基于实数值链接分析的 ESSC 融合算法*

王丽娟[1]，郝志峰[1,2]，蔡瑞初[2]，温 雯[2]
(1. 华南理工大学 计算机科学与工程学院，广州 510006；2. 广东工业大学 计算机学院，广州 510006)

摘 要：为了进一步提升 ESSC 聚类融合性能，采用实数值链接分析(real valued link analysis)计算聚类融合中模糊数据类的相似性。根据模糊决策及其相似性定义优化的融合信息，从而达到改进聚类性能的目的。实验选用了两个仿真数据库和五个 UCI 数据库。实验结果表明，基于实数值链接分析的 ESSC 聚类融合算法(RLA-ESSCE)的性能优于 K-means 聚类算法(KMC)、ESSC、ESSCE。

关键词：增强的软子空间聚类；聚类融合；实数值链接分析；聚类融合信息
中图分类号：TP181；TP301.6 **文献标志码**：A **文章编号**：1001-3695(2014)05-1366-04
doi:10.3969/j.issn.1001-3695.2014.05.019

Enhanced soft subspace clustering ensemble based on real valued link analysis

WANG Li-juan[1], HAO Zhi-feng[1,2], CAI Rui-chu[2], WEN Wen[2]
(1. *School of Computer Science & Engineering, South China University of Technology, Guangzhou* 510006, *China*; 2. *Faculty of Computer, Guangdong University of Technology, Guangzhou* 510006, *China*)

Abstract: In order to further improve the performance of ESSCE, real-valued link analysis had been proposed to compute the similarity between fuzzy clusters in ESSC clustering ensemble (RLA-ESSCE). The clustering ensemble information was refined according to fuzzy decision and its similarity. Therefore the performance according to refined clustering ensemble information had been improved. Experiments were conducted on two synthetic datasets and five UCI datasets. Experimental results show that RLA-ESSCE is better than K-means clustering (KMC), ESSC and ESSCE.

Key words: ESSC; clustering ensemble; real valued link analysis; clustering ensemble information

随着新型数据分析需求的不断出现，如生物信息学[1]和人类行为分析[2]，高维数据聚类成为聚类分析的研究热点。但是高维数据通常具有噪声、稀疏性和类存在于不同子空间的问题[3]，因此高维数据聚类是聚类分析的一个难题。软子空间聚类假定所有特征均参与聚类，但是对于不同数据类的贡献不同[4,5]。软子空间聚类算法通过子空间权重标示不同数据类的子空间，并在较大的特征值对应的子空间中寻找相应的数据类。因此软子空间聚类适用于解决高维数据聚类，尤其适用于类存在于不同子空间的问题。

经典的软子空间聚类算法如 EWKM[6] 和 LAC[7] 算法。这两种算法类似，均通过类内离散度和负的加权熵形成最终的软子空间聚类。Enhanced soft subspace clustering (ESSC)[8] 是 EWKM 的拓展算法，根据类内和类间的离散度以及负的加权熵形成软子空间聚类。而且 ESSC 中引入模糊系数，计算数据同时属于多个类的隶属度。实验表明 ESSC 的聚类性能优于 EWKM 算法。近两年国内对软子空间聚类较为关注。文献[9]提出基于差分演化算法的软子空间聚类；文献[10]提出特征加权距离与软子空间学习相结合的文本聚类新方法；文献[11]提出自适应的软子空间聚类算法。

