基于多扰动的局部自适应软子空间聚类融合算法

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

摘 要 提出基于随机初始化、参数扰动和特征子集映射的多扰动的局部自适应软子空间聚类(LAC)融合算法(MLACE)。MLACE具有以下特点:(i)多扰动融合:从初始化、参数和特征子集等不同侧面,探测数据内部结构,使之相互融合,从而达到改善聚类正确性的目的;(ii)融合信息提升:根据LAC算法输出的子空间权重矩阵,定义数据属于每一类的概率,形成提升的融合信息;(iii)融合一致性函数改进:融合信息的形式由0/1二值信息转换成[0,1]实值信息,因此,一致性函数采用了性能较优的实数值融合算法Fast global K-means来进一步改善融合正确性。实验选取2个仿真数据库和5个UCI数据库测试MLACE的聚类正确性,实验结果表明,MLACE聚类正确性优于K-means、LAC、基于参数扰动LAC融合算法(P-MLACE)。

关键词 聚类融合,软子空间聚类,局部自适应软子空间聚类,多扰动
中图法分类号 TP181 **文献标识码** A

Multiple Local Adaptive Soft Subspace Clustering Ensemble Based on Multimodal Perturbation

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

Abstract This paper proposed multiple local adaptive soft subspace clustering (LAC) ensemble (MLACE) based on multimodal perturbation. There are three merits in the proposed MLACE. Firstly, MLACE combines diversity and complement decisions generated by random initialization, parameter perturbation and feature subspace projection, so as to improve the accuracy of clustering. Secondly, the clustering ensemble information is refined. The probability of each instance belonging to all clusters is defined according to the subspace weight matrix from LAC. Thirdly, because the clustering ensemble information is refined from 0/1 binary value into [0,1]real value, the consensus function in clustering ensemble can adopt real valued clustering ensemble method Fast global K means, which can further improve the accuracy of clustering ensemble. Two synthetic datasets and five UCI datasets were chosen to evaluate the accuracy of MLACE. The experiment results show that MLACE is more accurate than K-means, LAC, Multiple LAC clustering ensemble based on parameter perturbation (P-MLACE).

Keywords Clustering ensemble, Soft subspace clustering, Local adaptive soft subspace clustering, Multimodal perturbation

1 引言

软子空间聚类算法[1,2]为每个数据学习相应的数据类,以及数据类所属的子空间。软子空间聚类算法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。因此,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。与全局无监督特征加权算法+聚类算法相比,软子空间聚类算法是前者的局部泛化,成为目前高维数据聚类的一个研究热点。

根据不同的权重确定方式,软子空间聚类算法被进一步分成两类,分别是:模糊加权软子空间聚类算法和熵加权软子空间聚类算法。模糊加权软子空间聚类算法属于早期算法,其中AWA[3]和W-K-means[4]是其经典算法,由香港大学的Zhexue Huang等人提出。熵加权软子空间聚类算法是近期的研究热点,其中局部自适应软子空间聚类算法(LAC)[5]和EWKM算法[6]是经典算法,分别由C. Domeniconi和Li Ping Jing提出。LAC和EWKM算法非常相似,二者均根据类内离散度和加权熵为每个特征指定相应类的权重;区别是LAC算法在目标函数中考虑了每个类的容量大小,而EWKM算法未考虑这一参数。除此之外,ESSC[7]由江南大学的Zhaohong Deng提出,权重学习过程不仅最小化类内离散度而且最大化类间离散度。FG-k-means算法[8]由哈工大的Xiaojun

到稿日期:2013-04-08 返修日期:2013-10-14 本文受国家自然科学基金(61070033,61100148,61202269),广东省自然科学基金(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课题(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)资助。
王丽娟(1978—),女,博士,讲师,主要研究方向为机器学习、数据挖掘,E-mail:ljwang@gdut.edu.cn;郝志峰(1968—),男,博士,教授,主要研究方向为机器学习、进化计算。

