# 基于链接的模糊聚类集成方法

杨 燕，冯晨菲，贾 真，王红军

(西南交通大学信息科学与技术学院 成都 610031)

**【摘要】**针对多数聚类集成方法忽视潜在信息或获取潜在信息方法复杂这一缺点，提出一种基于链接的模糊聚类集成方法。该算法首先利用模糊聚类算法建立集成信息矩阵，然后使用相应的链接方法将集成信息矩阵转化为反映数据相关性的权重图，最后运用图划分技术得到最终结果。实验结果表明，新提出的算法可以有效地获取潜在信息，同时提高聚类质量。

**关 键 词** 聚类集成； 模糊聚类； 链接； 潜在信息

**中图分类号** TP391 **文献标志码** A doi:10.3969/j.issn.1001-0548.2014.06.016

# A Link-Based Fuzzy Clustering Ensemble

YANG Yan, FENG Chen-fei, JIA Zhen, and WANG Hong-jun

(School of Information Science and Technology, Southwest Jiaotong University Chengdu 610031)

**Abstract** A link-based fuzzy cluster ensemble (LBFCE) is proposed to solve the problem that many clustering ensemble methods ignore the underlying information or acquire the underlying information by complex approaches. In the LBFCE, an ensemble information matrix is first built by primarily exploiting the results of fuzzy clustering, this matrix is then transformed into a weighted graph with data relations by appropriate link analysis, and at last a graph partitioning algorithm is employed to get the final clustering results. Experimental results show that the LBFCE algorithm may obtain the underlying information effectively and improve clustering performance.

**Key words** clustering ensemble; fuzzy clustering; link; underlying information

聚类是人们了解数据对象结构的基础性工具，它广泛应用于生物学、Web搜索及文本挖掘[1]等领域。数据对象根据最大化簇内相似、最小化簇间相似的原则进行聚类。尽管目前人们已经提出了许多的聚类算法及改进算法，但文献[2]中的“No Free Lunch”理论揭示了：没有哪一种单一的、超强的聚类算法能够发现任意形状和结构的簇；在给定一个数据集时，寻找适合于该数据集的聚类算法是一件较为困难的事情。

聚类集成将不同算法或者同一算法下使用不同的参数得到的结果进行合并，从而得到比单一算法更为优越的结果。文献[3]指出与单一聚类算法相比，聚类集成算法能够找到单一聚类算法难以找到的聚类结果；同时，降低噪声、孤立点等对聚类结果的影响，增强结果的鲁棒性和稳定性。文献[4]提出基于共联矩阵的EA(evidence accumulation)方法，该方法首先得到聚类成员的共联矩阵，然后采用基于MST(minimum spanning tree)的分级聚类算法得到最终的聚类结果。文献[5]在图分割思想的基础上提出了CSPA、HGPA、MCLA三种基于超图的算法。文献[6]通过建立投票机制解决聚类集成问题，根据投票结果得到聚类集成结果。文献[7]对蚁群算法进行改进并提出一种基于多蚁群算法的聚类集成方法。文献[8]在利用选择性聚类集成的多种有效性指标间的差异性的基础上提出一种基于相对评价指标的选择性聚类集成方法。文献[9]提出一种基于聚类集成的两阶段无监督分割方法并应用该算法从图像中提取感兴趣的区域。在利用模糊聚类算法作为基聚类器的模糊聚类集成方面的研究也卓有成效。文献[10]将三种模糊度量准则运用于模糊聚类集成，取得了较好的集成结果。文献[11]提出一种软投票聚类集成方法，该方法首先将模糊聚类算法结果标签对齐后对应相乘，然后将所得结果做归一化和硬化处理。不同于上述方法，文献[12]提出基于链接的聚类集成方法，该方法指出传统的集成方法在构建集成信息矩阵时忽略了簇与簇之间的相关性这一潜在信

收稿日期：2013 – 12 – 02；修回日期：2014 – 07 – 23
基金项目：国家自然科学基金(61170111, 61134002, 61003142)； 四川省科技支撑计划(2014S20207)
作者简介：杨燕(1964 –)，女，教授，博士生导师，主要从事数据挖掘、计算智能、集成学习方面的研究.