与其他软子空间聚类算法类似，ESSC 算法的性能受到初始化[12,13]、参数[8]和特征子空间[13,14]的影响。文献[8,15]提出聚类融合解决软子空间聚类中参数选取的难题。由于基聚类决策来自不同的初始化、不同的参数、不同特征，聚类融合比标准聚类算法更准确、更稳定、更健壮、更有意义。因此软子空间聚类融合是解决高维数据聚类的一种有效方法。但是文献[8,15]中的聚类信息直接来自于基聚类决策，仅包含了数据与类的隶属度；而忽略了不同基聚类间以及同一基聚类内部数据类的相似性。本文在文献[8,15]算法的基础上，提出实数值链接分析计算聚类融合中两两数据类的相似性，并借此优化融合信息。文献[16,17]通过链接分析[18,19]优化清晰 K-means 聚类(K-means clustering, KMC)的融合信息；而本文所提出实数值链接分析(link analysis)用于优化 ESSC 生成的模糊软子空间聚类融合信息，即本文算法可以看做是前者的泛化。通过两个仿真数据库和五个 UCI 数据库的实验表明：基于链接分析的 ESSC 聚类融合算法(LA-ESSCE)的性能优于 KMC、ESSC、ESSC 聚类融合算法(enhanced soft subspace clustering ensemble, ESSCE)。

1 ESSC 模糊软子空间聚类算法及其扰动

假定 N、D 和 K 分别表示数据个数、特征维度和聚类个数。$X=[x_i]_N$ 是数据集，其中 $x_i=\{x_{i1},x_{i2},\cdots,x_{iD}\}$ 是第 ith 个数据，

收稿日期：2013-06-18；**修回日期**：2013-08-02 **基金项目**：国家自然科学基金资助项目(61070033,61100148,61202269)；广东省自然科学基金资助项目(S2011040004804)；广东省科技计划项目(2010B050400011)；软件新技术国家重点实验室开放项目(KFKT2011B19)；广东高校优秀青年创新人才培育项目(LYM11060)；广州市科技计划项目(12C42111607,201200000031)；番禺区科技计划项目(2012-Z-03-67)

作者简介：王丽娟(1978-)，女，河北邢台人，讲师，博士(后)，主要研究方向为机器学习、数据挖掘(ljwang@gdut.edu.cn)；郝志峰(1968-)，男，江苏苏州人，教授，博士，主要研究方向为机器学习、进化计算；蔡瑞初(1983-)，男，浙江温州人，副教授，博士，主要研究方向为机器学习、生物信息学；温雯(1981-)，女，江西赣州人，副教授，博士，主要研究方向为机器学习、图像识别.

由 D 维特征描述。设 $\mathrm{Cen}=[\mathrm{Cen}_k]_K$ 表示聚类中心矩阵,其中 $\mathrm{Cen}_k=\{\mathrm{Cen}_{k1},\mathrm{Cen}_{k2},\cdots,\mathrm{Cen}_{kD}\}$ 表示第 kth 类聚类中心。设 $\boldsymbol{U}=[u_{ki}]_{K\times N}$表示数据隶属矩阵,其中 u_{ki}表示数据 $x_i(1\leqslant i\leqslant N)$属于第 kth 类的隶属度,$0\leqslant u_{ki}\leqslant 1$,而且要求一个数据属于所有类的隶属度之和为 1,即 $\sum_{k=1\cdots K}u_{ki}=1(1\leqslant i\leqslant N)$。设 $\boldsymbol{W}=[w_{kj}]_{K\times D}$表示子空间权重矩阵,其中 w_{kj}表示第 jth 维特征对于第 kth 类的贡献度,$0\leqslant w_{kj}\leqslant 1$,而且要求所有特征对于同一类的贡献加和为 1,即 $\sum_{j=1\cdots D}w_{kj}=1(1\leqslant k\leqslant K)$。

ESSC[8] 根据类内和类间离散度以及负的加权熵计算每维特征对于每个数据类的贡献度,其学习目标函数定义为

$$J_{\mathrm{ESSC}}(\boldsymbol{U},V,\boldsymbol{W},X)=J_{C-W}+\gamma E-\eta J_{S\_W}=\sum_{k=1}^{K}\sum_{i=1}^{N}u_{ki}^{m}\sum_{j=1}^{D}w_{kj}(\mathrm{Cen}_{kj}-x_{ij})^2+\gamma\sum_{k=1}^{K}\sum_{j=1}^{D}w_{kj}\log w_{kj}-\eta\sum_{k=1}^{K}\Big(\sum_{i=1}^{N}u_{ki}^{m}\Big)\sum_{j=1}^{D}w_{kj}(\mathrm{Cen}_{kj}-\overline{\mathrm{Cen}_j})$$

$$\text{s.t.}\quad 0\leqslant u_{ki}\leqslant 1,\sum_{k=1}^{K}u_{ki}=1,1\leqslant i\leqslant N;$$

$$0\leqslant w_{kj}\leqslant 1,\sum_{j=1}^{D}w_{kj}=1,1\leqslant k\leqslant K \tag{1}$$

