文章编号: 1001-0920 (2015) 05-0823-08 DOI: 10.13195/j.kzyjc.2014.0358

基于模糊测度和证据理论的模糊聚类集成方法

毕 凯, 王晓丹, 邢雅琼

(空军工程大学 防空反导学院，西安 710051)

摘 要: 针对现有集成方法在处理模糊聚类时存在的不足, 提出一种基于证据理论的模糊聚类集成方法. 以各聚类成员作为证据元, 以样本点间的类别关系作为焦元, 通过证据积累构造互相关矩阵. 考虑到模糊聚类对于各样本点的聚类有效性, 提出一种结合点模糊度和模糊贴近度的类别关系表示方法, 并以此作为各证据元的基本概率赋值函数. 最后基于互相关矩阵构造样本点间相似性关系, 并利用谱聚类算法对其聚类. 实验中通过与多种已有聚类集成方法的对比表明, 该方法具有较高的聚类性能.

关键词: 模糊聚类集成；模糊贴近度；模糊度；D-S 证据理论；互相关矩阵

中图分类号: TP391 **文献标志码:** A

Fuzzy clustering ensemble based on fuzzy measure and DS evidence theory

BI Kai, *WANG Xiao-dan*, *XING Ya-qiong*

(School of Air and Missile Defense，Air Force Engineering University，Xi'an 710051，China. Correspondent：BI Kai, E-mail：bk3039633@163.com)

Abstract: In order to solve the weakness of present ensemble methods for fuzzy clustering, a method of fuzzy clustering ensemble based on Dempster-Shafer(DS) theory is proposed, which takes every cluster member as the evidence, takes the relationship between pair of dates as the event and makes the co-matrix by accumulation of the evidence. Considering the classification effectiveness of each data point, a relationship based on the fuzzy degree and the fuzzy close-degree is proposed, and the new relationship is considered as the basic probability assignment for the evidence. Finally, the spectral clustering algorithm is used to make the final partition by similarity relationship based on the co-association matrix. Experimental results show that the method proposed is better than the current methods for clustering ensemble.

Keywords: fuzzy clustering ensemble; fuzzy close-degree; fuzzy degree; Dempster-Shafer evidence theory; co-association matrix

0 引 言

聚类集成是 Strehl 等[1]于 2002 年提出的, 由于具有较好的鲁棒性、新颖性、稳定性、并行性以及可扩展性[2]等优点, 近年来已成为机器学习领域的研究热点. 与有监督分类不同, 聚类过程没有训练样本, 缺乏类别规模和数据结构等先验信息, 因而存在一定程度的盲目性. 聚类集成以差异性聚类结果为处理对象, 构造聚类结果间的一致性度量, 期望通过组合多种差异的结果来强化相关信息、弱化无关信息, 进而获取更为合理的聚类结果. 当前聚类集成的研究主要集中在构造差异性聚类成员和设计共识函数上.

差异性是决定聚类集成效果的重要指标, 差异性越大越有利于削弱无关信息、保留相关信息, 进而形成良好的聚类结果. 常用的聚类成员生成方法有: 算法扰动[3]、参数扰动[4-5]以及样本扰动[6]等. 共识函数被用于描述聚类结果间关系, 由于聚类结果间缺乏必要的联系, 共识函数的设计一直是研究的重点. 如: 文献 [1] 利用超图的方法描述聚类成员间关系, 通过图划分实现集成; 文献 [4] 利用互相关矩阵描述聚类成员内样本间类别关系, 将集成过程近似为证据的积累过程; 文献 [7] 提出可以利用聚类结果的类别信息构造新的特征空间, 因而集成过程便转换为新的特征空间内的聚类过程.

虽然已有大量文献对聚类集成进行了深入研究,

收稿日期: 2014-03-16；**修回日期:** 2014-07-04.

基金项目: 国家自然科学基金项目(60975026, 61273275).

作者简介: 毕凯(1985−), 男, 博士生, 从事智能信息处理和机器学习的研究；王晓丹(1966−), 女, 教授, 博士生导师, 从事智能信息处理和机器学习等研究.

但多数文献的研究仅针对硬聚类算法, 而模糊聚类的研究相对较少. 如: 文献 [8] 给出了两种模糊数间的相似度量以描述聚类结果中样本间相互关系; 文献 [9] 给出了一种软投票的思想来描述证据的积累过程; 文献 [10] 在提出一种模糊谱聚类算法的基础上, 讨论了协方差对于模糊聚类集成的意义; 文献 [11] 基于模糊聚类给出了一种类别数的确定框架, 但在对模糊聚类结果进行处理时, 则是按照隶属度最大原则将结果去模糊化; 文献 [12] 基于投票决策理论, 讨论了 4 种投票方法, 即 sum voting、product voting、borda voting、copeland voting, 并将它们应用于模糊聚类集成, 但是这些投票方法均假设聚类成员的类别数相等, 这在许多实际情况中并不适用.

为了适用于模糊聚类结果的类别关系描述, 本文提出一种结合贴近度和点模糊度的类别关系描述方法, 同时将证据理论引入模糊聚类集成中, 通过各聚类成员的证据积累构造互相关矩阵. 最后通过在标准数据集上进行实验, 验证了本文方法的有效性.

1 相关理论介绍

1.1 证据理论

在证据理论中, 样本空间称为一个辨识框架, 常用 Θ 表示, 它是关于命题的彼此独立的可能解集, Θ 完备且其中元素互不相容, Θ 的幂集记为 2^{Θ}. 证据理论的基本问题是, 在确定辨识框架内确定一个先验未定元素属于 Θ 中某一子集的程度[13].

