具有稀疏特征的对象—属性子空间边缘重叠区域归属算法*

祝 琴, 陈 华
(南昌大学 管理科学与工程系, 南昌 330031)

摘 要: 通过分析具有稀疏特征的对象—属性子空间的特征,发现其边缘存在交叉重叠区域现象,为此,提出了基于聚类思想的具有稀疏特征的对象—属性子空间边缘的重叠区域归属算法(OASEDA),该算法能有效解决对象—属性子空间的独立性,算法根据子空间内部紧凑度和子空间之间分离度相对大小确定子空间边缘重叠区域的归属,并基于 K-means 算法结合权重理论设计了重叠区域归属判断目标函数,最后通过实验证明了该方法的有效性。

关键词: 具有稀疏特征的高维数据; 对象—属性子空间; 对象—属性子空间边缘重叠区域
中图分类号: TP301.6; TP391 **文献标志码**: A **文章编号**: 1001-3695(2013)01-0099-04
doi:10.3969/j.issn.1001-3695.2013.01.023

Object-attribute subspace with sparse feature edges detection

ZHU Qin, CHEN Hua
(*Dept. of Management Science & Engineering, Nanchang University, Nanchang* 330031, *China*)

Abstract: The overlapped regions among the identified objects-attributes subspaces by the traditional algorithm could influence the independence of these subspaces. In order to solve this defect, this paper developed the objects-attributes subspace edges detection algorithm(OASEDA) based on K-means. It designed the objective function of edge detection, algorithm with the information of within-cluster and between-cluster, and optimized the objective function by the weight theory. In the end, experimental results on synthetic datasets demonstrate that the accuracy of the proposed algorithm.

Key words: high-dimensional data with high dimension sparse feature; object-attribute subspace; overlapped region among object-attribute subspace

高维数据是比较常见的一种数据形式,具有稀疏性。随着应用的深入发展,如何从这些具有稀疏特征的高维数据集中挖掘出对用户有用的知识,是目前数据挖掘领域中重要的研究内容之一[1~4]。受维度效应的影响,传统的聚类算法不能适用于高维数据[5]。经典高维数据聚类算法包括网格聚类算法、密度聚类算法等,而近年来提出的子空间聚类算法因高效、准确的聚类结果而备受关注[6~8]。子空间聚类的前提和基础是子空间的识别。事实上,子空间识别的研究已经成为高维数据预处理的重要组成部分,并且子空间的质量直接影响最终的子空间聚类,因此,该问题已经引起了学者的关注,正成为当前高维数据聚类研究的热点和难点。二阶段联合聚类算法(MTPCCA)是从聚类的角度来研究具有稀疏特征的高维数据对象—属性子空间的识别问题,能够识别出具有较高质量的对象—属性子空间[9],但该算法识别出的子空间边缘容易出现边界不清的现象,即子空间边缘存在重叠区域。如图 1 所示,区域 C 既可以认为是对象—属性子空间 A 的边缘区域,也可以认为是对象—属性子空间 B 的边缘区域,这一部分本文定义为对象—属性子空间重叠区域 C。因此,确定该重叠区域的归属对提高子空间质量、减小具有稀疏特征的高维数据预处理时的搜索空间,甚至对具有稀疏特征的高维数据聚类都是非常重要的。鉴于这一点,本文针对具有稀疏特征的对象—属性子空间边缘重叠区域的归属问题提出了子空间边缘检测算法,提高子空间识别的质量。