式中:$\overline{\mathrm{Cen}}$为所有数据的平均值;J_{C-W}表示类内离散度;J_{S-W}表示类间离散度;E 表示负的加权熵;参数 γ 和 η 用于平衡类内离散度、类间离散度和负的加权熵之间的关系;参数 m 为模糊系数,用于生成模糊分割,通常设为 2。

根据式(1)可知,输入数据集 X、输出聚类隶属矩阵 $\boldsymbol{U}$、聚类中心矩阵 Cen 和子空间权重矩阵 $\boldsymbol{W}$。目标函数式(1)通过迭代最小二乘法逐个优化上述三个矩阵。求解过程类似于模糊 C-均值聚类算法,但是在推导过程中增加了子空间权重 $\boldsymbol{W}$ 的计算步骤。根据上述分析可知:

a)ESSC 采用迭代最小二乘法优化目标函数,因此聚类过程对于初始聚类中心选取敏感,不同的初始化将导致算法收敛到不同局部极值。

b)ESSC 算法的性能受到两个参数 γ 和 η 的影响,γ 和 η 的取值依赖于具体的数据库,但是在毫无先验信息的前提下,很难合理地确定参数 γ 和 η 的取值,文献[8,15]为了避免参数选取的难题,引入聚类融合获得较好的聚类性能。

c)虽然 ESSC 为每维特征学习相应数据类的权重,并在较大特征权重的子空间形成相应类的数据聚类。较小特征权重在聚类过程中的作用可以忽略不计。文献[14]的研究证明了软子空间聚类 FG-K-means 算法和特征选择能够进一步改善其聚类性能。

2　基于实数值链接分析的 ESSC 融合算法

2.1　ESSC 融合信息优化

设 $\Pi=\{\pi_1,\pi_2,\cdots,\pi_B\}$ 是具有 B 个基聚类的聚类融合。每个基聚类定义为 $\pi_q=\{C_1^q,C_2^q,\cdots,C_K^q\}$,其中 K 是基聚类形成聚类的类数,C_k^q 是第 qth 个基聚类中第 kth 数据类,其类心为 Cen_k^q,并要求数据类满足如下条件:$\cup_{k=1}^{K}C_k^q=X$。基聚类 π_q 的数据隶属信息为 $\boldsymbol{U}^q=[u_{ki}]^q(1\leqslant i\leqslant N)$;类所存在软子空间权重为 $\boldsymbol{W}^q=[w_{kj}]^q(1\leqslant j\leqslant D)$。聚类融合通过集成多个基聚类决策得到一致性聚类决策 $\pi^*=\{C_1^*,C_2^*,\cdots,C_K^*\}$,使其与所有基聚类决策具有最大的共享信息[13,20]。

聚类融合信息是决定聚类融合性能的关键因素。本文选用 ESSC 作为基聚类算法,其输出数据与类在相应软子空间 $\boldsymbol{W}^q$的模糊隶属矩阵 $\boldsymbol{U}^q$,及类心矩阵 Cen^q。目前常用的软子空间聚类融合算法直接将基聚类的隶属信息 $\boldsymbol{U}^q$作为融合信息,并从中提取融合决策。

