbol{P}\|\boldsymbol{Q})&=\sum_i\sum_{j\neq i}P_{ij}\log\frac{P_{ij}}{Q_{ij}}\\ &=\sum_i\sum_{j\neq i}\frac{\exp\left(-\delta_{ij}^2/\sigma\right)}{\sum\limits_k\sum\limits_{i\neq k}\exp\left(-\delta_{ij}^2/\sigma\right)}\\ &\quad\cdot\log\frac{\dfrac{\exp\left(-\delta_{ij}^2/\sigma\right)}{\sum\limits_k\sum\limits_{i\neq k}\exp\left(-\delta_{ij}^2/\sigma\right)}}{\dfrac{\left(1+\left\|\boldsymbol{y}_i-\boldsymbol{y}_j\right\|^2\right)^{-1}}{\sum\limits_k\sum\limits_{l\neq k}\left(1+\left\|\boldsymbol{y}_k-\boldsymbol{y}_l\right\|^2\right)^{-1}}}\end{aligned}\tag{5}$$

式(5)的梯度为

$$\frac{\delta C}{\delta \boldsymbol{y}_i}=4\sum_j\left(p_{ij}-q_{ij}\right)\left(\boldsymbol{y}_i-\boldsymbol{y}_j\right)\left(1+\left\|\boldsymbol{y}_i-\boldsymbol{y}_j\right\|\right)^{-1} \quad (6)$$

TSNE 采用梯度下降法最小化代价函数，通过从以原点为中心点，具有较小方差的等高斯分布随机采样低维映射点进行初始化。为了加速优化过程，避免陷入较差的局部最小值，在梯度中加入一个相对大的动量项。具体地，在梯度搜索的每次迭代中，为了确定映射点坐标变化，当前的梯度被加到上一步梯度的指数衰减和。带动量项的梯度更新规则为

$$\boldsymbol{Y}^{(t)}=\boldsymbol{Y}^{(t-1)}+\eta\frac{\delta C}{\delta \boldsymbol{Y}}+\alpha(t)\left(\boldsymbol{Y}^{(t-1)}-\boldsymbol{Y}^{(t-2)}\right) \quad (7)$$

其中，$\boldsymbol{Y}^{(t)}$ 表示第 t 次迭代的解，η 表示学习率，$\alpha(t)$ 表示第 t 次迭代的动量项。

2.2 算法设计

在获得 $\boldsymbol{H}$ 的行所对应的数据点的低维表示后，下一步是如何获得最终的聚类结果。KM 算法简单高效，但仅适于识别球型簇，且聚类结果受初始质心影响较大。层次聚类方法能够获得不同粒度的多层次结构，且无局部极值和初值选取问题。总体来看，平均链接(Average Linkage, AL)算法往往能够获得优于单链接(Single Linkage, SL)和全链接(Complete Linkage, CL)的聚类结果。由于簇的大小未知，本文采用加权平均链接算法(Weighted Average Linkage, WAL)进行聚类得到最终的结果，不妨将本文提出的 TSNE 结合 WAL 的聚类集成算法记为 TSNE+WAL，其主要步骤如表 1 所示，其中梯度迭代次数 T 一般设为 1000；当迭代次数 $t<250$ 时，动量项 $\alpha(t)=0.5$，当 $t\ge 250$ 时，$\alpha(t)=0.8$；学习率 η 初值为 100，每次迭代结束根据自适应学习率机制进行更新；维数 no_dims 一般设置为 2 或 3，本文实验中设置为 2。

TSNE+WAL 算法第 1 步生成聚类成员，时间复杂度为 $O(rkmn)$，第 2 步时间复杂度为 $O(rkn)$，第 3 步需要求解 n 个对象之间的距离，时间复杂度为 $O\left(rkn^2\right)$，第 4 步计算复杂度为 $O(rkn)$，第 5~6 步计算复杂度为 $O\left(Trkn^2\right)$，第 7 步时间复杂度为 $O(kn)$，因为 r，k 和 T 都为常数，所以 TSNE+WAL 算法的时间复杂度为 $O\left(mn+n^2\right)$。对于高维海量的数据集，n 和 m 都是数以万计的，此时 TSNE+WAL 算法第 1 步可以采用聚类工具箱 CLUTO (http://glaros.dtc.umn.edu/gkhome)高效实现。

