第45卷 第8期 西 安 交 通 大 学 学 报 Vol. 45 No. 8
2011年8月 JOURNAL OF XI′AN JIAOTONG UNIVERSITY Aug. 2011

采用仿射传播的聚类集成算法

王羡慧[1,3]，覃征[1,2]，张选平[1]，高洪江[4]
（1.西安交通大学电子与信息工程学院，710049，西安；2.清华大学计算机科学与技术系，100084，北京；
3.新疆大学信息科学与工程学院，830046，乌鲁木齐；4.鲁东大学信息科学与工程学院，264025，山东烟台）

摘要：针对K均值聚类随机初始聚类中心导致的聚类结果不稳定问题，提出一种基于仿射传播的聚类集成算法.该算法把每个聚类集成的成员个体结果看成是原始数据的一个属性，然后在其基础上对聚类成员个体的聚类结果进行加权集成，集成算法采用简单高效的仿射传播聚类，并且提出了直接集成、利用平均规范化互信息(NMI)和聚类有效性Silhouette指标进行加权集成.最后，运用Hungarian算法对仿射传播聚类集成的结果进行类别标签的统一和匹配.在加州大学尔湾分校数据集上进行了实验，结果表明，与集成前的K均值聚类及其他聚类集成算法相比，该算法能有效地提高聚类结果的准确性、鲁棒性和稳定性，建立起来的聚类集成算法具有良好的扩展性和灵活性，而且简单有效.

关键词：仿射传播；加权集成；K均值聚类；Hungarian算法
中图分类号：TP181 **文献标志码**：A **文章编号**：0253-987X(2011)08-0001-06

Cluster Ensemble Algorithm Using Affinity Propagation

WANG Xianhui[1,3], QIN Zheng[1,2], ZHANG Xuanping[1], GAO Hongjiang[4]
(1. School of Electronics and Information Engineering, Xi′an Jiaotong University, Xi′an 710049, China;
2. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China;
3. School of Information Science and Engineering, Xinjiang University, Urumqi 830046, China;
4. School of Information Science and Engineering, Ludong University, Yantai, Shandong 264025, China)

Abstract: The result of K-means cluster is instable for random initial clustering centers. A cluster ensemble algorithm based on affinity propagation is proposed, where the result of each cluster individual is regarded as a property of the original data. Following the new properties sets, the results of each cluster individual are carried out to a weighted ensemble, and simple and efficient affinity propagation cluster is chosen in the ensemble algorithm. Furthermore the direct ensemble, the ensemble to weighted ensemble from average normalized mutual information (NMI) and cluster validation indexes Silhouette are uniformly proposed. Finally, Hungarian algorithm is employed to unify and match the category labels for the results of affinity propagation cluster. The results of experiments on University of California Irvine data sets show the higher efficiency for improving the accuracy, robustness and stability of cluster results than the K means clustering before combination and the other clustering ensemble algorithms. The clustering ensemble algorithm gets more extendable and flexible.

Keywords: affinity propagation; weighted cluster ensemble; K means cluster; Hungarian algorithm

收稿日期：2011-03-20. 作者简介：王羡慧(1980—)，男，博士生，讲师；覃征(联系人)，男，教授，博士生导师. 基金项目：国家自然科学基金资助项目(60673024)；高等学校博士学科点专项科研基金资助项目(20100201110063)；国防“十一五”预研资助项目.
网络出版时间：2011-07-18 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20110718.1730.005.html

聚类分析是按照某种相似性测度(例如欧几里德距离)将多维数据分割成自然分组或者簇的过程,是人工智能、模式识别和机器学习的一个重要研究方向.在过去的几十年里,聚类分析已经成功应用于数据挖掘、图像分割、语音识别和信息检索等领域[1-3].K 均值[4-5]算法是所有聚类算法中的经典算法之一,应用广泛.K 均值聚类算法随机选取初始聚类中心,选取的点不同,聚类结果可能就不同.如果由初始聚类中心得到的分类严重偏离全局最优分类,算法就会陷入局部最优解.K 均值聚类算法对初始聚类中心的依赖性,导致了 K 均值聚类结果的不稳定性.