Chen提出，为每个特征及每个特征类学习权重。MOEASSC算法[9]由西安交大的Xia Hu提出，通过多目标进化优化实现软子空间聚类。厦门大学的陈黎飞提出一种自适应学习参数的软子空间聚类算法[10]。

上述软子空间聚类算法的性能不仅与算法自身性质相关，而且受到初始化、参数和特征子集映射的影响。本文希望研究基于随机初始化、参数扰动和特征子集映射的多扰动的软子空间聚类融合算法，使其能够从不同的侧面探测数据内部结构，最终达到改进聚类正确性的目的。但是目前聚类融合研究主要集中于K-means算法[12-14]，而针对软子空间聚类算法的研究较少。文献[11]提出了基于参数扰动的LAC聚类融合算法，记作P-MLACE。软子空间聚类融合算法重点解决如何利用子空间权重完善聚类融合信息，进而改善融合一致性函数，最终达到融合系统正确性改进的目的。

本文将以LAC算法为基聚类算法，提出了基于随机初始化、参数扰动和特征子集映射等多扰动的LAC聚类融合算法(MLACE)。为了充分利用LAC输出的子空间权重信息，构造每类的加权欧氏距离，计算数据属于所有类的概率，以此构成优化的融合信息。与文献[11]中算法P-MLACE相比，本文计算的数据属于类的概率具有较大的区分度，因此提供更清晰的融合信息。由于融合信息由传统的0/1二值转换成[0,1]区间的实值信息，本文将采用性能更优的实数值融合算法Fast global K-means[15]，而不是符号融合算法。实验表明，MLACE的聚类正确性优于K-means(KMC)，LAC，P-MLACE。

2 基于多扰动的局部自适应软子空间聚类融合算法

设n,m和K分别表示数据个数、特征空间维度和聚类的类数。$\boldsymbol{X}=[x_i]_n$是聚类数据集，其中$x_i=\{x_{i1},x_{i2},\cdots,x_{im}\}$ $(1\leqslant i\leqslant n)$是特征空间$R^m$的第$ith$个样例。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$C_k^q$是第$qth$个基聚类中第$kth$个数据类，并要求数据类满足如下条件：$\bigcup_{k=1}^{K}C_k^q=\boldsymbol{X}$。基聚类$\pi_q$对应的类心矩阵为$\boldsymbol{Cen}^q=[cen_k^q]_K$，数据分割矩阵为$\boldsymbol{U}^q=[u_{ki}^q]_K$ $(1\leqslant i\leqslant n)$，类所存在子空间权重矩阵为$\boldsymbol{W}^q=[w_{kj}^q]_K$ $(1\leqslant j\leqslant m)$。

聚类融合就是通过集成多个基聚类决策得到最优的一致性聚类决策$\pi^*=\{C_1^*,C_2^*,\cdots,C_K^*\}$，其与原始数据类具有最大的共享信息[14,16]。由于基聚类决策来自不同的初始化、不同的参数、不同的特征，聚类融合比标准聚类算法更准确、更稳定、更健壮、更有意义[17]。

聚类融合的两个关键因素分别是基聚类算法和一致性函数[14,16,17]。基聚类算法选用LAC算法，重点研究多样性设置，详细内容见2.1节。一致性函数用于描述和融合基聚类算法提供的融合信息，并且从中提取最终的一致性聚类结果。因此一致性函数包含两个研究点，分别是融合信息优化和融合算法的改进，详细内容见2.2节和2.3节。