3 实验

3.1 实验数据集和评价标准

实验中共使用了 8 组基准文本数据集，其中 7 组由 TREC(Text REtrieval Conference，文本检索大会，http://trec.nist.gov)提供，每组数据集中的文本类别标签唯一：数据集 tr11 包含 9 类，共 414 个文本，6429 个特征词；数据集 tr23 包含 6 类，共 204 个文本，5832 个特征词；数据集 tr41 包含 10 类，共 878 个文本，7454 个特征词；数据集 tr45 包含 10 类，共 690 个文本，8261 个特征词；数据集 hitech 包含 6 类，共 2301 个文本，10080 个特征词；数据集 reviews 包含 5 类，共 4069 个文本，18483 个特征词；数据集 sports 包含 7 类，共 8580 个文本，14870 个特征词。数据集 ng3 为 20 Newsgroups(http://qwone.com/~jason/20Newsgroups/)的子集，包含 3 类，每类约 1000 个文本，共 2998 个文本，15810 个特征词。

表 1 TSNE+WAL 聚类集成算法

输入：对象集 $\boldsymbol{D}=\{\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_n\}\in R^{n\times m}$（$n$ 为对象个数，m 为特征个数），簇个数 k，聚类成员个数 r，复杂度因子 Perp，迭代次数 T，学习率 η，动量 $\alpha(t)$，维数 no_dims。

(1)运行 K 均值算法 r 次，每次随机选取初始质心，生成 r 个聚类成员；

(2)构建超图的邻接矩阵 $\boldsymbol{H}$，将 $\boldsymbol{H}$ 的每一行看成一个对象，得到数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$；

(3)分别根据式(2)和式(3)计算 P_{ij} 和 $P_{j|i}$；

(4)初始解 $\boldsymbol{Y}^{(0)}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}$ 采样于正态分布 $\mathrm{N}\left(0,10^{-4}\mathrm{I}\right)$；

(5)for t=1 to T

(a)根据式(4)计算 Q_{ij}；

(b)根据式(6)计算梯度；

(c)根据式(7)计算 $\boldsymbol{Y}^{(t)}$；

end for

(6)得到低维数据表示 $\boldsymbol{Y}^{(\mathrm{T})}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_n\}$；

(7)使用 WAL 算法将 $\boldsymbol{Y}^{(\mathrm{T})}$ 划分为 k 个簇 $\boldsymbol{C}_1,\boldsymbol{C}_2,\cdots,\boldsymbol{C}_k$。

输出：聚类结果 $\pi=\{\boldsymbol{D}_1,\boldsymbol{D}_2,\cdots,\boldsymbol{D}_k$，其中 $\boldsymbol{D}_j=\{\boldsymbol{d}_i|\boldsymbol{y}_i\in\boldsymbol{C}_j,\ \boldsymbol{d}_i\in\boldsymbol{D},j=1,2,\cdots,k\}\}$。

因为文本的真实类别标签已知，本文采用外部指标规范化互信息(NMI)值、F 值(F-measure)和 AR 值(Adjusted Rand index)对算法进行综合评价。NMI 值是机器学习领域比较流行的评价指标，能够有效衡量聚类结果与真实类别标签的匹配程度。NMI 值越大，聚类结果和真实类别标签越匹配，当聚类结果和真实类别标签一一对应时，NMI 值达到最大值，1。F 值是信息检索与自然语言处理领域一个常用的评价文本分类和文本聚类算法的综合指标。F 值越大，文本聚类质量越高，反之越低。AR 值是常见的评价机器学习算法的外部指标之一，AR 值越大，聚类结果与真实类别标签越一致。

3.2 实验设计

首先对经过预处理的文本数据集进行 TF-IDF (Term Frequency-Inverse Document Frequency)加权[16]，然后运行使用余弦相似度的 KM 算法 100 次，每次生成 k 个簇(k 等于真实类别个数)，生成 100 个聚类成员，然后运行不同的聚类集成算法。