根据信息熵的定义,当数据与类的隶属信息趋向于 0/1,该模糊聚类决策较为清晰,此时的模糊决策较优。ESSC 是性能优越的软子空间模糊聚类算法,聚类融合中包含大部分基聚类决策中信息趋向于 0/1。根据最大隶属原则可知当数据所属的类的隶属度较大,趋向于 1;由于要求数据对所有类隶属度加和为 1,因此数据与其他类的隶属度值较小,趋向于 0,此时隶属度提供的聚类融合信息较少。因此,笔者希望在融合信息中保留数据所属类的最大隶属度,而忽略数据属于其他类的较小隶属度,这部分信息是根据链接分析得到数据类的相似性计算得到,定义为

$$RU^{b_s}(x_i,k_i)=\begin{cases}U_{k_{\max}}^{b_s}(x_i)(k_i=k_{\max}\mid U_{k_{\max}}^{b_s}(x_i)=\max\limits_{1\leqslant k_j\leqslant K}(U_{k_j}^{b_s}(x_i)))\\(U_{k_{\max}}^{b_s}(x_i))\times w^c(C_{k_i}^{b_s},C_{k_{\max}}^{b_s})(k_i\neq k_{\max})\end{cases} \tag{2}$$

式中:$k_{\max}$表示具有最大隶属度的类,即数据所属类;$U_{k_{\max}}^{b_s}(x_i)$表示在基聚类 π_{b_s}中最大隶属度;$w^c(C_{k_i}^{b_s},C_{k_{\max}}^{b_s})$表示链接分析得到数据类的相似度,定义见 3.2 节。根据式(2)定义的融合信息,保留了数据最大隶属度信息;较小的隶属度信息由数据的最大隶属度及所属类与其他类相似性的乘积替换。式(2)定义不仅增加了融合信息量,而且引入聚类融合中的高层次信息,提高了信息的质量。优化前聚类融合信息记做 $\boldsymbol{U}^q$,仅包含基聚类决策输出的数据与类的隶属信息;而优化后聚类融合信息记做 $\boldsymbol{RU}^q$,包含数据的隶属信息和数据类的相似性。

本文所提出的基于实数值链接分析的 ESSC 融合算法(RLA-ESSCE)重点研究基聚类算法和一致性函数。基聚类算法选用 ESSC,其多样性通过初始化、参数和特征子集扰动产生。一致性函数用于表示融合信息和提取融合结果。在一致性函数中引入链接分析,计算聚类融合所包含数据类的相似性,并借此优化聚类融合信息。根据优化后的融合信息 $\boldsymbol{RU}$,采用图聚类算法 SPEC[21] 融合基聚类决策得到最终的融合结果。基于实数值链接分析的 ESSC 聚类融合算法(RLA-ESSCE)的系统框图如图 1 所示。

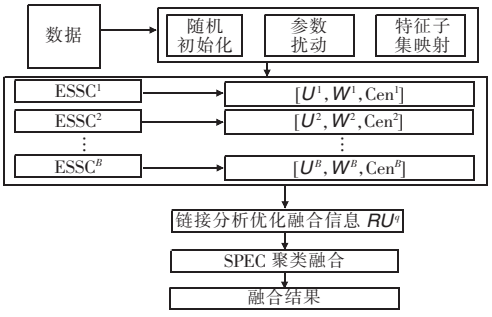
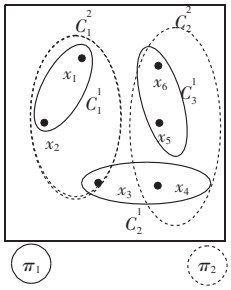


图 1　基于实数值链接分析的 ESSC 融合系统(RLA-ESSCE)

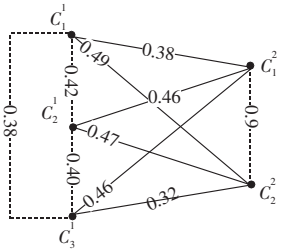
2.2　实数值链接分析

链接分析将聚类融合看做图 $G=\langle V,\boldsymbol{W}^c\rangle$,其中所有数据类看做图中的数据点,包含在点集 V 中;数据类的相似性看做图的边,包含在边集 $\boldsymbol{W}^c=\{w^c\}$,边的权重 w^c 表示数据类间的相似度。将根据图 2 所示数据说明链接分析如何计算数据类的相似性。图 2 所示数据聚类包含两个基聚类算法,分别是 $\pi_1=\{C_1^1,C_2^1,C_3^1\}$用黑色实线表示和 $\pi_2=\{C_1^2,C_2^2\}$用黑色虚线表示。图(a)为聚类融合的示例,(b)为根据链接分析提取的数据类的相似性,(c)为基聚类模糊决策矩阵 $\boldsymbol{U}$,(d)为优化后