2.1 局部自适应软子空间聚类(LAC)及其扰动

局部自适应软子空间聚类算法(LAC)属于熵加权软子空间聚类算法，其目标函数为：最小化类内离散度，同时最大化负的加权信息熵。最小化类内离散度使得对于某类具有较小离散度的特征获得较大的权重，否则反之。单独最小化离散度容易导致所有特征权重仅集中在一维特征上，而忽略其它特征，造成大量的信息损失。因此，在最小化离散度的同时最大化负的加权熵，从而保证足够多的特征参与聚类。LAC聚类算法的目标函数如式(1)所示：

$$J_{LAC}(\boldsymbol{U},\boldsymbol{Cen},\boldsymbol{W},\boldsymbol{X})=\sum_{k=1}^{K}\sum_{j=1}^{m}w_{kj}\boldsymbol{X}_{kj}+\gamma\sum_{k=1}^{K}\sum_{j=1}^{m}w_{kj}\log w_{kj}$$

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

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

$$X_{kj}=\frac{\sum_{i=1}^{n}u_{ki}(cen_{kj}-x_{ij})}{\sum_{i=1}^{n}u_{ki}} \tag{2}$$

式(1)中X_{kj}表示类内离散度，定义如式(2)所示。根据式(1)可知，输入数据集$\boldsymbol{X}$，输出聚类分割矩阵$\boldsymbol{U}$、聚类中心矩阵$\boldsymbol{Cen}$和子空间权重矩阵$\boldsymbol{W}$。上述3个变量的学习采用迭代最小二乘优化方式逐个优化。LAC算法的详细聚类过程参考文献[3]。LAC算法具有如下特点：

(i)LAC与KMC类似，均采用迭代最小二乘法求解，只是在K-means聚类过程中增加了子空间权重的学习步骤。因此，LAC与K-means算法的聚类过程均受到初始化聚类中心的影响，不同的聚类中心将导致算法收敛到不同的局部极小值。

(ii)式(1)中参数γ用于调整最小化类内离散度和最大化负的加权熵之间的平衡关系。参数γ取值的合理性将影响聚类正确性。但是参数γ取值依赖于具体的数据库，而且目前没有任何理论指导参数γ如何取值。因此文献[11]提出了基于参数γ不同取值的LAC融合算法，改进LAC算法的聚类正确性。

(iii)LAC聚类算法为每类数据学习相应的特征权重，并在较大权重的子空间形成相应类的数据聚类。具有较小权重的特征在聚类过程中的作用可以忽略不计。文献[8]的研究证明软子空间聚类FG-k-means算法+特征选择能够进一步改善其聚类正确性。

本文提出了基于随机初始化、参数扰动和特征子集映射的多扰动局部自适应软子空间聚类融合算法(MLACE)。其中，参数γ的取值范围为区间[0,1]。特征子集映射为每个基聚类算法随机选取$1/por$的特征子集用于聚类。由于每个基聚类具有不同的初始化中心、不同参数和不同的特征子集，因此其输出的类心矩阵$\boldsymbol{Cen}$、分割矩阵$\boldsymbol{U}$和子空间权重矩阵$\boldsymbol{W}$也有所不同。

2.2 优化融合信息

基于KMC算法的融合信息仅包含数据分割矩阵$\boldsymbol{U}$，而基于LAC聚类融合算法还包含子空间权重矩阵$\boldsymbol{W}$。因此，根据子空间权重，形成优化的融合信息，定义基于子空间权重的加权欧式距离，如式(3)：

$$d_{ik}=\sqrt{\sum_{j=1}^{m}w_{kj}(x_{ij}-cen_{kj})^2} \tag{3}$$

根据每类特征权重计算数据与其相应类心的加权欧式距离，并计算数据属于每一类的概率，如式(4)：

$$P(C_k\mid x_i)=\frac{\frac{1}{d_{ik}^2+\varepsilon}}{\sum_{q=1}^{K}\frac{1}{d_{iq}^2+\varepsilon}} \tag{4}$$