1 高维数据聚类边缘问题研究现状

与传统数据聚类算法相比,高维数据聚类算法能有效解决高维数据聚类问题,具有聚类效率高、准确度高等优点。但研究发现,高维数据聚类算法普遍存在边界效应现象,即聚类边界不清。这一问题引起了学者们的广泛关注,提出了诸多算法。从网格划分的角度研究,提出了如 MAFIA、GDCAP[10]、GCOD[11]和 CGDCP[8]等算法,这类算法在一定程度上克服了单纯网格划分后可能出现的类边缘划分不准的情况,但无法确定边缘网格内数据点的归属;从预设阈值角度研究聚类边界点归属问题,如 OptCLIQUE 聚类[12]、GDDEA[13]等算法,这类算法通过边界点的阈值函数解决边界点归属问题,但其聚类的精度对边界点的阈值非常敏感。此外,基于密度聚类的方法均在一定程度上受密度参数的影响,或者说边界部分的识别对密度参数是敏感的,因此其聚类边界的识别是有限的。

结合具有稀疏特征的对象—属性子空间的特点,本文从聚类的角度提出子空间边缘检测算法,以确定子空间边缘重叠区

收稿日期: 2012-05-04; **修回日期**: 2012-06-26 **基金项目**: 国家自然科学基金资助项目(60963008)

作者简介: 祝琴(1978-),女,江西临川人,讲师,博士,主要研究方向为数据挖掘、管理信息系统和计算机控制(zhuqin189@yahoo.com.cn); 陈华(1975-),女,安徽淮南人,教授,博士,主要研究方向为知识管理.

域的归属,解决高维数据聚类边界不清的问题。

2 具有稀疏特征的对象—属性子空间边缘重叠区域分析

如图 1 所示,对象—属性子空间重叠区域 C 是既是子空间 A 的边缘区域,又属于子空间 B 的边缘区域,即子空间 A 和子空间 B 边缘交叉重叠区域,$A \cap B = C$。这部分重叠区域 C 应属于子空间 A 还是 B,或者说 C 的归属问题直接关系到子空间 A 和 B 的构成,不仅影响到它们的质量,而且对最终数据挖掘的结果也会产生影响。如对具有稀疏特征的对象—属性子空间运行相关的数据挖掘算法时,该重叠区域 C 中的数据至少被扫描两次:作为子空间 A 的边缘被扫描一次,而作为子空间 B 的边缘同样被扫描一次,这样直接增加了该高维数据挖掘算法的复杂度。由于具有稀疏特征的对象—属性子空间分布的数据点是稀疏的,如果出现该子空间边缘重叠区域内分布全为 0 的数据点现象,如图 2 所示,即 $C = A \cap B$,且 $C = \{0,0;0,0\}$,重叠区域 C 中没有非零数据点分布。考虑到对象的这些零值属性对最终挖掘结果基本没有影响,因此这种情况不在本文的研究范围之内,也就是说本文研究的是子空间重叠区域内有非零数据点分布的情况。

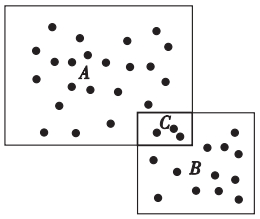

图1 对象—属性子空间重叠区域

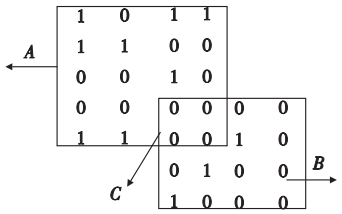

图2 重叠区域中零属性分布

3 基于内聚度和分离度的子空间边缘检测算法

综上所述,具有稀疏特征的对象—属性子空间之间的重叠区域会直接影响相邻子空间的组成,因此将影响子空间的质量。本文从聚类的角度提出针对具有稀疏特征的对象—属性子空间边缘检测算法,以解决子空间边缘重叠区域归属问题,实现对象—属性子空间的相对独立,提高子空间质量,从而改善具有稀疏特征的高维数据预处理效果,并为最终提高数据挖掘的质量奠定基础。

如图 1 所示的子空间 A 与 B 的边缘重叠区域 C,本文将其看做一个特殊的子空间进行研究,因此,确定子空间边缘重叠区域 C 的归属问题就转换为对象—属性子空间 C 的聚类问题。