息。针对这一问题提出了计算簇间相关性的算法，并将簇间相关性转换为数据与簇之间的相关性，最后取得了较好的结果，然而该获取潜在信息的算法较为复杂。

本文提出一种基于链接的模糊聚类集成方法(link-based fuzzy clustering ensemble，LBFCE)，该算法将表示数据与簇相关性的集成信息矩阵转化为反映数据之间相关性的权重图，这一设计简洁而有效地获取了潜在信息，在此基础上运用图划分技术使得最终的聚类质量得到提升。

# 1 聚类原理

聚类是一个极富挑战性的研究领域，聚类的输入可以用一组有序对 $(X,s)$ 来表示，这里 $X$ 表示数据对象集合， $s$ 是度量样本相似性的标准。聚类的输出是一个分区，可表示为 $\pi=\{C_1,C_2,\cdots,C_i\}$，其中 $C_i(i=1,2,\cdots,c)$ 是 $X$ 的子集，每个子集是一个簇，且满足 $\bigcup_{i=1}^{c}C_i=X$ 与 $C_i\cap C_j=\phi, i\neq j$。聚类的目标是使一个簇内的数据尽量相似而不同簇的数据尽量不相似。根据聚类算法的主要思路，文献[13]将聚类分为基于划分的方法、基于层次的方法、基于密度的方法和基于网格的方法4个类别。其中基于划分的方法应用较为广泛。

## 1.1 模糊C均值聚类算法

由文献[14]首次提出，文献[15]对其改进的模糊C均值聚类算法(fuzzy C-means，FCM)是用隶属度表示每一个数据对象隶属于某个簇程度的一个经典的基于划分的聚类学习算法。

FCM算法目标函数的一般形式为：

$$J(U,c_1,c_2,\cdots,c_c)=\sum_{i=1}^{c}J_i=\sum_{i=1}^{c}\sum_{j=1}^{n}u_{ij}^{m}\left\|c_i-x_j\right\|^2 \quad (1)$$

式中， $u_{ij}$ 表示数据对象 $j$ 属于簇 $i$ 的隶属度且满足条件， $\sum_{i=1}^{c}u_{ij}=1$， $c_i$ 表示第 $i$ 个簇的中心， $c$ 表示簇的个数； $x_j$ 表示第 $j$ 个数据对象； $m\in[1,\infty)$ 是加权指数。

由于目标函数的求解较复杂，这里给出最小化式(1)的两个必要条件：

$$c_i=\frac{\sum_{j=1}^{n}u_{ij}^{m}x_j}{\sum_{j=1}^{n}u_{ij}^{m}} \quad (2)$$

$$u_{ij}=\frac{1}{\sum_{k=1}^{c}\left(\frac{\left\|c_i-x_j\right\|}{\left\|c_k-x_j\right\|}\right)^{\frac{2}{m-1}}} \quad (3)$$

由上述两个条件可知，FCM算法是一个简单的迭代过程。这一过程可以表示为通过不断迭代式(2)与式(3)来更新聚类中心 $c_i$ 和隶属度 $u_{ij}$，再根据式(1)计算目标函数。如果目标函数值小于某个阈值或者它相对于上次目标函数的改变量小于某个阈值，则算法停止。

## 1.2 聚类集成

聚类集成主要包括两个阶段：1) 对 $N$ 个原始数据对象集合 $X=\{x_1,x_2,\cdots,x_N\}$ 用初始化不同的同种聚类算法运行 $M$ 次得到 $M$ 个有差异性的聚类结果，或者采用几种聚类算法得到这 $M$ 个有差异性的结果，表示为 $\prod=\{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2,\cdots,\boldsymbol{\pi}_M\}$。这里每个基聚类结果是簇的集合 $\pi_i=\{C_1^i,C_2^i,\cdots,C_{c^i}^i\}$ 表示在第 $i$ 次聚类中簇的个数为 $c^i$ 且满足 $\bigcup_{j=1}^{c^i}C_j^i=X$； 2) 共识函数设计阶段，其目的是对 $M$ 个有差异性的聚类结果 $\prod$ 进行融合，得到对于数据集合 $X$ 最终的集成结果 $\boldsymbol{\pi}^*$，一般来说集成结果 $\boldsymbol{\pi}^*$ 比单一聚类算法的结果要好。聚类集成示意图如图1所示。