如果函数 $m:2^{\Theta}\to[0,1]$, 满足 $m(\varnothing)=0$ 且 $\sum\limits_{A\subseteq\Theta}=1$, 则称 m 为该识别框架上的基本概率赋值函数 (BPA). A 为事件, $m(A)$ 为 A 的基本可信度. 事件 A 的不确定度可描述为 $[\mathrm{Bel}(A),\mathrm{Pl}(A)]$, Bel 和 Pl 分别为事件 A 的信任函数和似真函数, 其值由下式确定:

$$\begin{cases}\mathrm{Bel}(A)=\sum\limits_{B\subseteq A}m(B),\\ \mathrm{Pl}(A)=\sum\limits_{A\bigcap B=\varnothing}m(B).\end{cases}\tag{1}$$

可见二者分别为事件 A 可信程度的下限估计和上限估计.

对于同一辨识框架下的两证据, m_1 和 m_2 为相应的 BPA, 其焦元分别为 $A_1,A_2,\cdots,A_p$ 和 $B_1,B_2,\cdots,B_p$. 对于给定命题 $A\subseteq\Theta$, 两个证据可通过 Dempster 规则进行组合, 即

$$m(A)=m_1\oplus m_2(A)=\frac{1}{1-q}\sum_{A_i\bigcap B_j=A}m_1(A_i)m_2(B_j).\tag{2}$$

其中: $q=\sum\limits_{A_i\bigcap B_j=\varnothing}m_1(A_i)m_2(B_j)$ 为两证据的冲突度量, $q=0$ 时表示两证据完全没有冲突, $q=1$ 时表示两证据完全冲突, 上述合成规则不再适用; 系数 $1/(1-q)$ 为归一化因子, 避免在合成时将非 0 的信任赋给空集.

1.2 贴近度和模糊度

贴近度是描述两个模糊集相似程度的一种度量, 反映了两个模糊集间的距离[14-15], 下式为其通用表达式:

$$N(A,B)=1-\frac{1}{n^{\frac{1}{k}}}\Big(\sum_{i=1}^{n}|A(\mu_i)-B(\mu_i)|^k\Big)^{\frac{1}{k}}.\tag{3}$$

其中: n 为论域内元素个数; A、B 为两模糊集, 根据 k 的不同取值, 分别为海明贴近度、欧氏贴近度等.

模糊度是描述数据集模糊程度的度量[16], 其一般表达式为

$$\begin{cases}d_p(A)=\dfrac{2}{n^{\frac{1}{p}}}\Big(\sum\limits_{i=1}^{n}|A(\mu_i)-A_{\frac{1}{2}}(\mu_i)|^p\Big)^{\frac{1}{p}};\\ A_{\frac{1}{2}}(\mu_i)=\begin{cases}1,\ A(\mu_i)\geqslant 0.5;\\ 0,\ A(\mu_i)<0.5.\end{cases}\end{cases}\tag{4}$$

其中 $A_{\frac{1}{2}}(\mu_i)$ 为 $\lambda=0.5$ 时 A 的 λ 截集. 当 A 为普通集合时, $d_p(A)=0$; A 中各模糊值越接近于 0.5 就越模糊, 当 $A(\mu_i)=0.5$ 时, $d_p(A)=1$, 为最模糊情况.

2 模糊聚类集成

与硬聚类不同, 模糊聚类具有鲜明的特点. 在硬聚类中, 样本点与类别间的关系为一对一, 即每一个样本点只有唯一确定的类别; 而模糊聚类中, 样本点与类别间的关系是一对多, 样本点与各类别由隶属度相联系. 虽然隶属度代替确定类别能够充分描述样本间关系, 但是也给模糊聚类的集成提出了挑战.

2.1 模糊聚类的点模糊度

模糊聚类作为聚类算法的模糊拓展, 虽然与模糊集关系密切, 但其隶属度描述和模糊性度量具有自身特点.

记 x_i 为任意样本点, c 为类别数, μ_{ij} 为样本点 x_i 属于类别 j 的隶属度, 则有:

1) $\sum\limits_{j=1}^{c}\mu_{ij}=1$;

2) 当 $\mu_{ij}=1/c$ 时, 点 x_i 的模糊度最大;

3) 当存在某个 $\mu_{ij}=1$ 时, 点 x_i 的模糊度最小.

容易发现式 (4) 的模糊度描述并不适用于描述模糊聚类的结果.

划分系数 PC 和划分熵系数 PE 是两个描述模糊聚类有效性的常用指标, 二者形式为

$$\mathrm{PC}=\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c}\mu_{ij}^2,$$

$$\mathrm{PE}=-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{c}\mu_{ij}\log_a\mu_{ij},\tag{5}$$

其中 N 为数据集规模. PC 取值区间为 $[1/c,1]$, 越接近于 1, 聚类算法越硬, 即模糊度越小; 反之越接近于 $1/c$, 模糊度越大. PE 取值区间为 $[0,\log_a c]$, 越接近于 0, 结果模糊度越小; 越接近于 $\log_a c$, 模糊度越大.