实验包含两个部分，第 1 部分将本文算法与主流聚类集成算法进行对比，并对比直接进行 WAL 聚类集成与 TSNE+WAL 的结果。对比算法包括文献[4]提出的 CSPA, HGPA, MCLA，文献[8]提出的层次聚类算法 SL, CL, AL 和 Ward(记为 WL)，文献[7]提出的基于 KM 的算法。考虑到图划分算法中 CSPA 总体聚类效果最好，而层次聚类算法中 SL 易于产生链式效应，聚类质量较差，因此本文仅将 TSNE+WAL 与 CSPA, CL, AL, WL 和 KM 进行比较。第 2 部分比较聚类集体大小 r 变化时每种聚类集成算法获得的结果。本文实验中 CSPA 调用了图划分算法 METIS，不平衡因子 UB 取默认值 0.05，得到了稳定的聚类结果。层次聚类方法 CL, AL, WAL, WL 递归地合并对象，将数据集划分为嵌套的层次结构，获得了稳定的聚类结果。KM 算法获得的聚类结果不稳定，运行 10 次取最优结果。文献[13]指出，TSNE 算法的性能对复杂度因子 Perp 的变化较为鲁棒。根据 Perp 的定义，若 Perp 值太小，则每个低维嵌入点都成为孤立点，不能形成聚类结构；若 Perp 值太大，则所有低维嵌入点形成一个簇；合理的 Perp 值既能使相同簇内的点保持紧凑，又能使不同簇内的点更加远离。本文通过多组实验比较发现，对于不同数据集，Perp 的合理取值并不相同，本文设置 Perp 的依据是代价函数 $\mathrm{KL}(\boldsymbol{P}\,||\,\boldsymbol{Q})$ 的值在 0.01 ~0.20 之间。具体地，在数据集 tr11, tr23, tr41 和 tr45 上 Perp 设置为 50；在数据集 hitech 上 Perp 设置为 400；在数据集 reviews 和 sports 上 Perp 设置为 1200；在数据集 ng3 上 Perp 设置为 800。

3.3 实验结果

3.3.1 与主流聚类集成算法进行对比　图 1 显示了不同聚类集成算法获得的 NMI 值，F 值，AR 值及这 3 个评价指标值的平均值(Mean)。由图 1 可以发现：(1)在所有数据集上，TSNE+WAL 都获得了最高 NMI 值和 AR 值；除了在数据集 tr23 上获得了