式中，ε为[0,1]内一个较小的随机数，用于保证当数据与类

心距离为0时，仍可以计算数据属于该类的概率。数据属于某一类的概率由数据到该类的距离决定，当数据离类心越近，则属于该类的概率越大，否则反之。根据定义式(4)可知，$P(C_k|x_i)$具有如式(5)所示性质：

$$1\geqslant P(C_k|x_i)\geqslant 0,\ \sum_k P(C_k|x_i)=1 \tag{5}$$

计算数据到所有类的概率得到数据属于所有类后验概率矢量，如式(6)：

$$P_i=(P(C_1|x_i),P(C_2|x_i),\cdots,P(C_k|x_i))(1\leqslant i\leqslant n) \tag{6}$$

根据式(6)，定义映射 $x_i\rightarrow P_i(1\leqslant i\leqslant n)$，该映射将数据从原始的 m 维空间转换成数据与类的相关性 K 维空间，其中每一维空间对应一个数据类，形成完善的融合信息。

文献[11]提出的P-MLACE中数据属于类的概率是基于最大加权距离计算的。与文献[11]相比，本文所提出的MLACE计算数据属于某类概率具有较大的区分度，聚类结果清晰。比如，数据 x 到3类的加权距离分别为：{1,2,10}。P-MLACE和MLACE计算数据属于3类的概率如表1所列。P-MLACE计算得到数据 x 属于类1和类2的概率近似；而MLACE能够清晰地区分数据 x 属于类1和类2的概率，概率值更为合理。因此MLACE优化的融合信息优于P-MLACE。

表1 P-MLACE与MLACE形成融合信息对比

算法名称	类1	类2	类3
P-MLACE融合信息	0.5	0.45	0.05
MLACE融合信息	0.79	0.20	0.01

2.3 实值融合算法

优化后融合信息 P_i 将数据从原始的特征空间 R^m 转换到 BK 维的数据与类关系空间。聚类融合算法根据数据与类的关系聚类，得到最终一致的融合结果。由于融合信息由0/1二值关系拓展成[0,1]实数关系，因此聚类融合算法可以采用聚类性能较优的实数聚类融合算法：Fast global K-means。Fast global K-means是基于K-means算法的动态增量式聚类，得到了确定的快速全局最优聚类算法，而且融合结果不依赖于初始聚类中心。从 $k=1$ 类开始，动态增量式聚类每次迭代增加一个聚类中心，并从中选择当前最优的聚类结果，多次迭代直到指定的类数。为了降低算法的时间复杂度，提出了两种改进：(i) K-d tree 数据压缩；(ii) 预估聚类误差上界。Fast global K-means包含上述两个改进后在取得近似优化聚类结果的同时，大大降低了聚类过程所需的时间复杂度。算法的详细内容可以参考文献[15]。

综上，本文所提出的基于多扰动的局部自适应软子空间聚类融合算法(MLACE)其基聚类算法选用LAC。根据LAC的性质确定了多种扰动，分别是初始化、参数和特征子集。LAC不仅输出数据的类属信息，而且输出数据类的子空间权重，根据子空间权重计算数据属于所有类的概率，得到优化的聚类融合信息。由于融合信息从0/1二值信息提升为[0,1]的实数值信息，本文选用了聚类正确性较优的实数值融合算法Fast global K-means。MLACE的系统图如图1所示。

如图1所示，系统包含3部分，分别是扰动生成、基聚类决策以及聚类融合。其中时间复杂度最大的是基聚类决策过程。假定LAC聚类过程平均迭代次数为 r。LAC算法是K-means算法的拓展算法，即在聚类过程中增加了子空间权重学习步骤，其时间复杂性与K-means算法类似，时间复杂度为 $O(rnmK)$，适用于处理大规模数据聚类。MLACE中多个基聚类算法独立并行运算，如果采用并行算法实现，时间复杂度与单个聚类算法相当，为 $O(rnmK)$；如果采用串行算法实现，会降低多分类器融合算法的时间性能，其时间复杂度为 $O(BrnmK)$。当聚类大规模数据时，$B\ll rnmK$，因此MLACE的时间复杂度与单聚类算法相比仅增加了系数，不影响其处理大规模数据的时间性能。文献[15]指出：Fast global K-means引入2种改进后，能够快速取得与K-means算法多次随机初始化的最优聚类结果相当的聚类结果。融合阶段算法的时间复杂度与多个基聚类算法聚类的时间复杂度相比较小，不会对MLACE算法的时间复杂度产生影响。通过上述分析可知，MLACE的时间复杂度为 $O(BrnmK)$，其中 $B\ll rnmK$，适用于处理大规模数据。