式 (5) 通常用于描述模糊聚类结果的整体有效性, 其实质上是样本集内所有点模糊性的求和. 基于该思想, 这里给出两个模糊聚类中样本点的模糊度 (PFD) 指标: 基于划分系数的点模糊度 (PPC) 和基于划分熵系数的点模糊度 (PPE), 其表达式为

$$\mathrm{PPC}_i=\frac{c}{c-1}\Big(1-\sum_{j=1}^{c}\mu_{ij}^2\Big),$$

$$\mathrm{PPE}_i=\frac{-\sum_{j=1}^{c}\mu_{ij}\log_a\mu_{ij}}{\log_a c}.\tag{6}$$

式 (6) 中 PPC_i 和 PPE_i 的取值范围均为 $[0,1]$, 对于任意 x_i, 当 $\mu_{ij}=1/c$ 时取得最大值 1, 表示聚类方法不能判断该点类别, 此时类别信息最模糊; 当存在某个 $\mu_{ij}=1$, 其他 $\mu_{ij}=0$ 时, 取得最小值 0, 表示聚类方法给予该点确定的类别信息.

2.2　模糊聚类集成的互相关矩阵

聚类集成通常可描述为, 对于规模为 N 的样本集合 $X=\{x_1,x_2,\cdots,x_N\}$, 其 M 次差异性聚类结果可表示为 $\Pi=\{\pi_1,\pi_2,\cdots,\pi_M\}$, 其中 $\pi_i=\{C_1^i,C_2^i,\cdots,C_{c_i}^i\}$, c_i 为第 i 次聚类的结果类别数. 聚类集成的目的是将集合 Π 进行合并, 得到样本集合的最终划分结果[3].

共识函数通常被用于揭示 Π 中各类别向量间的关系, 常见的共识函数设计方法主要包括: 1) 互相关矩阵方法; 2) 图划分方法; 3) 代价函数方法. 其中互相关矩阵方法, 因其构造简单、直观、性能较为准确而得到广泛研究, 其形式如下式所示:

$$\mathrm{CoMatrix}=\begin{bmatrix}s_{11}&s_{12}&\cdots&s_{1N}\\s_{21}&s_{22}&\cdots&s_{2N}\\\vdots&\vdots&\ddots&\vdots\\s_{N1}&s_{N2}&\cdots&s_{NN}\end{bmatrix}.\tag{7}$$

互相关矩阵为 $N\times N$ 矩阵, $s_{ii}=1$, 当 $i\neq j$ 时, s_{ij} 为样本点 x_i 与 x_j 被划为同类的证据支持. 在硬聚类集成中, 样本点间类别关系是确定的 (同类或异类), 可利用投票法简单获取; 但在模糊聚类的集成中, 由于样本点的模糊类别表示, 使得投票法难以应用. 证据理论作为一种不确定性推理方法, 具有坚实的理论基础, 相比较贝叶斯方法等更适合于无先验信息的融合, 利用证据的积累作用缩小假设集合, 在区分不确定以及精确反映证据收集过程等方面显示了很大的灵活性. 本节讨论一种基于证据理论的互相关矩阵构造方法, 利用贴近度和点模糊度确定同一聚类成员内样本点间关系, 并利用 Dempster 规则通过证据积累构造互相关矩阵.

令 $m(S)$ 表示样本间的同类关系, $m(D)$ 表示样本间的异类关系, 考虑到模糊贴近度对于模糊集的相似程度描述, 对于任意两样本点 x_p 和 x_q, 其类别关系可描述为

$$m(S_{pq})=1-\frac{1}{c^k}\Big(\sum_{j=1}^{c}|\mu_{pj}-\mu_{qj}|^k\Big)^{\frac{1}{k}},$$

$$m(D_{pq})=\frac{1}{c^k}\Big(\sum_{j=1}^{c}|\mu_{pj}-\mu_{qj}|^k\Big)^{\frac{1}{k}}.\tag{8}$$

需要指出的是, 利用式 (8) 描述同一聚类成员中两样本点间类别关系, 虽然具有一定的合理性, 但却忽略了聚类对于两样本点的实际划分能力, 因而将导致过多不准确类别关系的引入.

例如, 假设某一聚类结果类别数为3, x_1 和 x_2 为任意两样本点, 它们与 3 个类别间关系如表 1 所示. 根据式 (8) 计算可得 ($k=1$, 即海明贴近度时) $m(S_{12})=0.993\,3$, $m(D_{12})=0.006\,7$. 由计算结果可以直观认为二者以极高的可能性属于同类, 但二者的隶属度实际上是划分不理想的表现, 可见利用式 (8) 描述其类别关系并不合理.

表 1　模糊聚类样本点类别关系

数据集	隶属度		
	Class1	Class2	Class3
x_1	0.33	0.33	0.34
x_2	0.34	0.33	0.33

利用式 (6) 计算表 1 中的两点, 结果如表 2 所示 (PPE 中 a 取自然数 e). 观察表 2, 聚类算法对于二者均具有极高的模糊度, 表明聚类结果并不能给予二者相对明确的类别.

表 2　样本点模糊度

样本点	模糊度	
	PPC	PPE
x_1	0.999 9	0.999 9
x_2	0.999 9	0.999 9

利用下式描述两样本点的组合模糊度 PFD_2:

$$\mathrm{PFD}_2=\mathrm{PFD}(x_1)+\mathrm{PFD}(x_2)-\mathrm{PFD}(x_1)\mathrm{PFD}(x_2).\tag{9}$$

其值在 $[0,1]$ 内取值, 该值越大表示模糊度越高, 反之则模糊度越低.

若将两样本点间类别关系视为贴近度和组合模糊度的函数, 则有

$$m'(S)=f(m(S),\mathrm{PFD}_2),$$

$$m'(D) = 1 - m'(S). \tag{10}$$

较为合理的式(10)应满足如下关系: 1) 当$\mathrm{PFD}_2 = 1$时, $m'(S)$与$m'(D)$应为0.5, 表示不能判断两样本点间类别; 2) 当$\mathrm{PFD}_2 = 0$时, 表示两样本点均有明确的类别关系, $m'(S) = m(S)$, $m'(D) = m(D)$; 3) PFD_2在$[0,1]$区间内单调递增时, $|m'(S) - 0.5|$的值趋近于0, 表明聚类结果的模糊性越高, 二者的类别关系越模糊.