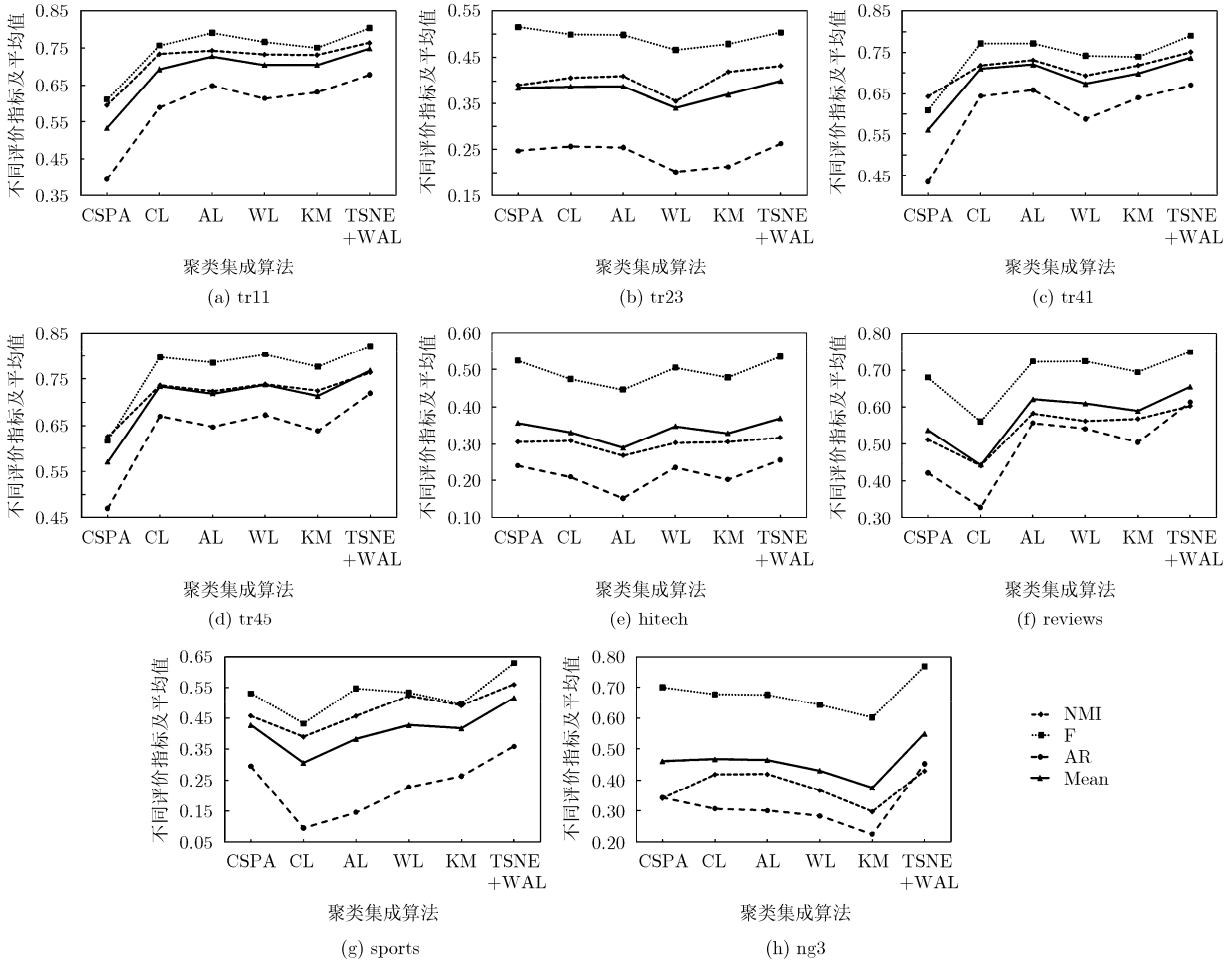

图 1 不同聚类集成算法获得的 NMI 值、F 值、AR 值及这 3 个评价指标值的平均值(Mean)

比 CSPA 略低的 F 值外，在其他数据集上都获得了最高的 F 值。(2)总体来看，TSNE+WAL 在所有数据集上都获得了最高平均值，验证了本文算法的优越性。

WAL 与 TSNE+WAL 在不同数据集上获得的结果如表 2 所示，其中“比值”表示 TSNE+WAL 获得的指标值与 WAL 的相应指标值的比值。若比值大于 1，则本文算法优于 WAL，且比值越大本文算法的优越程度越高；若比值小于 1，则 WAL 优于本文算法，且比值越大，WAL 的优越程度越高。由表 2 可见，TSNE+WAL 在所有数据集上获得的 NMI 值，F 值和 AR 值都高于 WAL(比值都大于 1)，比值在不同数据集上有高有低，特别地，在数据集 sports 上，比值高达 2.459，即 AR 值提升了 145.9%。这一实验结果表明 TSNE 能够在不同程度上提高 WAL 的聚类质量，因此，将 TSNE 引入聚类集成问题中是改善聚类集成效果的一种有效手段。