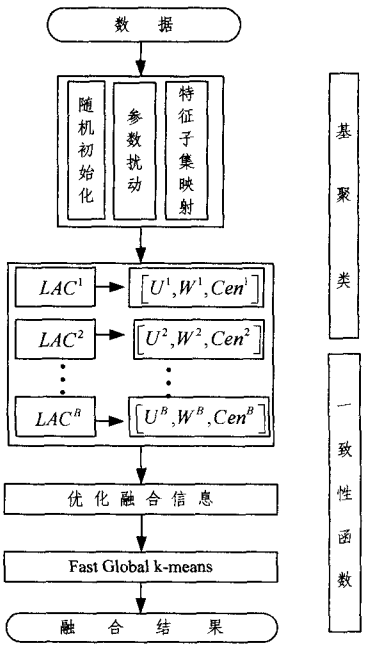
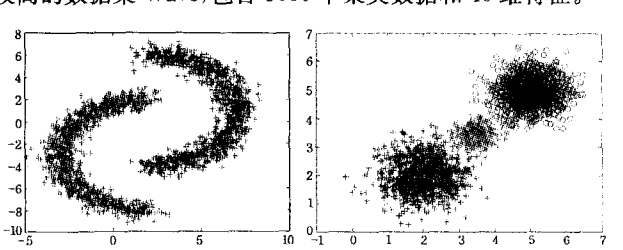


图1 基于多扰动的局部自适应软子空间聚类融合算法的示意图

3 实验

3.1 实验数据

本文选取2个仿真数据库和5个UCI数据库，分析所提出算法的融合正确性，数据库的介绍如表2所列。2个仿真数据集分别是2-banana和3-Gaussian，数据分布分别如图2(a)和图2(b)所示。需要注意两个仿真数据2-banana和3-Gaussian数据集最初包含2维数据，在实验过程中分别增加了5维和8维噪音特征。所选取的UCI实验数据中具有高维特征，如Sonar包含60维特征；同时具有大数据量和维数较高的数据集Wave，包含5000个聚类数据和40维特征。



(a) two banana shaped dataset (b) 3-Gaussian dataset

图2 2个仿真数据分布

表 2　2个仿真数据集和4个 UCI 数据介绍

数据库名	数据	特征	类	类型
2-banana	2000	7	2	Real-valued
3-Gaussian	2300	10	3	Real-valued
Breast cancer	569	30	2	Real-valued
Liver disorder	345	6	2	Real-valued
Sonar	208	60	2	Real-valued
Wave	5000	40	3	Real-valued
Wine	178	13	2	Real-valued

3.2　实验方法、参数和评价

本文实验平台为 DELL390 商用 PC 机，操作系统为 WinXP，编程语言为 Visual C++。实验将对比以下算法的聚类正确性：KMC、LAC、P-MLACE 以及本文所提出的 MLACE。P-MLACE 算法思想来源于文献[11]。为了比较多扰动和单一扰动算法的正确性，P-MLACE 算法仅根据不同的参数值构造融合系统，依据文献[11]定义的最大子空间加权距离计算数据属于类的概率，融合算法采用 Fast global k-means。