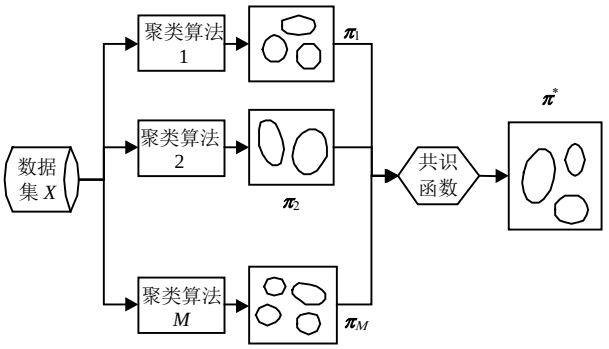


图1 聚类集成示意图

## 1.3 基于链接的聚类集成

数据对象之间的关系也即对象之间的相似性，相似性可以通过潜在的链接信息来获得。多种基于链接的相似性准则在Web文本分类上的应用已经比较成熟[16]。文献[17]成功地将基于链接的相似性准则运用在聚类上，提出多种基于链接的聚类集成方法。其原理都是先计算簇与簇之间的相关性，再将其转化为数据对象与簇之间的相关性进而表示为集成信息矩阵，最后采用相应的图划分技术得到最后结果。考虑到现有的模糊聚类算法的结果可以直接转化为表示数据对象与簇连接关系的二分图，但这

个二分图并不能体现聚类结果的融合信息。针对这一缺陷，本文设计一种链接方法将集成信息矩阵转化为体现信息融合和数据关系的权重图，再对权重图进行划分，详细算法将在第2节给出。

## 2　基于链接的模糊聚类集成方法

### 2.1　LBFCE算法

基于链接的模糊聚类集成算法(LBFCE)的第1步是构建集成信息矩阵，该信息矩阵实质上是一个表述数据对象与簇关系的二分图。假设对数据集 $X=\{x_1,x_2,\cdots,x_N\}$ 用初始化不同的FCM算法运行 $M$ 次得到 $M$ 个有差异性的结果，该结果的集合表示为 $\prod=\{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2,\cdots,\boldsymbol{\pi}_M\}$。基聚类结果 $\boldsymbol{\pi}_i(i=1,2,\cdots,M)$ 是 $c$ 个簇的集合，体现为一个 $N\times c$ 的矩阵。利用结果集将集成信息矩阵表示为 $\boldsymbol{\psi}(M)=[\boldsymbol{\pi}_1,\boldsymbol{\pi}_2,\cdots,\boldsymbol{\pi}_M]$，这里 $\boldsymbol{\psi}(M)$ 是一个 $N\times Mc$ 的矩阵，$Mc$ 表示簇的总个数。集成信息矩阵可以转化为二分图 $G=(V,E)$，顶点 $V$ 可分割为两个互不相交的子集 $(A,B)$，其中 $A$ 指数据对象，$B$ 表示簇，$E$ 代表隶属度。

为了不失一般性，随机给定一个数据对象集 $X'=\{x_1,x_2,x_3\}$，基聚类器个数 $M'=2$，基聚类簇个数 $c'=2$，假设聚类结果构成的集成信息矩阵表示如下：

$$\boldsymbol{\psi}(2)=\begin{pmatrix}0.6 & 0.4|0.7 & 0.3\\ 0.8 & 0.2|0.5 & 0.5\\ 0.1 & 0.9|0.2 & 0.8\end{pmatrix} \tag{4}$$