表 2 WAL 和 TSNE+WAL 在不同数据集上获得的结果

数据集	评价指标	WAL	TSNE+WAL	比值
tr11	NMI	0.748	0.765	1.022
	F	0.757	0.804	1.062
	AR	0.523	0.679	1.298
tr23	NMI	0.326	0.432	1.324
	F	0.448	0.504	1.125
	AR	0.176	0.263	1.497
tr41	NMI	0.741	0.751	1.013
	F	0.768	0.790	1.029
	AR	0.656	0.671	1.023
tr45	NMI	0.751	0.765	1.018
	F	0.805	0.822	1.021
	AR	0.692	0.719	1.040
hitech	NMI	0.282	0.317	1.124
	F	0.464	0.535	1.153
	AR	0.210	0.257	1.224
reviews	NMI	0.592	0.603	1.019
	F	0.731	0.749	1.025
	AR	0.573	0.613	1.070
sports	NMI	0.457	0.560	1.225
	F	0.542	0.630	1.162
	AR	0.146	0.359	2.459
ng3	NMI	0.428	0.470	1.098
	F	0.697	0.769	1.103
	AR	0.372	0.453	1.218

3.3.2 聚类集体大小变化时的实验结果 为了比较聚类集体大小对聚类集成算法的影响，本文从 100 个聚类成员中依次选取 $r = 10,20,\cdots,100$ 个聚类成员构建超边的邻接矩阵，并运行不同的聚类集成算法，得到的 NMI 值如图 2 所示，其中 Average 表示每个聚类集成算法在聚类集体大小取 10 个不同值的情况下获得的平均 NMI 值。

由图 2 可见：(1)在相同数据集上，随着聚类集体大小 r 的增加，同一聚类集成算法获得的 NMI 值上下起伏，而不是单调增加；不同聚类集成算法获得最高 NMI 值时的 r 值不一致，例如在数据集 tr11 上(如图 2(a)所示)，CSPA 在 $r = 10$ 时获得最高 NMI 值，CL 在 $r = 60$ 时获得最高 NMI 值，AL 和 KM 在 $r = 80$ 时获得最高 NMI 值，WL 在 $r = 90$ 时获得最高 NMI 值，TSNE+WAL 在 $r = 100$ 时获得最高 NMI 值。(2)对于同一聚类集成算法，在不同数据集上获得最高 NMI 值时的 r 值也不一致，例如，对于 TSNE+WAL，在数据集 tr11, hitech 和 ng3 上获得最高 NMI 值时 $r = 100$ (如图 2(a)，图 2(e)和图 2(h)所示)；在数据集 tr23 上获得最高 NMI 值时 $r = 30$ 和 $r = 80$ (如图 2(b)所示)；在数据集 tr41 上获得最高 NMI 值时 $r = 40$ (如图 2(c)所示)；在数据集 tr45 上获得最高 NMI 值时 $r = 60$ (如图 2(d)所示)；在数据集 reviews 上获得最高 NMI 值时 $r = 90$ (如图 2(f)所示)；在数据集 sports 上获得最高 NMI 值时 $r = 50$ (如图 2(g)所示)。(3)对于给定的聚类成员，要提高聚类集成的质量，可以从聚类成员选择方面着手，也可以从共识函数设计方法方面着手。(4)总体来看，TSNE+WAL 在聚类集体大小变化时在每组数据集上都获得了最高的平均 NMI 值，进一步验证了本文算法的优越性。

4 结束语

本文通过 TSNE 最小化超图邻接矩阵的行对应的高维数据点与低维映射点分布之间的 KL 散度，使得高维空间结构在低维空间得以保持，最后在低维空间设计层次聚类算法获得最终的聚类结果。在基准 TREC 数据集上的实验结果表明：(1)TSNE 能够有效提高层次聚类算法的聚类质量；(2)与其他主流聚类集成方法相比，本文提出的 TSNE 结合层次聚类算法的聚类集成方法获得了更加优越的结果。

本文的实验结果显示，随着聚类集体大小的增加，聚类集成算法获得的 NMI 值并不是单调增加。如何从聚类集体中选出部分聚类成员进行集成，获得比对聚类集体进行集成更加优越的结果是选择性聚类集成的关键问题，目前已经引起研究者们的重视[17,18]，也是本文进一步研究的重点。