基于上述分析, 这里给出如下3种样本间类别关系的描述:

$$m'(S) = \begin{cases} m(S) + \mathrm{PFD}_2 \dfrac{1 - 2m(S)}{2}, \ m(S) \neq 1; \\ 1 - \dfrac{\mathrm{PFD}_2}{2}, \ m(S) = 1; \end{cases} \tag{11}$$

$$m'(S) = \Big(\frac{1 + 2m(S)}{2}\Big)^{1-\mathrm{PFD}_2} - \frac{1}{2}; \tag{12}$$

$$m'(S) = \frac{1 - 2m(S)}{2} \log_2(\mathrm{PFD}_2 + 1) + m(S). \tag{13}$$

式(11)描述了模糊度与类别关系的线性变化, 式(12)描述了二者的指数变化, 式(13)描述了二者的对数变化, 容易证明式(11)～(13)满足上述3个条件. 利用它们重新计算例子中点间类别关系, 结果如表3所示. 由于考虑了样本点自身的模糊性, 调整后的点间类别关系$m'(S)$更为合理.

表 3 调整后类别关系

模糊度	调整函数		
	线性	指数	对数
PPC	0.5000	0.5000	0.5000
PPE	0.5000	0.5000	0.5000

以各聚类成员作为证据元, 以任意两样本点间类别关系作为焦元, 利用式(11)、(12)或(13)所求得的类别关系来确定BPA, 则互相关矩阵的计算步骤如下.

算法 1

输入: M个模糊聚类成员π_i;

输出: 互相关矩阵CoMatrix.

Step 1: 对于聚类成员π_i, 利用式(6)和(9)计算任意两点的组合模糊度;

Step 2: 利用式(8)计算π_i中任意两点间的模糊贴近度;

Step 3: 利用式(10)调整各聚类成员中样本点间类别关系;

Step 4: 以各聚类成员为证据元, 利用Step 3所求的样本点间类别关系为BPA, 利用Dempster规则进行证据融合, 求取互相关矩阵CoMatrix.

2.3 互相关矩阵的谱聚类

基于2.2节分析, 互相关矩阵本质上是模糊聚类成员对样本点间关系的证据积累, 是样本点间关系的整体描述. 因此, 本文以互相关矩阵作为样本点的相似性度量, 对其进行聚类以获得集成的最终结果.

谱聚类算法是最近出现的聚类算法, 具有识别非凸结构的能力, 实现简单且不会陷入局部最优, 能够避免维数过高而造成的奇异性问题[10], 近年来, 谱聚类在聚类集成中的应用受到了普遍关注[3,10]. 本文利用谱聚类算法对互相关矩阵所描述的样本相似性关系进行聚类以求取集成结果. 具体步骤如下.

算法 2

输入: 互相关矩阵CoMatrix, 尺度参数σ, 类别数K;

输出: 样本的类别划分.

Step 1: 构造相似性度量矩阵W, 即

$$W = \exp\Big(-\frac{\mathrm{dis}(x_i, x_j)}{2\sigma^2}\Big),$$

$$\mathrm{dis}(x_i, x_j) = 1 - \mathrm{CoMatrix}(i, j);$$

Step 2: 构造度矩阵D, 主对角线元素$D(i,i)$为W中第i行元素之和, 其他元素为0, 构造归一化拉普拉斯矩阵$L = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$;

Step 3: 对L进行特征分解, 求出前K个最大特征值所对应的特征向量l并构造矩阵$U = [l_1, l_2, \cdots, l_K]$;

Step 4: 离散化U, 确定样本点类别.

2.4 算法复杂性分析

记模糊度的单位计算时间为t_1, 贴近度的单位计算时间为t_2, 利用模糊度调整类别关系的单位时间为t_3, 证据合成的单位时间为t_4, 则互相关矩阵的计算时间为

$$T = MNt_1 + MN^2(t_2 + t_3) + (M-1)N^2(t_4). \tag{14}$$

可见构造互相关矩阵的时间复杂度为$O(MN^2)$, 与集成规模和样本规模的平方成正比. 谱聚类算法的时间复杂度为$O(N^3)$, 在处理较大规模数据集时同样需要消耗大量的运算时间.

Nystrom采样方法是一种可以有效降低互相关矩阵时间和空间复杂度的方法[3,17], 通过采样数据可将互相关矩阵拆分为采样数据间相互关系和采样数据与剩余数据间相互关系两部分, 进而利用Nystrom逼近的谱聚类算法进行二次聚类以获取集成结果. 设采样规模为m, 则采样情况下构造互相关矩阵的时间为

$$T_{\mathrm{Nystrom}} = MNt_1 + M(m^2 + m(N-m))(t_2 + t_3) +$$

$$(M-1)(m^2+m(N-m))t_4. \tag{15}$$

采样情况下互相关矩阵的时间复杂度可降低到$O(mMN)$, 而且Nystrom逼近的谱聚类[17]算法时间复杂度为$O(m^2N)$, 具有相对较低的运算复杂性, 此时可将算法应用于较大规模的数据分析.

3　实验与分析

为验证本文算法的有效性, 在标准数据集上进行实验. 表4为UCI数据集中的部分数据及描述, 实验过程中对各数据集进行归一化, 并对高维数据使用主成分分析法进行降维.