通过聚类集成可以有效地提高 K 均值聚类的准确性、鲁棒性和稳定性[6-8].聚类集成的关键问题是如何根据不同的聚类成员个体聚类结果得到更好的聚类集成结果.文献[9]最早明确定义了聚类集成问题,将聚类集成问题形式化为一个基于互信息的优化问题.然后,提出了获得高质量聚类集成结果的3种有效算法:基于聚类的相似度划分算法(CSPA)、超图划分算法(HGPA)和元聚类算法(MCLA).文献[7]提出了基于投票机制的聚类集成算法.但是,上面这一类算法都不具有可扩展性[10],因此限制了它们的应用范围.文献[10]提出了基于隐含变量的聚类集成模型,将聚类成员个体结果看成是原始数据的属性,在此处理方式基础之上建立聚类集成算法.这种新的视角,使得提出的聚类集成算法具有良好的扩展性和灵活性.

本文利用仿射传播聚类(AP)[11-13]的简单、高效和稳定,将其作为聚类集成算法,提出了一种基于仿射传播的聚类集成算法,可以有效地提高 K 均值聚类的准确性、鲁棒性和稳定性.

1 聚类集成问题

令数据集 $D=\{X_1,X_2,\cdots,X_N\}$,N 为数据点的个数.$\boldsymbol{X}_i=\{x_{i1},x_{i2},\cdots,x_{id}\}(i=1,2,\cdots,N)$ 为一个 d 维数据点.假设有 M 个聚类集成成员个体对数据集 D 进行聚类,得到 M 个长度为 N 的成员个体聚类结果,则聚类集成问题定义为

$$G:\{\boldsymbol{l}_i\mid l_1,l_2,\cdots,l_M\}\rightarrow\boldsymbol{L}^* \tag{1}$$

式中:$\boldsymbol{l}_i$ 和 $\boldsymbol{L}^*$ 是维数为 N 的向量;聚类集体 $l_1,l_2,\cdots,l_M$ 分别是 M 个聚类集成成员个体 $\boldsymbol{l}_i$ 的聚类结果;$\boldsymbol{L}^*$ 是 $l_1,l_2,\cdots,l_M$ 通过集成方法 G 产生的聚类集成结果.

本文借鉴文献[10]的思想,把每个聚类个体结果看成是原始数据的一个属性,然后在其基础上对 N 个 M 维的数据进行聚类,得到原始数据的聚类结果.通过此方法建立起来的聚类集成算法,具有良好的扩展性和灵活性,而且简单有效.由此,聚类集成问题转换为

$$G':\{\boldsymbol{y}_i\mid y_1,y_2,\cdots,y_N\}\rightarrow\boldsymbol{L}' \tag{2}$$

式中:$\boldsymbol{y}_i$ 是维数为 M 的向量,其每一维的分量是聚类成员个体的聚类结果,$\boldsymbol{y}_i$ 总共有 N 个;$\boldsymbol{L}'$ 是维数为 N 聚类集成结果,它是 $y_1,y_2,\cdots,y_N$ 通过集成方法 G' 产生的聚类集成结果.原则上,G' 可以选择任何聚类算法进行集成.

AP 是一种通过数据点之间的消息传递来发现聚类的新方法.与以往的聚类算法相比,AP 能在很短的时间内发现带有更低误差的聚类结果.因此,AP 被成功应用于人脸图像聚类、基因表达数据中的基因识别、文章中的关键句子识别和最优航空路线确定等问题中[11].因此,本文使用 AP 作为聚类集成算法来对聚类成员个体结果进行集成.

2 仿射传播聚类

AP 把一对数据点之间的相似度作为输入,在数据点之间交换有真实价值的消息,直到一个最优的类代表点集合(称为聚类中心或 exemplar)和聚类逐渐形成.此时,所有的数据点到其最近的类代表点的相似度之和最大.

给定数据集合 $D=\{X_1,X_2,\cdots,X_N\}$,$X_i=\{x_{i1},x_{i2},\cdots,x_{id}\}(i=1,2,\cdots,N)$ 为一个数据点.AP 算法以 N 个数据点之间的相似度矩阵 $\boldsymbol{S}=[s(i,k)]_{N\times N}$ 为基础进行聚类.本文选用欧氏距离作为相似度的测度指标,任意两点之间的相似度为两点欧氏距离平方的负数.例如,对于点 X_i 和点 X_k,则有

$$s(i,k)=-\parallel X_i-X_k\parallel^2$$
$$i,k=1,2,\cdots,N;\ i\neq k \tag{3}$$

AP 算法为每一个数据点 k 设定其参数选择 p_k $(k=1,2,\cdots,N)$作为输入,初始 $s(k,k)=p_k$,$(k=1,2,\cdots,N)$.p_k 越大,说明相应的数据点 k 被选中作为类代表点的可能性越大.AP 算法初始假设所有数据点被选中成为类代表点的可能性相同,所以 p_k 取相同值 p,即 $s(k,k)$取相同值 p.