的融合矩阵 RU。



(a)聚类融合示意图



(b)数据类关系图

U	C_1^1	C_2^1	C_3^1	C_1^2	C_2^2
x_1	0.8	0.1	0.1	0.6	0.4
x_2	0.7	0.2	0.1	0.8	0.2
x_3	0.3	0.6	0.1	0.7	0.3
x_4	0.3	0.5	0.2	0.4	0.6
x_5	0.1	0.2	0.7	1	0
x_6	0.3	0.1	0.6	0.9	0.1

(c)基聚类模糊决策

RU	C_1^1	C_2^1	C_3^1	C_1^2	C_2^2
x_1	0.80	0.36	0.30	0.60	0.54
x_2	0.70	0.32	0.27	0.80	0.72
x_3	0.27	0.6	0.22	0.70	0.63
x_4	0.23	0.5	0.19	0.54	0.60
x_5	0.27	0.26	0.7	1.0	0.90
x_6	0.23	0.22	0.6	0.90	0.81

(d)优化聚类融合信息矩阵

图 2 聚类融合示例

不同基聚类根据其共享数据计算数据类的相似性。如图 2(a)中数据类 C_1^1 和 C_1^2 分属两个不同基聚类算法,这两个聚类均共享数据 x_1 和 x_2,因此这两个数据类的相似度可以通过共享数据 x_1 和 x_2 计算。由于所处理聚类为模糊聚类,因此定义所有数据对于这两个数据类的最小隶属度之和与最大隶属度之和的比值为不同基聚类的相似性,定义为

$$w^c(C_{k_i}^{b_s},C_{k_j}^{b_t})=\frac{\sum_{1\leqslant x_i\leqslant N}(U_{k_i}^{b_s}(x_i)\wedge U_{k_j}^{b_t}(x_i))}{\sum_{1\leqslant x_i\leqslant N}(U_{k_i}^{b_s}(x_i)\vee U_{k_j}^{b_t}(x_i))}\begin{pmatrix}1\leqslant b_s,b_t\leqslant B,b_s\neq b_t\\1\leqslant k_i,k_j\leqslant K\end{pmatrix} \tag{3}$$

根据式(3)计算两个不同基聚类内数据类的相似性,如图 2(b)中的实线及其权重部分。

假定聚类没有退化的前提下,在同一基聚类内,不同数据类具有不同的类心,表示不同的数据类,根据共享数据计算数据类的相似度较小。从聚类融合的角度出发,同一基聚类内数据类的相似性可以根据其他基聚类的数据类计算相似性。图 2(a)所示数据类 C_1^1 和 C_2^1 属于同一个基聚类 π_1,数据类 C_1^1 和 C_2^1 分别表示不同的聚类中心,其具有共享数据的相似度非常小。如数据类 C_1^1 和 C_2^1 所包含的数据 x_1、x_2 和 x_3 在基聚类 π_2 中属于同一数据类 C_1^2,因此数据类 C_1^2 被看做数据类 C_1^1 和 C_2^1 的共享邻居。同一基聚类内部两个数据类相似性的计算依赖其他基聚类中的共享邻居计算,定义为

$$\mathrm{sim}(C_{k_i}^{b_s},C_{k_j}^{b_s},C_{k_q}^{b_t})=\frac{(w^c(C_{k_i}^{b_s},C_{k_q}^{b_t})+w^c(C_{k_j}^{b_s},C_{k_q}^{b_t}))}{(\sum_{1\leqslant k_r,k_q\leqslant K}w^c(C_{k_r}^{b_s},C_{k_q}^{b_t}))} \tag{4}$$