LAC 算法中参数 γ 的取值设为[0,1]区间的实数，实验将在[0,1]区间内均匀取值 10 次，最终的实验结果为 10 次实验结果的平均值。融合算法 MLACE 和 P-MLACE 涉及到 3 个参数。通常参数 γ 在[0,1]区间内随机取值，形成参数扰动；基聚类个数设为 $B=\{2,5,10,15,20,25,30\}$，后续实验将分析基聚类算法个数与融合正确性的关系。特征子集的比例设为 $1/por$，当选取部分特征融合时，可以忽略部分噪音特征；但是选取特征较少时，容易造成信息损失。因此融合算法中 $1/por$ 取值位于区间[0.5,1]，每个基聚类算法随机确定 $1/por$ 值，并随机选取相关特征，形成特征扰动。

实验结果的评价选用聚类正确性(Clustering accuracy，CA)和标准化互信息(Normalized mutual information，NMI)。CA 是一个简单常用的聚类性能评价指标，但是不能评价错误数据的类别分布。而 NMI 可以度量不同数据类的真实分布 π^{truth} 和聚类分布 π^* 之间的相似性。设 n_k 是真实数据分布 π^{truth} 中类 C_k 所包含的数据个数，n_l 是聚类分布 π^* 中类 C_l 所包含的数据个数，n_{kl} 是真实数据分布 π^{truth} 中类 C_k 和聚类分布 π^* 中类 C_l 所共同包含的数据个数。NMI 通过式(7)度量真实分布和聚类分布之间的相似性：

$$NMI(\pi^*,\pi^{truth})=\frac{\sum\limits_{k,l=1\cdots K} n_{kl}\log\left(\frac{n\cdot n_{kl}}{n_k\cdot n_l}\right)}{\sqrt{\left(\sum\limits_k n_k\log\frac{n_k}{n}\right)\left(\sum\limits_l n_l\log\frac{n_l}{n}\right)}} \tag{7}$$

当 NMI=1 时，聚类分布与真实分布一致，聚类结果最优。随着 NMI 的下降，聚类算法的性能也越来越差。

3.3　实验结论

K-means、LAC、P-MLACE 和 MLACE 算法聚类正确性 CA 如表 3 所列，聚类的标准化互信息 NMI 如表 4 所列。MLACE 处理不同基聚类个数的实验结果如图 3 所示。CA 和 NMI 在数值上有些差异，但是总体趋势一致。根据实验结果，可以得到以下实验结论。

(1)KMC 的聚类正确性受到初始化和噪音特征的影响，因此聚类结果较差，尤其是噪音特征影响严重。如 2-banana、3-Gaussian、Wave 和 Wine 数据库均含有噪音特征，KMC 与其它算法聚类正确性存在较大的差距。

(2)软子空间聚类算法 LAC 能够减少噪音特征对聚类的影响，但是却无法预估参数 γ 的取值。当参数 γ 的取值不同时，其聚类指标 CA 和 NMI 均存在差异，平均后聚类正确性低于融合算法的正确性，如 2-banana 和 3-Gaussian。

(3)融合算法的聚类正确性优于单个聚类算法 KMC 和 LAC 的正确性。由于融合算法能够从不同的角度探测数据的内部结构，避免算法陷入局部极小，因此融合算法的聚类正确性均优于单个聚类算法。但是 P-MLACE 仅采用了单一的扰动方式，因此其正确性改进有限。如 Liver disorder 数据库，P-MLACE 与 LAC 的两个聚类指标 CA 和 NMI 均相当。

(4)MLACE 算法正确性优于单聚类算法 KMC 和 LAC，以及融合算法 P-MLACE 算法。原因如下：(i)MLACE 引入了多个扰动，可以探测数据的多个侧面，避免了初始化、参数和特征子集对算法正确性的影响；(ii)MLACE 计算的优化融合信息的区分度大于 P-MLACE 算法；(iii)MLACE 采用了实数值融合算法，其聚类正确性优于符号聚类。综上可以确保 MLACE 聚类正确性较优。对于所有数据，MLACE 的聚类正确性最优，但是对于不同的数据库，MLACE 聚类正确性的改进不同。对于 Breast cancer 和 Liver disorder，MLACE 的聚类正确性改进较小；而对于 2 个仿真数据库和 Wine 数据库，MLACE 聚类正确性改进较大。