其对应的二分图如图2所示。

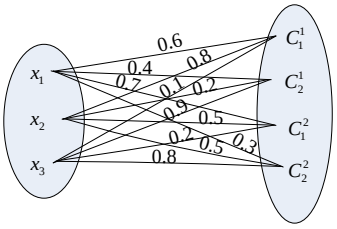


图2　表示数据对象与簇关系的二分图

集成信息矩阵并不能体现聚类结果的融合信息。针对这一缺陷，本文设计一种链接方法，基本思想是将集成信息矩阵转化为表示数据之间相关性的矩阵。该方法将基聚类结果进行巧妙融合，具体表述如下：

基聚类结果 $\boldsymbol{\pi}_i$，$i=1,2,\cdots,M$ 是一个 $N\times c$ 的矩阵，用 $\boldsymbol{\pi}_i(j)$，$j=1,2,\cdots,N$ 表示该矩阵的第 $j$ 行，那么 $\boldsymbol{\pi}_i(j)$ 实际上是一个 $1\times c$ 的矩阵，采用以下公式计算得到表示数据对象相关性的矩阵 $\boldsymbol{W}$。

$$\boldsymbol{W}(l,n)=\begin{cases}\sqrt{\dfrac{\sum_{i=1}^{M}\boldsymbol{\pi}_i(l)*\boldsymbol{\pi}_i^{\mathrm{T}}(n)}{M}} & l\neq n\\ 1 & l=n\end{cases} \tag{5}$$

式(5)中要求 $l=1,2,\cdots,N$ 与 $n=1,2,\cdots,N$，上标T表示矩阵的转置。通过式(5)将聚类结果信息进行融合，融合后的结果将转化为表示数据相关性的权重图。

在该事例中，利用式(5)对表示集成信息矩阵的式(4)进行处理得到如下相关性矩阵：

$$\boldsymbol{W}'=\begin{pmatrix}1 & 0.594\,4 & 0.516\,4\\ 0.594\,4 & 1 & 0.503\,3\\ 0.516\,4 & 0.503\,3 & 1\end{pmatrix} \tag{6}$$

矩阵 $\boldsymbol{W}'$ 表示数据对象之间的链接，体现了数据对象之间的相关性。$\boldsymbol{W}'$ 可以转化为表示数据对象之间权重图如图3所示。

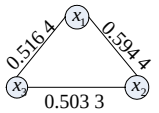


图3　数据对象相关性的权重图

在获得表示数据对象相关性的权重图之后，接下来的任务便是对这个权重图进行划分。本文使用METIS[18]图划分算法对权重图进行基于图论的聚类，得到最终的集成结果 $\boldsymbol{\pi}^*$。

LBFCE算法具体总结如下。

输入：数据集 $X=\{x_1,x_2,\cdots,x_N\}$，簇个数为 $c$

输出：数据对象标签集 $\boldsymbol{\pi}^*$

1) 对数据集 $X$ 运行 $M$ 个不同初始条件的FCM聚类算法，得到 $M$ 个有差异性的结果集 $\prod=\{\boldsymbol{\pi}_1,\boldsymbol{\pi}_2,\cdots,\boldsymbol{\pi}_M\}$；

2) 将结果集 $\prod$ 转化为表示数据与簇相关性的集成信息矩阵 $\boldsymbol{\psi}(M)=[\boldsymbol{\pi}_1,\boldsymbol{\pi}_2,\cdots,\boldsymbol{\pi}_M]$；

3) 利用式(5)将集成信息矩阵 $\boldsymbol{\psi}(M)$ 转化为表示数据对象之间相关性的矩阵 $\boldsymbol{W}$；

4) 利用METIS算法对矩阵 $\boldsymbol{W}$ 进行划分，得到最终的集成结果 $\boldsymbol{\pi}^*$。

### 2.2　关于标签对齐问题

标签对齐问题源于在多个基聚类划分结果中对于相同的簇使用不同的数字进行标注。例如，标签 $L_1=[1,1,2,1,3,3]^{\mathrm{T}}$ 与标签 $L_2=[2,2,3,2,1,1]^{\mathrm{T}}$ 实质上表示的是相同的聚类结果。如何将基聚类划分结果转化成统一类标后的结果就是标签对齐所要解决的问题。针对这一问题，许多学者给出了相应的处理方法(当然有些集成算法不需要进行标签对齐，如文献