表 4　实验数据集

数据集	样本规模	特征维数	类别数
Iris	150	4	3
Ecoli	336	7	8
Wine	178	13	3
Thyroid	215	5	3
Soybean	306	35	4
Glass	214	9	6
Diabetes	768	8	2
Sonar	208	60	2
Ionosphere	351	34	2
Segment	2 310	19	7
Sat	2 000	36	6

3.1　实验设计

首先通过互相关矩阵的图形化描述和对比, 分析模糊度对于构造互相关矩阵的意义; 其次讨论集成规模和成员差异性对集成结果的影响, 进而在11个标准数据集上与典型聚类集成算法进行对比以验证本文方法的有效性; 最后就算法效率进行对比.

实验中利用Fowlkes-Mallows指标评价聚类性能, 即

$$\mathrm{FM}=\sqrt{\frac{a}{a+b}\frac{a}{a+c}}. \tag{16}$$

记两次聚类运算分别为π_1和π_2. a表示既在聚类π_1中属于同一类, 又在聚类π_2中属于同一类的样本对个数; b表示在聚类π_1中属于同一类, 但在聚类π_2中不属于同一类的样本对个数; c表示在聚类π_1中不属于同一类, 但在聚类π_2中属于同一类的样本对个数. FM的取值范围是$[0,1]$, 越接近于1表示两聚类结果一致性越高, 越接近于0则表示一致性越低. 实验环境为CPU: Pentium D 3.2 GHz, 内存1 G, 仿真软件为Matlab 7.10.0.

3.2　实验结果与分析

实验中除特别说明外, 聚类成员均由模糊C均值(FCM)通过扰动类别参数和扰动初始中心点生成, 类别参数在$[10,30]$区间内随机选取. 利用算法2对互相关矩阵进行集成时, 尺度参数σ在区间$[0.08,3]$内以0.02为步长遍历取值, 选择其中最优集成结果[4].

3.2.1　互相关矩阵对比

为获取图形化的直观结果, 在Iris、Wine和Ecoli三个数据集上分别利用Hamming贴近度方法、Euclid贴近度方法以及对二者进行模糊度调整(PPE/线性)的方法生成互相关矩阵. 在集成规模为50的情况下, 各方法构造的互相关矩阵如图1所示. 为便于观察, 各数据集中的数据均按类标签顺序化.

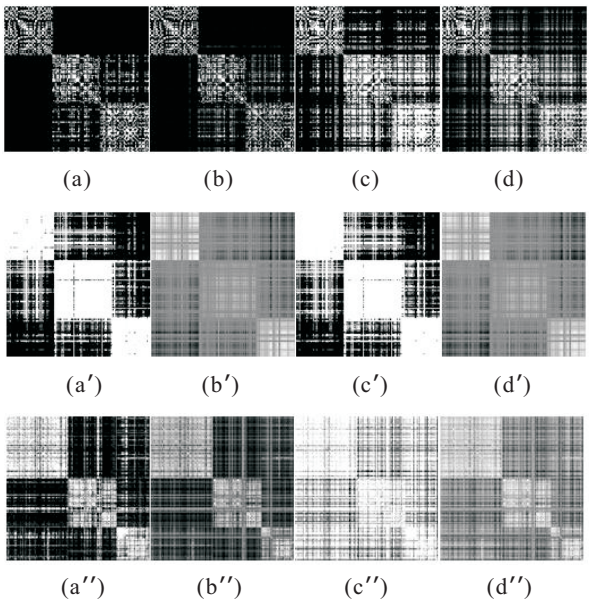

图 1　互相关矩阵对比图

图1中, (a)～(d)为Iris数据集实验结果, (a′)～(d′)为Wine数据集实验结果, (a″)～(d″)为Ecoli数据集实验结果. (a) (a′) (a″)和(c) (c′) (c″)分别为利用Hamming贴近度和Euclid贴近度所构造的互相关矩阵; (b) (b′) (b″)和(d) (d′) (d″)分别为利用模糊度调整的Hamming贴近度和Euclid贴近度所构造的互相关矩阵. 图1中高亮度区域表示样本点间类别相关性较强, 低亮度区域表示类别差异性较强, 灰色区域则表示类别信息不明确. 对比图1中(b) (b′) (b″)和(a) (a′) (a″)以及(d) (d′) (d″)和(c) (c′) (c″)可以看到, 对于模糊度较高的点对, 不仅异类样本点间的同类关系得到了抑制, 而且异类样本点间的异类关系、同类样本点间的同类关系、同类样本点间的异类关系也得到了抑制, 但由于保留了样本点间较合理的同类和异类关系(对比图1中(c″)和(d″)中主对角线上的方形区域), 使得聚类性能有所提高.

为进一步量化描述模糊度调整对于互相关矩阵的意义, 分别利用PPC/线性、PPC/指数、PPC/对数、PPE/线性、PPE/指数、PPE/对数6种方法对上述3个数据集的Hamming贴近度和Euclid贴近度互相关矩

表 5 利用图 1 互相关矩阵的集成结果

Datasets	Hamming	PPC/线性	PPC/指数	PPC/对数	PPE/线性	PPE/指数	PPE/对数
Iris	0.840 7	0.698,9	0.683 4	0.683 2	0.911 5	0.921 5	0.916 7
Wine	0.920 5	0.679 5	0.759 2	0.757 2	0.911 3	0.866 5	0.847 0
Ecoli	0.732 7	0.612 7	0.643 0	0.606 7	0.740 0	0.776 8	0.708 6
Datasets	Euclid	PPC/线性	PPC/指数	PPC/对数	PPE/线性	PPE/指数	PPE/对数
Iris	0.683 2	0.917 5	0.910 3	0.841	0.935 5	0.935 5	0.938 1
Wine	0.871 1	0.680 1	0.776 9	0.734	0.898 9	0.866 5	0.847 0
Ecoli	0.669 2	0.607 8	0.587 2	0.564 5	0.780 3	0.824 8	0.821 1

阵进行调整, 并利用算法 2 获得集成结果, 如表 5 所示. 表 5 中 Hamming、Eculid、PPE/线性所在列对应于图 1 中所示互相关矩阵. 由对比可知, 利用 PPE/线性调整所得的互相关矩阵具有更好的聚类结果 (Wine 数据集的 Hamming 贴近度调整除外). 进一步, 对比表 5 中的 6 种方法可以发现, PPE/线性和 PPE/指数方法通常具有较好的实验结果.