(5)基聚类的个数 B 对融合系统 MLACE 的聚类正确性 CA 的影响如图 3 所示。当基聚类个数较少时，融合正确性不是很稳定，但是随着基聚类个数的增加，融合正确性逐渐趋于稳定。对于 2 个仿真数据库而言，$B=10$，MLACE 的聚类正确性趋于稳定。

表 3　4种算法聚类正确性 CA 比较

数据库名	KMC	LAC	P-MLACE	MLACE
2-banana	0.83	0.88	0.94	0.95
3-Gaussian	0.70	0.84	0.86	0.88
Breast cancer	0.85	0.88	0.89	0.90
Liver disorder	0.55	0.56	0.56	0.57
Sonar	0.53	0.54	0.56	0.57
Wave	0.51	0.73	0.75	0.77
Wine	0.70	0.90	0.91	0.93

表 4　4种算法聚类标准化互信息 NMI 比较

数据库名	KMC	LAC	P-MLACE	MLACE
2-banana	0.63	0.74	0.80	0.82
3-Gaussian	0.66	0.73	0.74	0.76
Breast cancer	0.70	0.73	0.73	0.75
Liver disorder	0.40	0.41	0.41	0.41
Sonar	0.37	0.37	0.38	0.39
Wave	0.48	0.55	0.57	0.58
Wine	0.53	0.72	0.73	0.75

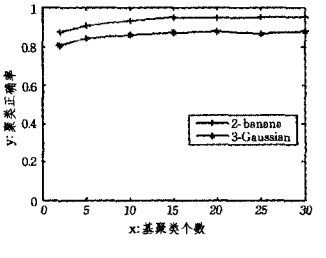


图 3　不同基聚类个数的 MLACE 处理仿真数据聚类正确性示意图

结束语　本文研究了基于多扰动的 LAC 聚类融合算法(MLACE)。实验表明：MLACE 的聚类正确性优于 KMC、LAC 和 P-MLACE。MLACE 算法正确性的改进取决于以下 3 个方面，分别是：(i)基于初始化、参数和特征子集映射的多

扰动；(ii)融合信息计算准确；(iii)实数值融合算法性能优越。本文重点研究 LAC 的聚类融合算法，但是本文所提出的软子空间聚类融合算法的框架不限于 LAC 算法。未来，将深入研究其它软子空间聚类融合算法及其参数之间的关系。另外，本文实验中选取特征子集的比例 $1/por$ 为随机值，对于不同数据库该值可能有所不同，未来我们将同时深入研究不同数据库与选取特征子集的比例 $1/por$ 之间的关系。

参 考 文 献

[1] Kriegel H P, Kroger P, Zimek A. Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering [J]. ACM Transactions on Knowledge Discovery from Data, 2009, 3(1): 1-58

[2] Parsons L, Haque E, Liu H. Subspace Clustering for High Dimensional Data: A Review [J]. ACM SIGKDD Explorations Newsletter-Special issue on learning from imbalanaced datasets, 2004, 6(1): 90-105

[3] Huang J Z. Automated variable weighting in K-means type clustering [J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657-668

[4] Gan G, Wu J. A convergence theorem for the fuzzy subspace clustering (FSC) algorithm [J]. Pattern Recognition, 2008, 41(6): 1939-1947

[5] Domeniconi C. Locally adaptive metrics fore clustering high dimensional data [J]. Data mining knowledge discovery, 2007, 14: 63-97

[6] 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

[7] 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

[8] Chen X J, Ye Y M, Xu X F, et al. A feature group weighting method for subspace clustering of high-dimensional data [J]. Pattern Recognition, 2012, 45(1): 434-446