[5]等)。LBFCE算法不需要对基聚类结果进行标签对齐处理，具体解释如下。

如前所述，LBFCE算法首先将 $M$ 个未进行标签对齐的基聚类结果 $\boldsymbol{\pi}_i$ ( $i=1,2,\cdots,M$ )转化为集成信息矩阵 $\boldsymbol{\psi}(M)$，然后利用式(7)将潜在信息进行融合。在式(5)的计算中 $\boldsymbol{\pi}_i(l)$ 与 $\boldsymbol{\pi}_i(n)$ 所表示的是第 $i$ 个基聚类结果中第 $l$ 行与第 $n$ 行的数据向量。假设 $\boldsymbol{\pi}_i(l)=[\alpha_1,\alpha_2,\cdots,\alpha_c]$， $\boldsymbol{\pi}_i^{\mathrm{T}}(n)=[\beta_1,\beta_2,\cdots,\beta_c]^{\mathrm{T}}$，于是有：

$$\boldsymbol{\pi}_i(l)*\boldsymbol{\pi}_i^{\mathrm{T}}(n)=\sum_{q=1}^{c}\alpha_q*\beta_q \tag{7}$$

由式(7)可知无论参数 $\alpha_q$ 与 $\beta_q$ 在向量 $\boldsymbol{\pi}_i(l)$ 与 $\boldsymbol{\pi}_i^{\mathrm{T}}(n)$ 中的位置发生怎样的变化，都不会影响式(7)的求和结果，而模糊聚类结果的标签对齐过程的实质就是矩阵 $\boldsymbol{\pi}_i$ 中列与列互换的过程，因此本文算法不需要对基聚类结果进行标签对齐，这也是本文算法的优点之一。

# 3 实验结果与分析

## 3.1 测试数据集选取

实验选用人工数据集，UCI数据集[19]，KDDCup99[20]三种来源的10个测试数据对本文算法进行评价。4D3K和8D5K[21]数据集是基于高斯分布模型随机产生的人工数据；中间7组来源于UCI的真实数据集，其中cmc是数据Contraceptive-method-choice的缩写；Kdd99sub是KDDCup99的一个子集。所有测试数据集的相关信息如表1所示。

**表1 实验测试数据集相关信息描述**

| 数据集 | 样本个数 | 属性 | 分类数 | 来源 |
|---|---|---|---|---|
| 4D3K | 80 | 4 | 3 | 人工 |
| 8D5K | 1 000 | 8 | 5 | 人工 |
| iris | 150 | 4 | 3 | UCI |
| wine | 178 | 13 | 3 | UCI |
| cmc | 1 473 | 9 | 3 | UCI |
| glass | 214 | 9 | 6 | UCI |
| segment | 2 310 | 19 | 7 | UCI |
| Heart-statlog | 270 | 12 | 2 | UCI |
| vehicle | 846 | 18 | 4 | UCI |
| Kdd99sub | 1 280 | 41 | 3 | KDDCup99 |

## 3.2 实验设计

设置聚类成员个数 $M$ =10，运行 $M$ 次FCM算法，每一次随机产生隶属度矩阵 $\boldsymbol{U}$，并使用式(2)计算 $c$ 个初始聚类中心，模糊程度加权指数 $m=2$，最大迭代次数100次，迭代终止条件 $\xi\leqslant 1\times e^{-5}$。由于初始聚类中心的随机性，导致单次集成结果的偶然性。为了降低结果的偶然性，最终的取值为集成算法运行次数 $H=20$ 结果的平均值。

实验将本文算法与7种集成算法进行比较。7种集成算法的来源分别是：文献[5]中基于超图的聚类集成算法CSPA、HGPA、MCLA，文献[11]中提出的SVCE算法以及文献[17]中阐述的三种在完全链接(CL)下的基于CTS、SRS、ASRS三种相似性准则的聚类集成算法。这里使用文献[17]中的三种算法而非文献[12]中的单种算法的原因在于文献[12]中所提出的算法为基于CTS相似性准则算法的一个特例，因此本文使用相似性准则更为广泛的文献[17]中的算法。值得注意的是由于CSPA、HGPA、MCLA三种算法不接受模糊聚类结果，因此需要将基聚类结果做硬化处理。