3.2.2 参数讨论与方法分析

本节重点讨论集成规模与成员差异性对于集成结果的影响. 实验包括两部分: 一是固定聚类成员的类别参数范围, 即成员差异性, 研究集成规模与结果间的关系; 二是固定集成规模, 研究成员差异性与集成结果间的关系. 需要指出的是, 本实验仅讨论基于 Hamming 贴近度的多种模糊调整方法, 实际中 Euclid 贴近度也将得到类似的结果.

首先, 固定聚类成员的类别变化范围 [10, 30], 研究规模与集成结果间的关系. 由于生成聚类成员的随机性, 取 20 次实验的平均值, 结果如图 2 所示, 横坐标为集成规模, 纵坐标为 FM 指标. Iris 数据集中, PPC/线性、PPC/指数、PPC/对数 3 条曲线随集成规模的增大而缓慢升高, PPE/线性、PPE/指数、PPE/对数 3 条曲线随集成规模的增大而缓慢下降. Wine 和 Thyroid 数据集中, PPC/线性、PPC/指数、PPC/对数 3 条曲线随集成规模的增大变化较小, 而 PPE/线性、PPE/指数、PPE/对数 3 条曲线随集成规模的增大而缓慢升高. 虽然数据规模与集成结果间的关系随着数据集和方法的不同而差异较大, 但是观察图 2 可以发现, 当集成规模大于一定阈值时 (如 40), 上述 3 个数据集中的 6 条曲线变化均趋于平稳.

固定集成规模为 50, 研究成员差异性与集成结果间的关系, 取 20 次实验的平均值, 结果如图 3 所示, 横坐标为类别扰动范围, 纵坐标为 FM 指标. Iris 数据集中, 6 条曲线随差异性的增大缓慢波动, PPC/线性、PPC/指数、PPC/对数 3 条曲线在 [0, 60] 差异度时聚类结果最差. Wine 数据集中, PPE/线性曲线随差异性的增大先升高后下降, 在 [10, 30] 差异度时取得最

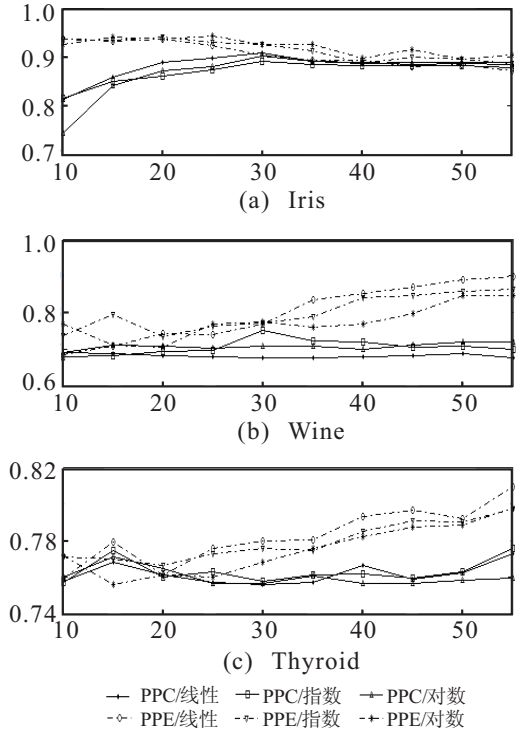

图 2 规模与集成结果间关系

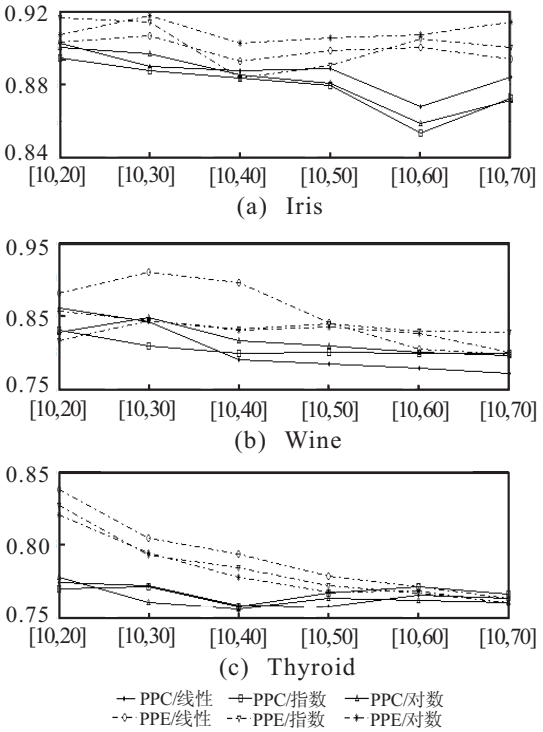

图 3 差异性与集成结果间关系

大值, 其余 5 条曲线整体呈波动下降, 但变化范围不大. Thyroid 数据集中, PPC/线性、PPC/指数、PPC/对数 3 条曲线随差异性增大缓慢波动, 未呈现明显趋势, PPE/线性、PPE/指数、PPE/对数 3 条曲线随差异性的增大而显著下降.