假设数据集合 D 的聚类标记为 $c=\{c_1,c_2,\cdots,c_n\}$,AP 可以被看作一个搜索能量函数最小值的方法

$$E(c)=-\sum_{i=1}^{N}s(i,c_i) \tag{4}$$

式中：聚类标记 c_i 指数据点 i 的典型样例；$s(i,c_i)$ 指数据点 i 和它的典型样例的相似度.

AP 算法中数据点之间有两种消息交换，分别定义为响应度矩阵 $\boldsymbol{R}=[r(i,k)]_{N\times N}$ 和效用度矩阵 $\boldsymbol{A}=[a(i,k)]_{N\times N}$. AP 算法的循环迭代过程就是这两种消息交换迭代更新的过程，每种信息考虑了不同种类的竞争. AP 的信息更新公式如下

$$r(i,k)\leftarrow s(i,k)-\max_{\text{s.t. } k'\neq k}\{a(i,k')+s(i,k')\} \tag{5}$$

$$\text{IF } i\neq k, a(i,k)\leftarrow \min\Big\{0, r(k,k)+\sum_{\text{s.t. } i'\notin\{i,k\}}\max\{0,r(i',k)\}\Big\} \tag{6}$$

$$a(k,k)\leftarrow \sum_{\text{s.t. } i'\neq k}\max\{0,r(i',k)\} \tag{7}$$

AP 算法在每一次循环迭代过程中，$r(i,k)$ 和 $a(i,k)$ 被设置成为：$(1-\lambda)$ 乘以当前迭代过程中的更新值加上 λ 乘以上一步迭代的结果. 阻尼因子取值范围为 $[0,1]$，缺省值为 0.5. 假设当前迭代次数为 t，信息更新为

$$r^{(t)}(i,k)=(1-\lambda)r^{(t)}(i,k)+\lambda r^{(t-1)}(i,k) \tag{8}$$

$$\text{IF } i\neq k,\ a^{(t)}(i,k)=(1-\lambda)a^{(t)}(i,k)+\lambda a^{(t-1)}(i,k) \tag{9}$$

$$a^{(t)}(k,k)=(1-\lambda)a^{(t)}(k,k)+\lambda a^{(t-1)}(k,k) \tag{10}$$

3　基于仿射传播的聚类集成算法

聚类集成可以分为两个阶段：聚类成员个体生成阶段和对聚类成员个体结果进行集成阶段.

3.1　聚类成员个体生成

在聚类成员个体生成阶段，可以通过不同的方法生成成员个体. 文献[9]使用 3 种方法来生成个体：①通过同一算法在数据集的不同特征子集上进行聚类生成个体；②通过同一算法在数据的不同子集上进行聚类生成个体；③通过不同的聚类算法在同一个完整的数据集上进行聚类生成个体. Fred[6,14]利用 K 均值聚类的随机选择初始聚类中心来生成个体. 本文利用 K 均值聚类的随机性，运行多次，每次随机选择初始聚类中心来产生不同的聚类个体.

3.2　聚类集成算法

按照式(1)，文献[9]提出了 3 种基于图划分的聚类集成算法：①CSPA 将数据点作为图的顶点、成对数据点之间的相似性作为边的权重，利用图划分算法 METIS 来得到聚类结果；②HGPA 将数据点作为超图的顶点、每个聚类成员个体的每个簇作为一条超边，利用超图划分算法 HMETIS 来得到聚类结果；③MCLA 将每个聚类成员个体的每个簇作为图的顶点、成对簇之间共有相同数据点的比例作为边的权重，利用图划分算法 METIS 压缩超边集得到元簇，最后根据数据点的比例来决定聚类结果. 按照式(2)，本文使用仿射传播聚类对聚类成员个体结果进行集成. 表 1 为聚类集成 G，表 2 为聚类集成 G'. 聚类集成 G 表示将聚类集成成员个体的聚类结果 $l_1,l_2,\cdots,l_M$ 通过集成方法 G 产生聚类集成结果. 聚类集成 G' 表示将聚类个体结果看成是原始数据的一个属性，通过对 $y_1,y_2,\cdots,y_N$ 使用集成方法 G' 产生聚类集成结果.