## 3.3 评价标准与实验结果分析

实验采用外部评价标准RI(rand index)[22]和相对评价标准Ocq(overall cluster quality)[23]对8种情况的聚类结果进行评价。RI有如下定义：

$$\mathrm{RI}(\boldsymbol{\pi}^*,\prod{}')=\frac{n_{11}+n_{00}}{n_{11}+n_{10}+n_{01}+n_{00}} \tag{8}$$

式中，$\prod{}'$ 表示真实标签集；$n_{11}$ 指数据对象对在集合 $\boldsymbol{\pi}^*$ 中在同一个簇中且在 $\prod{}'$ 集合中也在同一个簇的数目；$n_{00}$ 指在数据对象对在集合 $\boldsymbol{\pi}^*$ 中在不同簇中且在 $\prod{}'$ 集合中也在不同簇的数目，同理可推知 $n_{10}$ 与 $n_{01}$ 所表示的含义。

聚类综合质量Ocq是将聚类密集性与聚类邻近性组合后的一种评价标准，其定义为：

$$\mathrm{Ocq}(\theta)=1-[\theta\times\mathrm{Cmp}+(1-\theta)\times\mathrm{Prox}] \tag{9}$$

式中，$\theta\in[0,1]$ 是平衡聚类密集性与聚类邻近性的权值，$\theta=0.5$ 表示两种评价有相等的权值；Cmp 代表聚类密集性，它的定义为：

$$\mathrm{Cmp}=\frac{1}{c}\sum_{i=1}^{c}[\mathrm{var}(c_i)/\mathrm{var}(X)] \tag{10}$$

式中，$c$ 为簇个数；$\mathrm{var}(c_i)$ 表示簇 $c_i$ 的方差；$\mathrm{var}(X)$ 是数据集 $X$ 的方差。聚类邻近性被定义为：

$$\mathrm{Prox}=\frac{1}{[c(c-1)]}\sum_{i=1}^{c}\sum_{j=1,j\neq i}^{c}\exp[-d^2(x_{c_i},x_{c_j})/2\sigma^2] \tag{11}$$

式中，$\sigma$ 为高斯常数，为简化计算，取 $2\sigma^2=1$；$x_{c_i}$ 表示簇 $c_i$ 的中心；$d(x_{c_i},x_{c_j})$ 表示簇 $c_i$ 与 $c_j$ 中心之间的距离。表2与表3分别给出运行20次各种算法后每个数据集上的平均RI值与平均Ocq值。最好的结果加粗标记，未获取的值用“N/A”表示。

由表2可以看出，在10个数据集上，LBFCE算法有8次获得了最好的平均RI值，在另外两个数据集上也取得了较好的结果(均为第3位)。从表3可知

LBFCE算法在4个数据集上取得了最好的平均Ocq值，在其他数据集上的表现也较为良好。综合表2和表3的结果所得：LBFCE算法通过有效利用潜在信息能取得较好的集成结果。

**表2　各种算法的平均RI值**

| 数据集 | CSPA | HGPA | MCLA | SVCE | CTS-CL | SRS-CL | ASRS-CL | LBFCE |
|---|---|---|---|---|---|---|---|---|
| 4D3K | 0.848 1 | 0.707 4 | 0.877 5 | 0.877 5 | 0.750 7 | 0.876 1 | 0.816 0 | 0.922 8 |
| 8D5K | 1 | 0.669 5 | 1 | 1 | 1 | 1 | 1 | 1 |
| iris | 0.879 7 | 0.735 6 | 0.879 7 | 0.885 9 | 0.808 9 | 0.810 6 | 0.790 9 | 0.912 4 |
| wine | 0.685 0 | 0.618 5 | 0.710 5 | 0.705 5 | 0.545 7 | 0.587 7 | 0.579 9 | 0.711 8 |
| cmc | 0.554 0 | 0.549 9 | 0.558 2 | 0.558 5 | 0.452 1 | 0.530 5 | 0.480 2 | 0.559 5 |
| glass | 0.705 0 | 0.691 5 | 0.706 3 | 0.707 0 | 0.570 6 | 0.611 9 | 0.614 4 | 0.736 7 |
| segment | 0.837 5 | 0.818 1 | 0.838 3 | 0.744 4 | 0.683 4 | 0.825 9 | 0.721 8 | 0.834 2 |
| Heart-statlog | 0.512 7 | 0.498 6 | 0.515 3 | 0.515 3 | 0.504 4 | 0.509 8 | 0.505 6 | 0.516 7 |
| vehicle | 0.659 6 | 0.625 2 | 0.649 7 | 0.656 2 | 0.570 2 | 0.572 4 | 0.589 0 | 0.650 7 |
| KddCup99sub | 0.712 9 | 0.695 2 | 0.672 7 | 0.718 3 | 0.448 1 | 0.595 5 | 0.522 4 | 0.725 9 |