3.1 算法思想

根据牛顿万有引力原理,设某一物体的受力分析如图 3 所示,由力学知识可知 $F = F_1 + F_2$,而物体合力 $F' = F - F_3$,假设物体的运动轨迹是圆周时,则 F 即为向心力,而 F_3 则为离心力。该物体的运动轨迹取决于物体所受向心力和离心力的相对大小,即合力的大小。如果 $F' > 0 (F > F_3)$,即受到的向心力大于离心力,则物体继续做圆周运动;若 $F' < 0 (F < F_3)$,所受到的向心力小于离心力,则物体做离心运动。

聚类中的内聚度与物理学中的向心力具有相似性,分离度则与离心力相仿。受牛顿第二定律思想的启发,本文提出具有稀疏特征对象—属性子空间边缘检测算法(objects-attributes subspace edges detection algorithm,OASEDA)来研究子空间边缘重叠区域的归属问题,因此子空间边缘重叠区域应归属到哪个相邻子空间取决于归属后的子空间内聚度与分离度的相对大小。

3.2 子空间边缘重叠区域划分策略

结合具有稀疏特征的对象—属性子空间形状为矩形的特点,将相邻子空间沿重叠区域部分边界进一步细分。如图 4 所示,将重叠区域 C 沿相邻子空间 A 和 B 分别细分为 A_1 和 A_2、B_1 和 B_2。

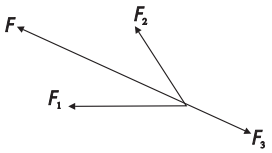

图3 受力分析

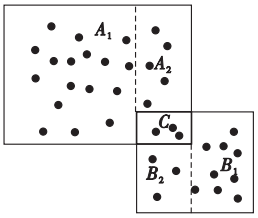

图4 子空间分块

3.3 OASEDA

OASEDA 基于 K-means,从子空间内部紧凑度与子空间之间差异度的角度研究子空间边缘重叠区域归属问题。本文首先设计计算重叠区域 C 的归属系数的目标函数,并结合对象属性的权重关系优化目标函数,其值作为相似度(归属度)。具体步骤如下。

通过二阶段联合聚类 MTPCCA 完成对象—属性子空间的识别,其实质是对具有稀疏特征的高维数据的预处理,识别出的子空间仍然具有稀疏特征。其中央为数据密集区,且密度沿中央向边缘部分递减,边缘重叠区域内分布的数据点可能更加稀疏,即边缘重叠区域内分布的非零数据点数目很小。

当前对于权重的研究主要分为模糊函数、信息熵两类。本文利用模糊函数来研究属性的权重[14,15]。

定义 1 重叠区域归属系数 γ 为交叉重叠区域与相邻对象—属性子空间之间的相似度。

归属系数 γ 越大,相似度越大,则认为该交叉重叠区域应与其对象—属性子空间聚为一类。

归属系数 γ:重叠区域的归属系数即是该区域与子空间相似度。

根据 K-means 思想,γ 越大的两个区域聚为一类。其计算公式为

$$\gamma = \frac{1}{J(X,C,W)} \tag{1}$$

其中:X 为子空间 A 或 B;C 是子空间 A 和 B 之间的重叠区域;W 为熵权重系数。故有

$$\gamma_1 = \frac{1}{J(A,C,W_1)}, \gamma_2 = \frac{1}{J(B,C,W_2)}$$

如果 $\gamma_1 > \gamma_2$,则 $A \supset C$;即重叠区域 C 与子空间 A 更相似,故重叠区域 C 应与子空间 A 合并,或者说重叠区域 C 应归属到子空间 A;否则 $B \supset C$,重叠区域 C 归属子空间 B。