表 1　聚类集成 G

数据点	聚类标签			
	l_1	l_2		l_M
X_1	$l_1(X_1)$	$l_2(X_1)$	$\cdots$	$l_M(X_1)$
X_2	$l_1(X_2)$	$l_2(X_2)$	$\cdots$	$l_M(X_2)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
X_N	$l_1(X_N)$	$l_2(X_N)$	$\cdots$	$l_M(X_N)$

表 2　聚类集成 G'

数据点	聚类标签				$\boldsymbol{y}_i$
	l_1	l_2		l_M	
X_1	$l_1(X_1)$	$l_2(X_1)$	$\cdots$	$l_M(X_1)$	y_1
X_2	$l_1(X_2)$	$l_2(X_2)$	$\cdots$	$l_M(X_2)$	y_2
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
X_N	$l_1(X_N)$	$l_2(X_N)$	$\cdots$	$l_M(X_N)$	y_N

3.3　聚类成员个体聚类结果加权集成

由于聚类成员个体之间的差异性，不同聚类成员个体对最终聚类集成结果的影响力不同. 现有的聚类集成算法将每个聚类成员个体的结果平等对待，这忽略了集成个体的差异性. 因此，本文对聚类成员个体的结果在聚类集成 G' 的基础上进行加权集成. 假设聚类成员个体 $\boldsymbol{l}_i(i=1,2,\cdots,M)$ 的加权权重为 $w^{(l_i)}$，本文提出 4 种加权方法.

(1)平等对待聚类成员个体，直接集成，即权值相等

$$w_1{}^{(l_i)}=\frac{1}{M} \tag{11}$$

(2)利用聚类成员个体结果之间的规范化互信

息[8-9]来表示聚类成员个体的差异度,然后利用平均规范化互信息来加权. 假设有两个聚类成员个体的聚类结果为 $l^{(a)}$ 和 $l^{(b)}$,则两者的规范化互信息表示为

$$I(l^{(a)},l^{(b)}) = \sum_{h=1}^{k^{(a)}}\sum_{l=1}^{k^{(b)}} n_{h,l}\,\mathrm{lb}\left(\frac{nn_{h,l}}{n_h^{(a)}\,n_l^{(b)}}\right)\Big/ \left(\left(\sum_{h=1}^{k^{(a)}} n_h^{(a)}\,\mathrm{lb}\,\frac{n_h^{(a)}}{n}\right)\left(\sum_{l=1}^{k^{(b)}} n_l^{(b)}\,\mathrm{lb}\,\frac{n_l^{(b)}}{n}\right)\right)^{1/2} \tag{12}$$

式中:$k^{(a)}$ 和 $k^{(b)}$ 分别为聚类结果 $l^{(a)}$ 和 $l^{(b)}$ 中簇的个数;$n_{h,l}$ 为同时位于 $l^{(a)}$ 的 h 簇和 $l^{(b)}$ 的 l 簇中数据点的个数;$n_h^{(a)}$ 为 $l^{(a)}$ 的 h 簇中点的个数;$n_l^{(b)}$ 为 $l^{(b)}$ 的 l 簇中点的个数. 对于单个聚类成员个体 $\boldsymbol{l}_i$,其平均规范化互信息表示为

$$I_A^{(l_i)} = \frac{1}{M-1}\sum_{j=1,j\neq i}^{M} I(l_i,l_j) \tag{13}$$

利用 $I_A^{(l_i)}$ 来加权,权值表示为

$$w_2{}^{(l_i)} = I_A^{(l_i)}\Big/\sum_{i=1}^{M} I_A^{(l_i)}$$

$$\sum_{i=1}^{M} w_2{}^{(l_i)} = 1,\quad w_2{}^{(l_i)} > 0,\quad i = 1,2,\cdots,M \tag{14}$$