式中:$C_{k_i}^{b_s}$、$C_{k_j}^{b_s}$是基聚类 π_{b_s} 中的数据类。$C_{k_q}^{b_t}$是基聚类 π_{b_t} 中的数据类,而且 $b_s\neq b_t$,则 $C_{k_q}^{b_t}$为 $C_{k_i}^{b_s}$、$C_{k_j}^{b_s}$的共享邻居。分子是 $C_{k_i}^{b_s}$和 $C_{k_j}^{b_s}$与共享邻居 $C_{k_q}^{b_t}$的相似性。分母是共享邻居 $C_{k_q}^{b_t}$的贡献率,当共享邻居与其他数据类的相似性越大,则其贡献越小;否则反之。上述分母定义在社交网络中网页的相似性计算中得到证明[22]。根据式(4)计算 $C_{k_i}^{b_s}$和 $C_{k_j}^{b_s}$与所有共享邻居的相似性,得到 $C_{k_i}^{b_s}$和 $C_{k_j}^{b_s}$的相似性定义为

$$\mathrm{sim}(C_{k_i}^{b_s},C_{k_j}^{b_s})=\sum_{b_t=1}^{B}\sum_{k_q=1}^{K}\mathrm{sim}(C_{k_i}^{b_s},C_{k_j}^{b_s},C_{k_q}^{b_t}) \tag{5}$$

从中寻找最大数据类的相似性,根据式(6)将相似度标准化到区间[0,1]:

$$w^c(C_{k_i}^{b_s},C_{k_j}^{b_s})=\frac{\mathrm{sim}(C_{k_i}^{b_s},C_{k_j}^{b_s})}{\max(\mathrm{sim}(C_{k_i}^{b_s},C_{k_j}^{b_s}))}\times df \tag{6}$$

式中:df 为衰减因子,通常设置为 $df=0.9$。根据式(6)计算得到同一基聚类内两个数据类的相似性,如图 2(b)中虚线及其权重所示。进一步,式(2)根据图 2(b)(c)的信息计算得到优化后的融合信息 RU,如图 2(d)所示。

3 实验

3.1 实验数据

本文选取两个仿真数据库和五个 UCI 数据库,分析所提出 RLA-ESSCE 算法的融合性能,数据库的介绍如表 1 所示。

表 1 两个仿真数据集和四个 UCI 数据介绍

数据库	数据	特征	聚类个数	数据类型
2-banana	2000	7	2	实值型
3-Gaussian	2300	10	3	实值型
Breast cancer	569	30	2	实值型
Liver disorder	345	6	2	实值型
Sonar	208	60	2	实值型
Wave	5000	40	3	实值型
Wine	178	13	2	实值型

3.2 实验方法及参数

实验将对比 KMC、ESSC、ESSCE 以及 RLA-ESSCE 算法的聚类性能。ESSC 算法思想来源于文献[8]。ESSCE 是选用 ESSC 作为基聚类算法,基聚类算法的多样性考虑不同初始化、参数和特征子集;聚类融合信息为基聚类决策信息,聚类融合采用 SPEC 算法融合。

上述算法中 ESSC、ESSCE 和 RLA-ESSCE 涉及到两个聚类参数,即 γ 和 η。根据文献[6~8]中的实验结论,参数 γ 和 η 的取值为[0,1]区间的实数。在没有先验信息的前提下,参数 γ 和 η 的取值很难确定,因此文献[15]提出了基于参数扰动 LAC 聚类融合算法。本文将参数 γ 和 η 的取值作为多样性的一种扰动方式,在上述取值区间内随机确定两个参数的取值。除此之外,聚类融合的算法具有参数:基聚类个数 B 和选取特征子集的比例 1/por。这两个参数依赖不同的数据库有不同的取值。但是在聚类融合的实验中,基聚类个数取值较小,算法性能逐渐退化为标准聚类算法;随着聚类个数的增加,算法性能逐渐提升,但是融合信息包含冗余信息,算法的时空复杂度增大。因此本文设基聚类个数为经验值 $B=20$。在聚类融合实验中,仅选取部分特征融合,可以减少噪声特征的影响,但是当选取特征较少势必造成信息损失。本文中选取特征子集的比例 1/por 设为区间[0.5,1]。以下所分析的实验结果是 20 次实验结果的平均值。