重叠区域的归属系数计算包括子空间的内部紧凑度和子空间之间的分离度两部分。K-means[16] 是一种 EM 型算法[17],它在迭代过程中不断更新数据集的划分,用于优化以下目标函数:

$$R_0(C,V) = \sum_{k=1}^{K} \sum_{\tilde{x}_i \in C_k} \| \tilde{x}_i - \tilde{v}_k \|^2 \tag{2}$$

K-means 在全空间搜索数据集的最优划分,记号$\hat{x}_i$、$\hat{v}_k$ 分别表示全空间中的第 i 个数据点和 k 个划分的中心,$\|\cdot\|$ 表示 L_2范数。

1)子空间内紧凑度的计算

a)将对象—属性空间沿重叠区域划分子空间,得对象—属性子空间 A、B 和 C,如图 4 所示,则 $A=A_1\cup A_2$,$B=B_1\cup B_2$,$A\cap C=\varnothing$,$B\cap C=\varnothing$,$A\cap B=\varnothing$。

b)根据借鉴 Huang 等人[18]提出的软子空间聚类 FWKM 算法中的内紧凑度计算思想,则子空间的内紧凑度计算方法为:设 $A'=A\cup C$, 则子空间 A'的数据空间 $DB=\{X_1,X_2,\cdots,X_N\}$,其中 $X_i=\{X_{i1},X_{i2},\cdots,X_{iD}\}$,$X_i$ 为 $D(D>1)$维数据空间的第 i 个数据点($i=1,2,\cdots,N$)。$N(N>1)$表示数据点数目,$K(K>1)$是给定的簇数目,本文取 $K=2$。

$$J(C,V,W)=\sum_{k=1}^{K}\sum_{j=1}^{D}w_{kj}^{\beta}\sum_{x_i\in C_k}(x_{ij}-v_{kj})^2 \tag{3}$$

$$\text{s.t.}\quad u_{ij}\in\{0,1\},\sum_{i=1}^{C}u_{ij}=1,0\leqslant w_{ij}\leqslant 1,\sum_{k=1}^{D}w_{ik}^{\tau}=1$$

其中:$w_k=\langle w_{k1},w_{k2},\cdots,w_{kD}\rangle$和 $v_i=\langle v_{k1},v_{k2},\cdots,v_{kD}\rangle$分别表示 C_k 的维度权值和簇中心向量,且$\sum_{j=1}^{D}w_{kj}=1(k=1,2,\cdots,K)$;$V=\{v_{kj}\}_{k\times D}$和 $W=\{w_{kj}\}_{k\times D}$是两个矩阵;β 为用户定义的加权参数,其作用是调节权值的影响力。

2)簇间分离度计算

如图4 所示 $C\cap B=\varnothing$,为了计算子空间 C 与 B 的分离度,设 $B'=B\cup C$,$C=2$。这样样本空间由原子空间 A、B 和 C 变成子空间 A'和 B',相应地,样本数据发生了变化,对应的 N 和 D 都变为 N'和 D'。

$$J_{s_fw}=\sum_{i=1}^{C}\Big(\sum_{j=1}^{N'}u_{ij}^{m}\Big)\sum_{k=1}^{D'}w_{ik}^{\tau'}(v_{ik}-v_{Ok})^2 \tag{4}$$

$$v_0=\frac{\Big(\sum_{j=1}^{N}X_j\Big)}{N}$$

$$\text{s.t.}\quad u_{ij}^{'}\in\{0,1\},\sum_{i=1}^{C}u_{ij}^{'}=1,0\leqslant w_{ij}^{'}\leqslant 1,\sum_{k=1}^{D}w_{ik}^{\tau'}=1$$