观察两组实验结果, 可以得到如下结论: 1) 对比 3 个数据集中 6 种方法, PPE/线性、PPE/指数、PPE/对数 3 种方法要普遍优于 PPC/线性、PPC/指数、PPC/对数方法; 2) 集成规模与结果并不存在严格对应关系, 但在集成规模较小时, 集成结果波动性较大, 当集成规模足够大时, 集成结果趋于稳定; 3) 成员差异性与集成结果间具有一定联系, 合理的成员差异性有利于形成良好的结果, 过大的差异性将大量的错误分类信息引入集成中, 导致集成性能的下降.

3.2.3　对比实验与分析

为进一步验证本文所提出考虑模糊度的聚类集成方法的有效性, 分别在表 4 中 11 个标准数据集上进行实验, 并与 FCM、CSPA、HGPA、MCLA 方法[1], EAC/AL (Average-Linkage) 方法[4]以及基于 Jacard 测度[8]的模糊聚类集成方法进行对比. 聚类成员的获取和算法 2 中尺度参数的设置均如前文所述. 需要指出的是, 由于 CSPA、HGPA、MCLA、EAC/AL 四种方法只适用于硬聚类集成, 这里按最大隶属度原则硬化模糊聚类结果. 本文方法 AH 和 AE 利用 PPE/线性方法调整 Hamming 和 Euclid 贴近度来获取互相关矩阵并将其用于集成. 前 9 个小规模数据集利用谱聚类算法实现二次聚类, 后 2 个较大规模数据集利用 Nystrom 逼近的谱聚类算法实现二次聚类, 采样规模为 200 (Jacard 方法采用相同的二次聚类方法).

令集成规模为 50, 取 20 次实验的平均值, 结果如表 6 所示, 其中最优聚类结果均用黑体表示. 可以看到本文方法在 11 个数据集中的 8 个取得了最优聚类结果. 对比 FCM 可以看到, 除 Thyroid 数据集外, 本文方法均取得了优于单聚类算法的聚类结果. 对比其他集成算法, 在 Thyroid 和 Glass 两个数据集上, 本文方法劣于 EAC/AL; 在 Sonar 数据集上本文方法劣于 Jacard. 需要指出, 本文方法中, 聚类成员的有效性、差异性以及样本点的模糊度是影响集成结果的重要因素. 较低的聚类成员有效性 (如 Glass 数据集) 通过集成虽然能够获得聚类性能的提升, 但空间不大, 对于数据划分的意义并不明显. 聚类集成要求聚类成员在错分样本的类别划分上应表现出一定的差异性, 当这些差异不明显或表现在正确划分的样本点上时 (如 Thyroid 数据集), 将导致集成性能的下降. 本文方法中模糊度被作为调整样本点间类别关系的重要依据, 而 Sonar 数据集中, FCM 聚类结果具有极高的模糊度, 这使得调整过程中丢失大量相关信息, 因而集成结果不甚理想. 综上所述, 与已有聚类集成方法相比, 本文方法能够得到更好的集成结果, 但当聚类成员有效性较差、成员差异性不合理或者聚类结果模糊度过高时, 集成结果可能不理想.

3.2.4　运算时间对比

针对 Iris、Diabetes、Segment 三个规模差异较大的数据集进行实验, 讨论算法的时间复杂性. 本文算法 AH 和 AE (如上节所述) 在 Iris 和 Diabetes 数据集上利用谱聚类进行二次聚类, 在 Segment 数据集上利用 Nystrom 逼近的谱聚类进行二次聚类, 并与 CSPA、HGPA、MCLA、EAC/AL 等算法进行对比. 仅考虑集成过程的运算时间, 结果如表 7 所示. 因为 CSPA 和 EAC/AL 的运算时间分别与 N^2 和 N^3 成正比, 开销较大, 所以对于较大规模数据集 Segment 没有给出运算时间.

表 6　集成结果对比

Datasets	Single	Ensemble methods compared					Method proposed	
	FCM	CSPA	HGPA	MCLA	EAC/AL	Jacard	AH	AE
Iris	0.818 8	0.843 0	0.729 3	0.783 5	0.834 7	0.775 1	0.917 5	**0.922 1**
Ecoli	0.503 0	0.543 7	0.574 1	0.537 9	0.779 9	0.649 0	0.701 3	**0.780 6**
Wine	0.900 1	0.841 8	0.702 1	0.898 5	0.898 9	0.782 1	**0.900 3**	0.896 8
Thyroid	**0.875 9**	0.477 1	0.576 9	0.491 2	0.832 8	0.769 7	0.812 2	0.810 2
Soybean	0.521 8	0.437 5	0.429 4	0.512 3	0.616 4	0.645 6	**0.647 1**	0.611 6
Glass	0.353 0	0.324 5	0.327 8	0.335 0	**0.550 7**	0.387 5	0.384 2	0.527 2
Diabetes	0.581 9	0.592 4	0.524 9	0.581 1	0. 593 9	0.654 3	**0.734 8**	0.734 2
Sonar	0.502 3	0.500 5	0.498 1	0.502 2	0.502 3	**0.692 3**	0.568 2	0.567 0
Ionosphere	0.603 1	0.573 7	0.602 8	0.601 2	0.611 8	0.656 5	0.677 8	**0.683 7**
Segment	0.576 8	—	0.386 7	0.578 5	0.610 9	0.605 9	0.586 8	**0.613 0**
Sat	0.608 0	—	0.346 4	0.531 4	0.625 7	0.610 0	0.611 9	**0.630 4**