[9] Xia Hu, Zhuang Jian, Yu De-hong. Novel Soft Subspace Clustering with Multi-objective Evolutionary Approach for High-dimen-sional Data [J]. Pattern Recognition, 2013, 46(9): 2562-2575

[10] 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

[11] Domeniconi C, Al-Razgan M. Weighted Cluster Ensembles: Methods and Analysis [J]. ACM Trans. Knowledge Discovery from Data, 2009, 2(4): 1-40

[12] Fred A L N, Jain A K. Combining Multiple Clusterings Using Evidence Accumulation [J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27(6): 835-850

[13] Kuncheva L I, Vetrov D P. Evaluation of stability of k-means cluster ensembles with respect to random initialization [J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2006, 28(11): 1798-1808

[14] Fern Xiaoli Z, Brodley Carla E. Random projection for high dimensional data clustering: a cluster ensemble approach [C] // Proceedings of 20th International Conference on Machine learning (ICML2003). Washington, DC, USA: AAAI Press, 2003: 186-193

[15] Likas A, Vlassis N, Verbeek J J. The Global k-Means Clustering Algorithm [J]. Pattern Recognition, 2003, 36: 451-461

[16] Strehl A, Ghosh J. Cluster Ensembles—A Knowledge Reuse Framework for Combining Multiple Partitions [J]. Journal of Machine Learning Research, 2002, 3: 583-617

[17] Iam-On 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

(上接第 205 页)

[2] Liben-Nowell D, Kleinberg J. The link-prediction problem for social networks[J]. Journal of the American society for information science and technology, 2007, 58(7): 1019-1031

[3] Al Hasan M, Chaoji V, Salem S, et al. Link prediction using supervised learning[C] // SDM'06: Workshop on Link Analysis, Counter-terrorism and Security. 2006

[4] Fire M, Tenenboim L, Lesser O, et al. Link prediction in social networks using computationally efficient topological features[C] // Privacy, security, risk and trust (PASSAT), 2011 IEEE third international conference on and 2011 IEEE third international confernece on social computing (SocialCom). 2011: 73-80

[5] Heckerman D, Geiger D, Chickering D M. Learning Bayesian networks: The combination of knowledge and statistical data [J]. Machine learning, 1995, 20(3): 197-243

[6] Taskar B, Abbeel P, Koller D. Discriminative probabilistic models for relational data[C] // Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence. 2002: 485-492

[7] Hopcroft J, Lou T, Tang J. Who will follow you back?: reciprocal relationship prediction[C] // Proceedings of the 20th ACM international conference on Information and knowledge management. 2011: 1137-1146

[8] Kunegis J, Lommatzsch A. Learning spectral graph transformations for link prediction[C] // Proceedings of the 26th Annual International Conference on Machine Learning. 2009: 561-568

[9] 樊鹏翼，王晖，姜志宏，等. 微博网络测量研究[J]. 计算机研究与发展，2012，49(4)：691-699

[10] 周刚，邹鸿程，熊小兵，等. MB-SinglePass：基于组合相似度的微博话题检测[J]. 计算机科学，2012，39(10)：198-202

[11] Adamic L A, Adar E. Friends and neighbors on the web[J]. Social networks, 2003, 25(3): 211-230

[12] Katz L. A new status index derived from sociometric analysis [J]. Psychometrika, 1953, 18(1): 39-43

[13] Breiman L. Random forests[J]. Machine learning, 2001, 45(1): 5-32

[14] Olshen R, Breiman L, Friedman J H, et al. Classification and Regression Trees[M]. Wadsworth International Group, 1984

[15] Breiman L. Bagging predictors[J]. Machine learning, 1996, 24(2): 123-140

[16] 苏金树，张博锋，徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报，2006，17(9)：1848-1859

[17] Chawla N V, Japkowicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets[J]. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 1-6

[18] 中国爬盟[EB/OL]. http://www.cnpameng.com, 2012