3)结合对象属性权重构造目标函数

根据 EM 的原理,常用的解决办法是将基于 W 和 V 的局部最优转换为解决目标函数 J 的最优化。结合式(2)和(3),为计算 W 和 V 的局部最优值,在 J 的基础上引入 W_{ij}和 $W_{ij}^{'}$的约束条件构造拉格朗日优化函数 J:

$$\begin{aligned}J(C,V,W,U,W')=&\sum_{k=1}^{K}\sum_{j=1}^{D}w_{kj}^{\beta}\sum_{x_i\in C_k}(x_{ij}-v_{kj})^2-\\&\sum_{i=1}^{C}\Big(\sum_{j=1}^{N'}u_{ij}^{m}\Big)\sum_{k=1}^{D'}w_{ik}^{\tau'}\left(v_{ik}-\frac{\Big(\sum_{j=1}^{N}X_j\Big)}{N}\right)^2+\\&\lambda_1\sum_{j=1}^{N'}\Big(\sum_{i=1}^{C}u_{ij}-1\Big)+\lambda_2\sum_{i=1}^{C}\Big(\sum_{k=1}^{D}w_{ik}^{\tau}-1\Big)+\\&\lambda_3\sum_{i=1}^{C}\Big(\sum_{k=1}^{D}w_{ik}^{\tau'}-1\Big)\end{aligned} \tag{5}$$

其中,λ_1、λ_2、λ_3 为拉格朗日乘子。

$$\frac{\partial J}{\partial w_{ij}}=0,\frac{\partial J}{\partial v_{ij}}=0,\frac{\partial J}{\partial w_{ij}^{'}}=0$$

$$v_{kj}=\frac{1}{|C_k|}\sum_{x_i\in C_k}^{n}x_{ij}$$

$$W_{kj}=\left[\sum_{l=1}^{n}X_i\left[\frac{\sum_{x_i\in C_k}^{n}(x_{ij}-v_{kj})^2}{\sum_{x_i\in C_k}^{n}(x_{il}-v_{kl})^2}\right]^{\frac{1}{\beta-1}}\right]^{-1}$$

同理,计算 γ_2 时,内紧凑度 $B'=B\cup C$,分离度 $A'=A\cup C$。

3.3.1　算法步骤

以上分析的具体步骤如下:

a)建立对象—属性二维表,根据实际对象的取值情况预设对象属性取值的阈值为 b,并进行归一化处理,获得原对象—属性稀疏二维表。

b)由对象—属性稀疏二维表建立具有稀疏特征的对象—属性空间。

c)运用二阶段联合聚类算法完成对象—属性子空间的识别,并判断识别出的子空间边缘是否存在重叠区域,如果有,则进行步骤 d)。

d)判断对象—属性子空间边缘的重叠区域内分布的对象其属性取值情况,如果其属性取值不完全为 0,则进行步骤 e),否则算法结束。

e)根据子空间边缘重叠区域划分策略,将相邻子空间 A 和 B 沿重叠区域 C 分块,如图 4 所示,则 $A=A_1\cup A_2$,$B=B_1\cup B_2$,$A\cap B=C$。

f)分别计算边缘重叠区域 C 的归属系数 γ_1 和 γ_2 值。根据式(4)分到计算边缘重叠区域 C 的归属系数值。如果 $\gamma_1>\gamma_2$,则 $C\supset B$,子空间 C 应与子空间 A 合并,C 归属到子空间 A;否则 $C\subset B$,C 归属到子空间 B。

3.3.2　算法时间复杂度分析

算法 OASEDA 时间复杂度可以表示为:$T=O(nmk)$。其中:n 为具有稀疏特征的对象—属性子空间中数据集所含的对象个数;m 为描述对象的属性数目;k 为与边缘重叠区域相邻的子空间数目。

4　算例分析

假设有 8 个客户对象,记为 $O_i(i\in\{1,2,\cdots,8\})$。描述每个对象的属性有 10 个,分别为该对象对 10 种产品的订购量,记为 $A_j(j\in\{1,2,\cdots,10\})$,如表 1 所示。现在需要根据这 8 个客户对 10 种产品订购的情况进行对象维和属性维的预处理,识别其中的对象—属性子空间。