(3)利用 $I_A^{(l_i)}$ 的倒数来加权,权值定义如下

$$w_3{}^{(l_i)} = \frac{1}{I_A^{(l_i)}}\Big(\sum_{i=1}^{M}\frac{1}{I_A^{(l_i)}}\Big)^{-1}$$

$$\sum_{i=1}^{M} w_3{}^{(l_i)} = 1,\quad w_3{}^{(l_i)} > 0,\quad i = 1,2,\cdots,M \tag{15}$$

(4)聚类有效性 Silhouette[15-16]指标可以用来评价聚类结果的质量,令聚类的 Silhouette 指标值为 S. 本文利用 Silhouette 指标值 S 来加权,权值定义如下

$$w_4{}^{(l_i)} = S^{(l_i)}\Big/\sum_{i=1}^{M} S^{(l_i)}$$

$$\sum_{i=1}^{M} w_4{}^{(l_i)} = 1,\quad w_4{}^{(l_i)} > 0,\quad i = 1,2,\cdots,M \tag{16}$$

3.4 聚类集成结果准确性评价

聚类成员个体产生的聚类结果使用 AP 聚类进行加权集成后,为了评价聚类集成结果的准确性,需要对聚类集成结果和数据的分类结果进行类别标签的统一和匹配. 本文使用 Hungarian[17]算法完成统一和匹配.

聚类集成结果经过类别标签的统一和匹配后,可以使用 Micro-Precision[18]测度来进行准确性评价. Micro-Precision 测度 P_{MP} 定义如下

$$P_{MP} = \frac{1}{N}\sum_{i=1}^{Q} a_i \tag{17}$$

式中:Q 为聚类集成结果中类别的个数;a_i 为分类正确的某一类数据点的个数.

4 实 验

4.1 实验数据集

本文实验使用加州大学尔湾分校(UCI)标准数据集[19],9 个数据集信息如表 3 所示.

表 3 实验使用的 UCI 数据集信息

数据集	样本数	特征数	类别数
iris	150	4	3
ionosphere	351	34	2
wine	178	13	3
wpbc	198	33	2
bupa	345	6	2
glass	214	9	6
balance-scale	625	4	3
wdbc	569	30	2
Image_segmentation	2 100	19	7

4.2 实验结果

本文选择推荐的 20 个聚类集成成员个体进行集成[8],即 $M=20$. 聚类成员个体由 K 均值聚类每次随机初始化聚类中心产生,聚类中心的个数为数据集的类别数. 为了增加实验结果的可靠性,所有的实验结果为运行 50 次的平均值. 对比实验采用 K 均值聚类算法、AP 算法、CSPA 算法、HGPA 算法、MCLA 算法. 本文所提出的算法为 APE-w_1、APE-w_2、APE-w_3 和 APE-w_4,其中 APE 代表基于仿射传播的聚类集成方法,w_1、w_2、w_3、w_4 分别代表 4 种加权集成方法. 表 4 列出了所有对比集成算法的实验结果,实验结果为 50 次运行实验的均值和标准差.

根据表 4,比较 APE 聚类集成算法和集成前的 K 均值算法. 由表 4 可知,在数据集 iris、wine 和 balance-scale 上,4 种 APE 聚类集成算法的聚类准确性明显优于集成前的 K 均值算法;在数据集 ionosphere、wpbc、bupa 和 wdbc 上,4 种 APE 聚类集

成算法获得了和集成前 K 均值算法同等的性能；在数据集 Image_segmentation 上，APE-w_2 和 APE-w_3 的聚类准确性优于集成前的 K 均值算法，APE-w_1 和 APE-w_4 的低于集成前的 K 均值算法；在数据集 glass 上，4 种 APE 聚类集成算法的聚类准确性都低于集成前的 K 均值算法.