3.3 实验结果分析

KMC、ESSC、ESSCE 和 RLA-ESSCE 算法正确性的对比,如表 2 所示。

表 2 KMC、ESSC、ESSCE 和 LA-ESSCE 四种算法实验结果比较

数据库	KMC	ESSC	ESSCE	RLA-ESSCE
2-banana	0.83	0.88	0.92	0.96
3-Gaussian	0.70	0.84	0.87	0.89
Breast cancer	0.85	0.89	0.90	0.91
Liver disorder	0.55	0.57	0.58	0.58
Sonar	0.53	0.55	0. 57	0.58
Wave	0.51	0.73	0.75	0.78
Wine	0.70	0.91	0.91	0.91

根据实验结果,可以得到以下实验结论:

a)对于含有噪声数据和类存在于不同子空间的数据库,KMC 的聚类性能较差,如 KMC 聚类 3-Gaussian 和 Wave 数据库的结果显著低于其他聚类算法的聚类结果。

b)所有数据库的 ESSC 聚类结果均优于 KMC 的聚类结果。ESSC 算法通过子空间权重减少不同数据类的噪声特征对于聚类的影响,但是却无法预估参数 γ 和 η 取值。当参数 γ 和 η 在[0,1]区间随机取值时,ESSC 算法的平均性能低于 ESSCE 的融合性能,如两个仿真数据库和 Breast Cancer 数据库。

c)对于大多数数据库 ESSCE 的性能优于单个聚类算法 KMC 和 ESSC,仅对于 Wine 数据库,ESSCE 和 ESSC 取得了相同的性能。聚类融合算法从不同的初始化、不同的参数、不同特征探测数据内部规律,因此聚类融合算法的聚类性能均优于单个聚类算法。但是 ESSCE 中融合信息完全依赖基聚类决策,忽略了聚类融合的高层次信息,因此其聚类性能低于 RLA-ESSCE。

d)对于所有数据库 RLA-ESSCE 优于 KMC、ESSC 和 ESSCE。对于两个仿真数据库和 Wave 数据库,RLA- ESSCE 聚类性能改进较大;即使对于 Wine 和 Liver disorder 数据库,RLA-ESSCE 取得了和 ESSCE 同样的最优的聚类性能。RLA-ESSCE 性能优越的主要原因:(a)RLA-ESSCE 根据多个扰动探测数据,避免了初始化、参数和特征子集对算法性能的影响,因此其性能优于单聚类算法;(b)RLA-ESSCE 采用链接分析优化融合信息,不仅增加融合信息量,而且提升融合信息层次,因此融合结果优于 ESSCE 融合算法。

4　结束语

实数值链接分析算法提取 ESSC 模糊决策中的数据类相似性,并借此信息优化聚类融合信息。ESSC 的模糊决策不仅包含数据类隶属信息,而且包含了子空间权重信息,为聚类融合提供更丰富的融合信息。实数值链接分析算法处理的数据域由原来 0/1 二值清晰决策拓展到[0,1]实数值的模糊决策,是传统链接分析算法的拓展算法。根据模糊隶属矩阵和数据类相似性,优化的融合信息保留最大类的隶属度,修正较小类隶属度为其隶属类间的相似性,从而引入了聚类融合中所包含的高层次的融合信息。仿真数据和 UCI 数据实验表明,RLA-ESSCE 的性能优于 KMC、ESSC、ESSCE。

实数值链接分析算法中不同基聚类和同一基聚类的数据类的相似性是两个研究重点。不同基聚类中数据类的相似性采用了取小取大算子,未来将深入研究其他相似性度量策略。同一基聚类中数据类的相似采用了共享邻居的方式计算,除此之外网络学习中还有很多因素和学习策略可以借鉴提高实数值链接分析算法提取数据类相似性的信息量,也将是笔者未来研究的一个重点。

参考文献:

[1] JIANG Da-xin, TANG Chun, ZHANG Ai-dong. Cluster analysis for gene expression data: a survey [J]. IEEE Trans on Knowledge and Data Engineering, 2004, 16(11): 1370-1386.