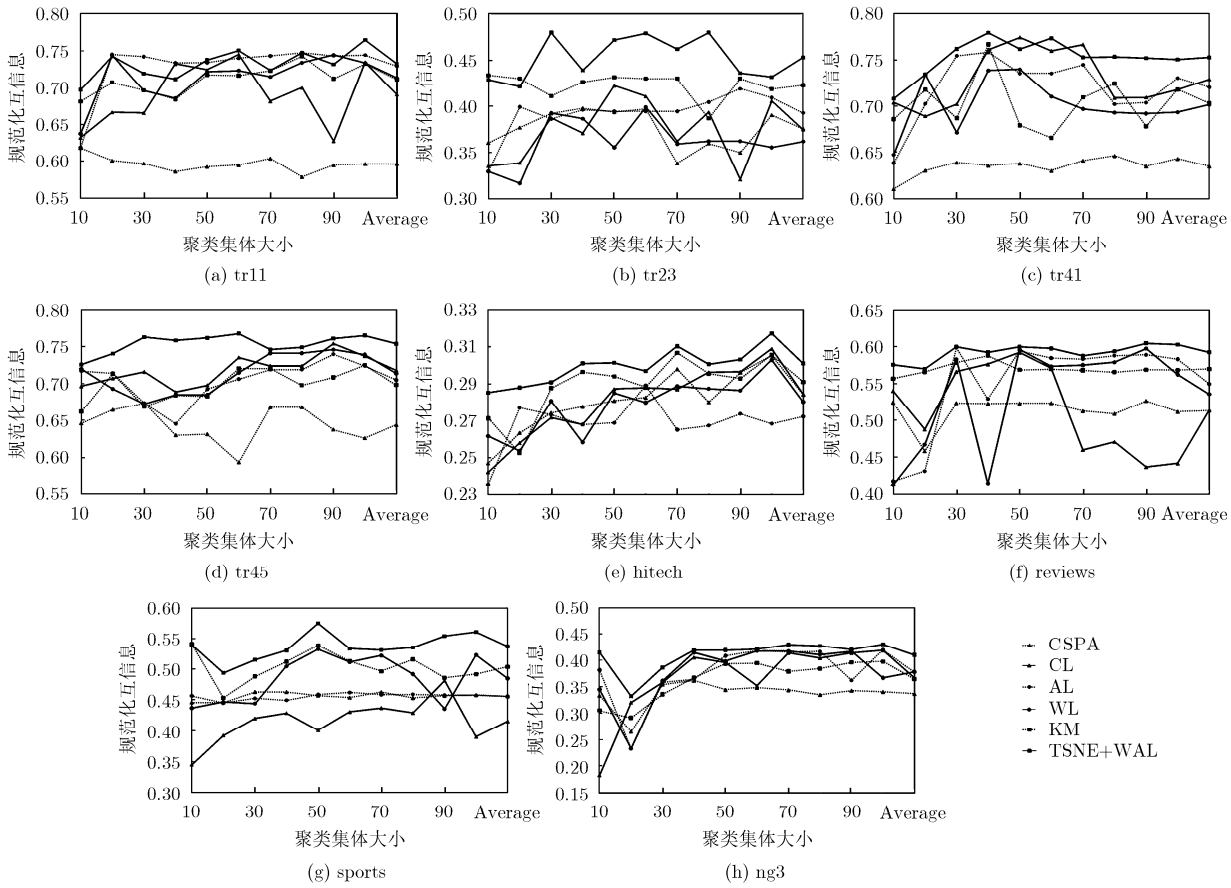

图 2 当聚类集体大小变化时不同聚类集成算法获得的 NMI 值

参 考 文 献

[1] JAIN A K, MURTY M N, and FLYNN P J. Data clustering: A review[J]. *ACM Computing Surveys*, 1999, 31(3): 264–323.

[2] JAIN A K. Data clustering: 50 years beyond K-means[J]. *Pattern Recognition Letters*, 2010, 31(8): 651–666.

[3] 汪晓锋, 刘功申, 李建华. 基于模糊聚类的多分辨率社区发现方法[J]. 电子与信息学报, 2017, 39(9): 2033–2039. doi: 10.11999/JEIT161116.
WANG Xiaofeng, LIU Gongshen, and LI Jianhua. Multiresolution community detection based on fuzzy clustering[J]. *Journal of Electronics & Information Technology*, 2017, 39(9): 2033–2039. doi: 10.11999/JEIT 161116.