表 4　不同聚类集成算法的实验结果

数据集	聚类准确度								
	K 均值	CSPA	HGPA	MCLA	APE-w_1	APE-w_2	APE-w_3	APE-w_4	AP
iris	0.819 1±0.027 1	0.870 5±0.018 1	0.602 4±0.089 9	0.893 3±0.000 0	0.893 3±0.000 0	0.8933±0.000 0	0.893 3±0.000 0	0.893 3±0.000 0	0.893 3±0.000 0
ionos-phere	0.712 3±0.000 0	0.678 1±0.000 0	0.584 0±0.000 0	0.712 3±0.000 0	0.712 3±0.000 0	0.712 3±0.000 0	0.712 3±0.000 0	0.712 3±0.000 0	0.709 4±0.000 0
wine	0.674 7±0.011 3	0.680 3±0.006 0	0.550 8±0.083 2	0.702 2±0.000 0	0.701 0±0.008 7	0.702 2±0.000 0	0.696 9±0.018 0	0.701 0±0.008 7	0.719 1±0.000 0
wpbc	0.636 4±0.000 0	0.560 6±0.000 0	0.535 4±0.000 0	0.636 4±0.000 0	0.636 4±0.000 0	0.636 4±0.000 0	0.636 4±0.000 0	0.636 4±0.000 0	0.601 0±0.000 0
bupa	0.553 6±0.000 0	0.562 3±0.000 0	0.507 5±0.005 3	0.553 6±0.000 0	0.553 6±0.000 0	0.553 6±0.000 0	0.553 6±0.000 0	0.553 6±0.000 0	0.553 6±0.000 0
glass	0.517 8±0.007 0	0.409 3±0.027 6	0.375 0±0.043 3	0.466 4±0.035 2	0.474 2±0.024 5	0.476 4±0.023 7	0.472 4±0.023 2	0.491 2±0.031 9	0.537 4±0.000 0
balance-scale	0.510 5±0.016 2	0.514 2±0.021 6	0.413 3±0.009 1	0.500 6±0.031 9	0.511 5±0.053 7	0.518 7±0.044 6	0.513 9±0.060 0	0.513 7±0.052 2	0.500 8±0.043 3
wdbc	0.854 1±0.000 0	0.669 6±0.000 0	0.514 9±0.000 0	0.854 1±0.000 0	0.854 1±0.000 0	0.854 1±0.000 0	0.854 1±0.000 0	0.854 1±0.000 0	0.848 9±0.000 0
Image_ segmentation	0.528 2±0.011 6	0.435 8±0.064 6	0.546 7±0.025 2	0.520 4±0.021 6	0.486 1±0.115 8	0.532 5±0.033 1	0.523 3±0.045 5	0.489 1±0.097 5	0.492 6±0.000 0

由表 4 可知，在数据集 iris、ionosphere、wpbc、glass 和 wdbc 上，4 种 APE 聚类集成算法都获得了优于 CSPA、HGPA、MCLA 或者同等的聚类准确度；在数据集 wine 上，APE-w_2 和 MCLA 获得了此数据集上的最好聚类集成性能，APE-w_1、APE-w_4 和 APE-w_3 的集成性能略差，但仍优于 CSPA 和 HGPA；在数据集 balance-scale 上，APE-w_2 获得了此数据集上的最好聚类集成性能；在数据集 bupa 上，虽然 CSPA 获得了此数据集上的最好聚类集成性能，但是 4 种 APE 聚类集成算法都优于 HGPA 且等于 MCLA；在数据集 Image_segmentation 上，虽然 HGPA 获得了此数据集上的最好聚类集成性能，但是 4 种 APE 聚类集成算法都优于 CSPA，APE-w_2 和 APE-w_3 优于 MCLA.

由表 4 可知，在数据集 iris、ionosphere、wpbc、bupa、balance-scale 和 wdbc 上，4 种 APE 聚类集成算法都获得了与 AP 算法同等或更优的聚类准确度；在数据集 Image_segmentation 上，APE-w_2 和 APE-w_3 优于 AP 算法，但是 AP 算法优于 APE-w_1 和 APE-w_4；在 wine 和 glass 上，AP 算法优于 4 种 APE 聚类集成算法.

根据表 4，在数据集 iris、ionosphere、wpbc、bupa 和 wdbc 上，4 种 APE 加权集成算法获得了同等的聚类准确度；在数据集 wine、balance-scale 和 Image_segmentation 上，APE-w_2 都获得了数据集上 4 种 APE 聚类集成算法中的最好性能；在数据集 glass 上，APE-w_4 获得了 4 种 APE 聚类集成算法中的最好性能.

综上所述，在大多数数据集上，APE 聚类集成算法都有效地提高了集成前聚类结果的准确性. 与 CSPA、HGPA、MCLA 经典聚类集成算法相比，在大多数数据集上，APE 能够获得更优的聚类集成结果. APE 与 AP 算法相比，在较为复杂的高维数据集（ionosphere、wpbc、wdbc、Image_segmentation）上，通过引入集成学习，APE 可以更加有效地提高 AP 算法的聚类准确性，这说明 APE 对处理高维复

杂数据更有优势. 4 种 APE 加权集成算法中，APE-w_2 在较多的数据集上获得了最优性能，这说明利用平均规范化互信息来加权集成是一种较好的方案.