**表3　各种算法的平均Ocq值**

| 数据集 | CSPA | HGPA | MCLA | SVCE | CTS-CL | SRS-CL | ASRS-CL | LBFCE |
|---|---|---|---|---|---|---|---|---|
| 4D3K | 0.772 2 | 0.576 2 | 0.820 3 | 0.817 6 | 0.776 9 | 0.815 3 | 0.798 7 | 0.783 8 |
| 8D5K | 0.644 3 | 0.003 8 | 0.644 3 | 0.644 1 | 0.644 3 | 0.644 3 | 0.644 3 | 0.644 3 |
| iris | 0.780 8 | 0.588 1 | 0.810 4 | 0.808 4 | 0.759 6 | 0.786 6 | 0.770 8 | 0.807 6 |
| wine | 0.716 9 | 0.604 4 | 0.795 8 | 0.792 4 | 0.767 9 | 0.769 6 | 0.765 7 | 0.796 7 |
| cmc | 0.635 1 | 0.522 1 | 0.685 7 | 0.686 2 | 0.654 7 | 0.658 3 | 0.665 1 | 0.687 7 |
| glass | 0.583 9 | 0.549 8 | N/A | 0.624 1 | 0.705 7 | 0.708 0 | 0.703 7 | 0.648 4 |
| segment | 0.747 4 | 0.705 2 | 0.744 9 | 0.785 2 | 0.755 6 | 0.713 8 | 0.741 5 | 0.762 2 |
| Heart-statlog | 0.566 8 | 0.501 1 | 0.585 1 | 0.581 1 | 0.572 9 | 0.550 8 | 0.559 3 | 0.589 6 |
| vehicle | 0.692 9 | 0.500 3 | N/A | 0.772 8 | 0.759 4 | 0.742 0 | 0.755 8 | 0.714 4 |
| KddCup99sub | 0.679 2 | 0.592 3 | 0.638 3 | –0.391 9 | 0.786 0 | 0.795 4 | 0.773 2 | 0.699 3 |

## 4　结 束 语

本文提出一种基于链接的模糊聚类集成方法LBFCE。实验结果表明，这种基于链接的模糊聚类集成算法能有效获取潜在信息，同时提升聚类质量。该算法通过有针对性地设计链接方法，避免了在处理基聚类划分结果时进行标签对齐。在今后的工作中将考虑加入半监督信息来改善集成效果，同时也将考虑进一步提高算法的效率。

本文还得到西南交通大学牵引动力国家重点实验室自主研究课题(2012TPL_T15)的支持，在此表示感谢。

## 参 考 文 献

[1] 朱君, 曲超, 汤庸. 利用单词超团的二分图文本聚类算法[J]. 电子科技大学学报, 2008, 37(3): 439-442.
ZHU Jun, QU Chao, TANG Yong. Clustering algorithm of bipartite graph partition based on word hyperclique[J]. Journal of University of Electronic Science and Technology of China, 2008, 37(3): 439-442.

[2] WOLPERT D H, MACREADY W G. No free lunch theorems for search[R]. Technical Report SFI-TR -9502010, Santa Fe Institute, 1995.