表 1　8 个客户对 10 种产品的订购量

客户	产品 1	产品 2	产品 3	产品 4	产品 5	产品 6	产品 7	产品 8	产品 9	产品 10
1	0	0	0	0	0	0	60	0	0	0
2	0	180	0	260	90	0	0	0	0	180
3	150	0	360	300	0	500	70	0	160	0
4	0	0	0	0	100	0	0	80	0	350
5	0	120	180	120	0	60	560	0	300	420
6	320	0	0	280	600	0	0	0	270	0
7	0	500	350	350	0	480	500	120	0	450
8	400	0	0	0	0	0	480	0	0	380

a)对表 1 进行归一化处理,使所有对象的属性取值在[0,1]区间。经过标准化处理后的数据如表 2 所示。

表 2　8 个客户对 10 种产品的订购量归一化

客户	产品 1	产品 2	产品 3	产品 4	产品 5	产品 6	产品 7	产品 8	产品 9	产品 10
1	0	0	0	0	0	0	0.6	0	0	0
2	0	0.18	0	0.26	0.09	0	0	0	0	0.18
3	0.15	0	0.36	0.3	0	0.5	0.07	0	0.16	0
4	0	0	0	0	0.1	0	0	0.08	0	0.35
5	0	0.12	0.18	0.12	0	0.06	0.56	0	0.3	0.42
6	0.32	0	0	0.28	0.6	0	0	0	0.27	0
7	0	0.5	0.35	0.35	0	0.48	0.5	0.12	0	0.45
8	0.4	0	0	0	0	0	0.48	0	0	0.38

b)设稀疏判断阈值 $b_j=0.2$,对表 2 进行归一化处理得到

其稀疏特征表，如表 3 所示。

表 3 8 个客户订购 10 种产品的稀疏特征值表

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
O_1	0	0	0	0	0	0	1	0	0	0
O_2	0	0	0	1	0	0	0	0	0	0
O_3	0	0	1	1	0	1	0	0	0	0
O_4	0	0	0	0	0	0	0	0	0	1
O_5	0	0	0	0	0	0	1	0	1	1
O_6	1	0	0	1	1	0	0	0	1	0
O_7	0	1	1	1	0	1	1	0	0	1
O_8	1	0	0	0	0	0	1	0	0	1

c)根据 8 个客户对 10 种产品订购情况的稀疏特征表，得到对象—属性空间图，如图 5 所示。

	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
O_1	0	0	0	0	0	0	1	0	0	0
O_2	0	0	0	1	0	0	0	0	0	0
O_3	0	0	1	1	0	1	0	0	0	0
O_4	0	0	0	0	0	0	0	0	0	1
O_5	0	0	0	0	0	0	1	0	1	1
O_6	1	0	0	1	1	0	0	0	1	0
O_7	0	1	1	1	0	1	1	0	0	1
O_8	1	0	0	0	0	0	1	0	0	1

图 5 8 × 10 对象—属性空间

d)运用二阶段协同聚类算法对对象—属性空间进行聚类分割，获得的高维稀疏对象—属性子空间，如图 6 所示，即对象—属性子空间 A 和 B，且 $C = A \cap B$，即对象—属性子空间的重叠区域 C。

e)基于 OASEDA 思想，根据式(5)计算。将交叉重叠区域 C 作为一个独立区域研究，由图 6 可得其中数据点的分布，令

$$P = A - C = \left\{ \begin{matrix} (1,5),(2,5),(2,6),(3,4),(3,5), \\ (3,6),(3,8),(4,5),(4,6),(5,7) \end{matrix} \right\}$$

$$C = \{(5,5),(6,4)\}$$

$$Q = B - C = \left\{ \begin{matrix} (5,2),(5,3),(6,2),(7,3),(7,4), \\ (8,1),(8,2),(8,3),(8,5),(9,4) \end{matrix} \right\}$$