5 结 论

本文提出一种基于仿射传播的聚类集成算法. 该算法把每个聚类集成的成员个体结果看成是原始数据的一个属性，然后在其基础上对聚类成员个体的聚类结果进行加权集成，并采用简单高效的仿射传播聚类. 通过此方法建立起来的聚类集成算法，具有良好的扩展性和灵活性，而且简单有效. 当运行聚类分析任务时，在没有先验知识的情况下，在大多数数据集上，本文所提算法有效地提高了集成前聚类结果的准确性，并且获得了优于或等于传统的 CSPA、HGPA、MCLA 的聚类结果. 在 UCI 数据集上的实验结果表明，与集成前的 K 均值聚类及其他聚类集成算法相比，本文方法能有效地提高聚类结果的准确性、鲁棒性和稳定性.

参考文献：

[1] XU R, WUNSCH D. Survey of clustering algorithms [J]. IEEE Transactions on Neural Networks, 2005, 16 (3): 645-678.

[2] OMRAN M G H, ENGELBRECHT A P, SALMAN A. An overview of clustering methods[J]. Intelligent Data Analysis, 2007, 11 (6): 583-605.

[3] 孙吉贵，刘杰，赵连宇. 聚类算法研究[J]. 软件学报，2008，19(1)：48-61.
SUN Jigui, LIU Jie, ZHAO Lianyu. Clustering algorithms research [J]. Journal of Software, 2008, 19 (1): 48-61.

[4] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C]// Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, California, USA: University of California Press, 1967: 281-297.

[5] 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 Zhimao, GU Guochang. Two spectral algorithms for ensembling document clusters[J]. Acta Automatica Sinica, 2009, 35(7): 997-1002.

[6] FRED A, JAIN A. Combining multiple clusterings using evidence accumulation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27 (6): 835-850.

[7] ZHOU Z H, TANG W. Clusterer ensemble [J]. Knowledge-Based Systems, 2006, 19 (1): 77-83.

[8] 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 Huilan, KONG Fansheng, LI Yixiao. An analysis of diversity measures in clustering ensembles[J]. Chinese Journal of Computers, 2007, 30 (8): 1315-1324.

[9] STREHL A, GHOSH J. Cluster ensembles: a knowledge reuse framework for combining multiple partitions [J]. The Journal of Machine Learning Research, 2002 (3): 583-617.

[10] 王红军，李志蜀，成飚，等. 基于隐含变量的聚类集成模型[J]. 软件学报，2009，20 (4)：825-833.
WANG Hongjun, LI Zhishu, CHENG Biao, et al. A latent variable mode for cluster ensemble [J]. Journal of Software, 2009, 20 (4): 825-833.

[11] FREY B J, DUECK D. Clustering by passing messages between data points [J]. Science, 2007, 315 (5814): 972-976.

[12] FREY B J, DUECK D. Response to comment on "clustering by passing messages between data points" [J]. Science, 2008, 319 (5864): 2.

[13] MEZARD M. Computer science: where are the exemplars? [J]. Science, 2007, 315 (5814): 949-951.

[14] FRED A. Finding consistent clusters in data partitions [M]. Heidelberg, German: Springer, 2001: 309-318.

[15] HRUSCHKA E R, CAMPELLO R, FREITAS A A, et al. A survey of evolutionary algorithms for clustering[J]. IEEE Transactions on Systems, Man and Cybernetics: Part C Applications and Reviews, 2009, 39 (2): 133-155.

[16] KAUFMAN L, ROUSSEEUW P J. Finding groups in data: an introduction to cluster analysis [M]. New York: Wiley, 1990: 7-64.

[17] KUHN H W. The Hungarian method for the assignment problem[J]. Naval Research Logistics Quarterly, 1955, 2 (2): 83-97.

[18] MODHA D S, SPANGLER W S. Feature weighting in k-means clustering[J]. Machine Learning, 2003, 52 (3): 217-237.

[19] FRANK A, ASUNCION A. UCI machine learning repository [EB/OL]. [2010-12-22]. http://archive.ics.uci.edu/ml.

（编辑 杜秀杰 武红江）