[3] TOPCHY A, JAIN A K, PUNCH W. A mixture model for clustering ensembles[C]//Proceedings of the 4th SIMA International Conference on Data Mining. Florida: [s.n.], 2004: 379-390.

[4] FRED A, JAIN A K. Data clustering using evidence accumulate[C]//Proceedings of the 16th International Conference on Pattern Recognition. Quebec: [s.n.], 2002, 4: 276-280.

[5] STREHL A, GHOSH J. Cluster ensembles: a knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research, 2003, 3(3): 583-617.

[6] ZHOU Zhi-hua. Ensemble methods: foundations and algorithms[M]. Boca Raton: CRC Press, 2012.

[7] YANG Y, KAMEL M. An aggregated clustering approach using multi-ant colonies algorithms[J]. Pattern Recognition, 2006, 38(7): 1278-1289.

[8] NALDI M C, CARVALHO A C P L F, CAMPELLO R J G B. Cluster ensemble selection based on relative validity indexes[J]. Data Mining and Knowledge Discovery, 2013, 27(2): 259-289.

[9] RAFIEE G, DLAY S S, WOO W L. Region-of-interest extraction in low depth of field images using ensemble clustering and difference of Gaussian approaches[J]. Pattern Recognition, 2013, 46(10): 2685-2699.

[10] YANG Lin-yun, LV Hai-rong, WANG Wen-yuan. Soft cluster ensemble based on fuzzy similarity measure[C]// IMACS Multiconference on Computational Engineering in Systems Application. Beijing: [s.n.], 2006(2): 1994-1997.

[11] WANG Hai-sheng, YANG Yan, WANG Hong-jun, et al. Soft voting cluster ensemble[C]//Proceedings of the 11th International Conference on Multiple Classifier Systems, Lecture Notes in Computer Science. Nanjing: [s.n.], 2013(7827): 307-318.

[12] IAM-ON N, BOONGONE T, GARRETT S, et al. Link-based cluster ensemble approach for categorical data clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2012(24): 413-425.

[13] HAN Jia-wei, KAMBER M, PEI Jian. Data mining: concepts and techniques[M]. San Francisco: Morgan Kaufmann Press, 2012.

[14] DUNN J C. A Fuzzy relative of the isodata process and its use in detecting compact well-separated clusters[J]. Journal of Cybernetics, 1973, 3(3): 32-57.

[15] BEZDEK J C. Pattern recognition with fuzzy objective function algorithm[M]. New York: Plenum Press, 1981.

[16] CALADO P, CRISTO M, GONCALVES M A, et al. Link-based similarity measure for the classification documents[J]. Journal of the American Society for Information Science and Technology, 2006, 57(2): 208-221.

[17] IAM-ON N, GARRETT S. LinkCluE: A matlab package for link-based cluster ensemble[J]. Journal of Statistical Software, 2010, 36(9): 1-36.

[18] KARYPIS G, KUMAR V. Multilevel K-way partitioning scheme for irregular graphs[J]. Parallel Distributed Computing, 1998, 41(2): 278-300.

[19] ASUNCION A, NEWMAN D J. “UCI machine learning repository” school of information and computer science, university of california[DB/OL]. (2007-06-02). http://www. ics.uci.edu/ ~mlearn/ MLRepository.html.

[20] KDD Cup. 1999 Data[DB/OL]. (1999-05-13). http:// kdd.ics.uci.edu/databases /kddcup99/kddcup99.html.

[21] STREHL Alexander. Prof. Dr. Alexander’s Data-sets [DB/OL]. (2003-03-27). http://www. lans. ece. utexas. Edu/ ~strehl.

[22] RAND W M. Objective criteria for the evaluation of clustering methods[J]. Journal of American Statistical Association, 1971, 66(336): 846-850.

[23] 杨燕, 靳蕃, KAMEL M. 聚类有效性评价综述[J].计算机应用研究, 2008, 25(6): 1630-1632.
YANG Yan, JIN Fan, KAMEL M. Survey of clustering validity evaluation[J]. Application Research of Computer, 2008, 25(6): 1630-1632.

编 辑 蒋 晓