[2] LIU Zong-yi, SARKAR S. Improved gait recognition by gait dynamics normalization[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2006, 28(6): 863-876.

[3] DONOHO D. High-dimensional data analysis: the curses and blessings of dimensionality[C]//Proc of the 21st American Mathematical Society-Mathematical Challenges Century. 2000: 1-32.

[4] PARSONS L, HAQUE E, LIU H. Subspace clustering for high dimensional data: a review [J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 90-105.

[5] KRIEGEL H, KROGER P, ZIMEK A. Clustering high-dimensional data: a survey on subspace clustering, pattern based clustering, and correlation clustering[J]. ACM Trans on Knowledge Discovery from Data, 2009, 3(1): 1-58.

[6] DOMENICONI C, GUNOPULOS D, MA Sheng, *et al*. Locally adaptive metrics fore clustering high dimensional data [J]. Data Mining Knowledge Discovery, 2007, 14(1): 63-97.

[7] JING L P. An entropy weighting K-means algorithm for subspace clustering of high dimensional sparse data [J]. IEEE Trans on Knowledge and Data Engineering, 2007, 19(8): 1026-1041.

[8] DENG Z H, CHOI K S, CHUNG F L, *et al*. Enhanced soft subspace clustering integrating within-cluster and between-cluster information [J]. Pattern Recognition, 2010, 43(3): 767-781.

[9] 毕志升,王甲海,印鉴. 基于差分演化算法的软子空间聚类[J]. 计算机学报,2012,35(10):2116-2128.

[10] 王骏,王士同, 邓赵红. 特征加权距离与软子空间学习相结合的文本聚类新方法[J]. 计算机学报,2012,35(8):1655-1665.

[11] 陈黎飞, 郭躬德, 姜青山. 自适应的软子空间聚类算法[J]. 软件学报,2010,21(10):2513-2523.

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

[13] TOPCHY A, JAIN A K, PUNCH W. Clustering ensembles: models of consensus and weak partitions [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(12): 1866-1881.

[14] CHEN Xiao-jun, YE Yun-ming, XU Xiao-fei, *et al*. A feature group weighting method for subspace clustering of high-dimensional data [J]. Pattern Recognition, 2012, 45(1): 434-446.

[15] DOMENICONI C, Al-RAZGAN M. Weighted cluster ensembles: methods and analysis[J]. ACM Trans on Knowledge Discovery from Data, 2009, 2(4): 1-40.

[16] IAM O N, BOONGOEN T, GARRETT S, *et al*. A link-based approach to the cluster ensemble problem[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2011, 33(12): 2396-2409.

[17] IAM O N, BOONGOEN T. Improved link-based cluster ensembles [C]//Proc of IEEE World Congress on Computational Intelligence. 2012: 10-15.

[18] BOONGOEN T, SHEN Q, PRICE C. Disclosing false identity through hybrid link analysis[J]. Artificial Intelligence and Law, 2010, 18(1): 77-102.

[19] 张宪超,徐雯,高亮,等. 一种结合文本和链接分析的局部 Web 社区识别技术[J]. 计算机研究与发展,2012,49(11):2352-2358.

[20] FERN X Z, BRODLEY C E. Random projection for high dimensional data clustering: a cluster ensemble approach [C]//Proc of the 20th International Conference on Machine Learning. Washington DC: AAAI Press, 2003: 186-193.

[21] NG A, JORDAN M, WEISS Y. On spectral clustering: analysis and an algorithm[J]. Advances in Neural Information Processing Systems, 2001, 14: 849-856.

[22] ADAMIC L A, ADAR E. Friends and neighbors on the Web [J]. Social Networks, 2003, 25(3): 211-230.