[4] STREHL A and GHOSH J. Cluster ensembles: A knowledge reuse framework for combining multiple partitions[J]. *Journal of Machine Learning Research*, 2002, 3: 583–617.

[5] ZHOU Zhihua and TANG Wei. Clusterer ensemble[J]. *Knowledge-Based Systems*, 2006, 19(1): 77–83.

[6] 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): 1315–1323.
LUO Huilan, KONG Fansheng, and LI Yixiao. An analysis of diversity measures in clustering ensembles[J]. *Chinese Journal of Computers*, 2007, 30(8): 1315–1323.

[7] WU Junjie, LIU Hongfu, XIONG Hui, *et al*. K-means based consensus clustering: A unified view[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2015, 27(1): 155–169. doi: 10.1109/TKDE.2014.2316512.

[8] FRED A and LOURENGO A. Cluster ensemble methods: From single clusterings to combined solutions[J]. *Studies in Computational Intelligence*, 2008, 126(1): 3–30.

[9] XU Sen, CHAN Kungsic, Gao Jun, *et al*. An integrated K-means Laplacian cluster ensemble approach for document datasets[J]. *Neurocomputing*, 2016, 214(6): 495–507. doi: 10.1016/j.neucom.2016.06.034.

[10] YU Zhiwen, LI Le, LIU Jiming, *et al*. Adaptive noise immune cluster ensemble using affinity propagation[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2015, 27(12): 3176–3189. doi: 10.1109/TKDE.2015.2453162.

[11] 褚睿鸿, 王红军, 杨燕, 等. 基于密度峰值的聚类集成[J]. 自动化学报, 2016, 42(9): 1401–1412. doi: 10.16383/j.aas.2016.c150864.
CHU Ruihong, WANG Hongjun, YANG Yan, *et al*. Clustering ensemble based on density peaks[J]. *Acta Automatica Sinica*, 2016, 42(9): 1401–1412. doi: 10.16383/j.aas.2016.c150864.

[12] BERIKOV V and PESTUNOV I. Ensemble clustering based on weighted co-association matrices: Error bound and convergence properties[J]. *Pattern Recognition*, 2017, 63: 427–436. doi: 10.1016/j.patcog.2016.10.017.

[13] MAATEN L V D and HINTON G. Visualizing data using t-SNE[J]. *Journal of Machine Learning Research*, 2008, 9(11): 2579–2605.

[14] MAATEN L V D. Learning a parametric embedding by preserving local structure[C]. Proceedings of the 12th International Conference on Artificial Intelligence and Statistics, Clearwater Beach, Florida, USA, 2009: 384–391.

[15] MAATEN L V D. Accelerating t-SNE using tree-based algorithms[J]. *Journal of Machine Learning Research*, 2014, 15(1): 3221–3245.

[16] SALTON G and BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. *Information Processing and Management*, 1998, 24(5): 513–523.

[17] FERN X Z and LIN W. Cluster ensemble selection[J]. *Statistical Analysis & Data Mining*, 2008, 1(3): 128–141.

[18] ZHAO Xingwang, LIANG Jiye, and DANG Chuangyin. Clustering ensemble selection for categorical data based on internal validity indices[J]. *Pattern Recognition*, 2017, 69(4): 150–168. doi: 10.1016/j.patcog.2017.04.019.

徐 森：男，1983 年生，副教授，研究方向为模式识别与人工智能、机器学习.

花小朋：男，1975 年生，副教授，研究方向为模式识别与人工智能、机器学习.

徐 静：女，1982 年生，副教授，研究方向为网络安全、智能信息处理.

徐秀芳：女，1973 年生，高级实验师，研究方向为数据挖掘、智能信息处理.

皋 军：男，1971 年生，教授，研究方向为模式识别与人工智能、机器学习.

安 晶：女，1982 年生，副教授，研究方向为模式识别与人工智能、机器学习.