表 7 运算时间对比 s

Datasets	规模/维数/类别数	CSPA	HGPA	MCLA	EAC/AL	AH	AE
Iris	150/4/3	0.120 2	0.521 7	0.393 7	9.765 3	1.108 7	1.222 5
Diabetes	768/8/2	0.728 5	0.409 3	0.869 3	107 1.6	30.895 1	33.456 2
Segment	2310/19/7	—	1.453 6	2.129 8	—	27.338 0	29.226 2

观察表 7 可以看到: 对于小规模数据集 Iris, 本文方法运算时间略高于 CSPA、HGPA 和 MCLA, 但低于 EAC/AL; 对于较大规模数据集 Diabetes, 由于利用谱聚类算法对互相关矩阵进行聚类, 本文方法运算时间显著提高. 但是在大规模数据集 Segment 的集成过程中, 由于采用 Nystrom 采样方法, 使得运算时间有所下降. 虽然本文方法在各数据集上的运算时间均高于 HGPA 和 MCLA, 但考虑到相对较高的集成精度 (如上节讨论), 这种时间开销仍具有一定的实际意义.

4 结 论

集成方法已成为当前聚类领域的研究热点. 模糊聚类成员间共识关系设计以及模糊关系的证据积累是将集成方法扩展到模糊聚类中的关键问题. 为此, 本文提出了一种结合模糊度和模糊贴近度的类别关系表示方法, 并利用证据理论实现模糊类别关系的证据积累. 实验结果表明, 本文方法与经典方法相比, 不仅可以构造模糊聚类集成的共识函数, 而且具有更好的聚类性能.

参考文献(References)

[1] Strehl A, Ghosh J. Cluster ensembles — A knowledge reuse framework for combining multiple partitions[J]. J of Machine Learning Research, 2002, 3(3): 583-617.

[2] 卢志茂, 李纯, 张琦. 近邻传播的文本聚类集成谱算法[J]. 哈尔滨工程大学学报, 2012, 33(7): 899-905.
(Lu Z M, Li C, Zhang Q. A document cluster ensemble spectral algorithm based on affinity propagation[J]. J of Harbin Engineering University, 2012, 33(7): 899-905.)

[3] 周林, 平西建, 徐森, 等. 基于谱聚类的聚类集成算法[J]. 自动化学报, 2012, 38(8): 1335-1342.
(Zhou L, Ping X J, Xu S, et al. Cluster ensemble based on spectral clustering[J]. Acta Automatica Sinica, 2012, 38(8): 1335-1342.)

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

[5] 王羡慧, 覃征, 张选平, 等. 采用仿射传播的聚类集成算法[J]. 西安交通大学学报, 2011, 45(8): 1-6.
(Wang X H, Qin Z, Zhang X P, et al. Cluster ensemble algorithm using affinity propagation[J]. J of Xi'an Jiaotong University, 2011, 45(8): 1-6.)

[6] 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 W, Zhou Z H. Bagging-based selective cluster ensemble[J]. J of Software, 2005, 16(4): 496-502.)

[7] Alexander T, Anil K J, William P. Clustering ensembles: Models of consensus and weak partitions[J]. IEEE Trans on Pattern Analsis and Machine Intelligence, 2005, 27(12): 1866-1881.

[8] Yang L Y, Lü H R, Wang W Y. Soft cluster ensemble based on fuzzy similarity measure[C]. IMACS Conf on Computational Engineering in Systems Applications. Beijing, 2006: 1994-1997.

[9] Wang H S, Yang Y, Wang H J, et al. Soft-voting clustering ensemble[C]. Lecture Note in Computer Science, Multiple Classifier Systems. Nanjing, 2013: 307-318.

[10] Jia J H, Liu B X, Jiao L C. Soft spectral clustering ensemble applied to image segmentation[J]. Front Computer Science China, 2011, 5(1): 66-78.

[11] Mok P Y, Huang H Q, Kwok Y L, et al. A robust adaptive clustering analysis method for automatic identification of clusters[J]. Pattern Recognition, 2012, 45(8): 3017-3033.

[12] Sevillano X, Alias F, Socoro J C. Positional and confidence voting-based consensus functions for fuzzy cluster ensembles[J]. Fuzzy Sets and Systems, 2012, 193: 1-32.

[13] Dezert J, Wang P, Tchamova A. On the validity of Dempster-Shafer theory[C]. Proc of the 15th Int Conf on Information Fusion. Singapore, 2012: 655-660.

[14] Kwang H L, Song Y S, Lee K M. Similarity measure between fuzzy sets and between elements[J]. Fuzzy Sets and Systems, 1994, 62: 291-293.

[15] 何建华, 刘耀林, 俞艳, 等. 基于模糊贴近度分析的不确定拓扑关系表达模型[J]. 测绘学报, 2008, 37(2): 212-216.
(He J H, Liu Y L, Yu Y, et al. The topological relation model for indeterminate geographical objects based on fuzzy close-degree[J]. Acta Geodaetica et Cartographica Sinica, 2008, 37(2): 212-216.)

[16] 杨纶标, 高英仪. 模糊数学原理及应用[M]. 广州: 华南理工大学出版社, 2005: 31-33.
(Yang L B, Gao Y Y. The principle and application of fuzzy mathematics[M]. Guangzhou: South China University of Technology Publisher, 2005: 31-33.)

[17] Charless F, Serge B, Fan C, et al. Spectral grouping using the nystrom method[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2004, 26(2): 214-225.

(责任编辑: 李君玲)