设区域 P 和 Q 的中点分别为 O_1 和 O_2，则 $O_1 = (3, 5.6)$，$O_2 = (7.1, 2.9)$，设权值的取值为

$w_1 = w_2 = \cdots = w_{10} = 0.1$

$w_1' = w_2' = \cdots = w_{10}' = 0.1$

所以 $J_1 = 1.592$，$J_2 = 1.124$

因为 $\gamma = 1/J$

所以 $\gamma_1 < \gamma_2$，$B \supset C$

即交叉重叠区域 C 应归属到相邻子空间 B。算例中由于 $y_1 < y_2$，因此 $C \subset B$，如图 7 所示。因此，原对象—属性空间识别出的子空间分别为 B、A_1、A_2 和 A_3。

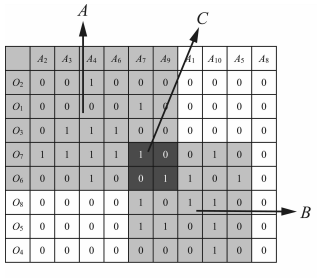

图6 别出的8×10 对象—属性子空间

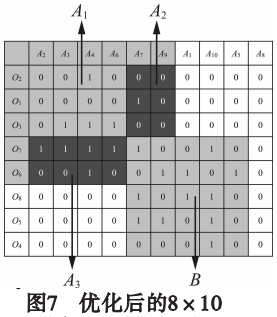

图7 优化后的8×10 对象—属性子空间

5 实验及结果分析

某钢铁销售公司某产品有 26 个客户对象，记为 O_i($i \in \{1,2,\cdots,26\}$)，该产品有 45 种型号，即描述每个对象的属性有 45 个，记为 A_j($j \in \{1,2,\cdots,45\}$)，26 个客户订购这 45 种产品的数量如表 4 所示，其对应的对象—属性空间图如图 8 所示。现需要对该对象—属性空间进行数据预处理。

表 4 26 个客户订购 45 种产品的数量

对象序号	取值为 1 的属性序号集	对象序号	取值为 1 的属性序号集
1	2,3,4,6,12,23,25,26,30,32,45	14	5,19,24,31,33,38,41
2	5,16,19,24,27,33,38,44	15	8,10,15,18,29,35,37
3	4,6,7,13,15,25,26,28,35,45	16	2,4,6,7,12,15,23,28,30,32
4	1,3,9,10,15,22,29,37	17	10,13,15,17,36,40,43
5	3,8,9,18,22,34,35,37,42	18	9,15,18,22,34,35,42
6	11,16,21,27,31,33,38,41	19	11,16,21,24,27,31,41,44
7	5,11,19,21,24,31,38,44	20	1,3,8,10,22,29,34,35,37
8	1,3,8,9,10,15,18,22,29,34,35,37,42	21	5,11,19,24,31,33,41,44
9	1,9,10,15,22,29,34,35,42	22	2,3,6,12,13,15,25,26,28,35
10	1,8,10,15,18,29,37,42	23	2,3,6,12,23,25,28,30,32,35,45
11	11,19,21,24,27,33,38,41,44	24	4,8,16,18,23,38,39,42,
12	3,4,6,12,15,23,25,26,32,35,45	25	4,7,12,13,15,23,26,30,32,35,45
13	2,4,7,12,13,25,26,28,30,32,45	26	4,7,12,15,23,26,30,32,35

运用二阶段联合聚类 MTPCCA 算法识别出的子空间如图 9 所示。其中子空间 A 和 B 边缘存在重叠区域 C。

将子空间边缘重叠区域 C 作为一个独立对象，运用 OASEDA 研究其归属问题，计算对应的归属系数 γ_1 和 γ_2。由于 $\gamma_1 < \gamma_2$，因此 $C \supset B$，即重叠区域 C 应合并到子空间 B，如图 10 所示。OASEDA 解决了重叠区域 C 的归属问题，不仅提高了相邻子空间 A 和 B 的质量，而且提高了具有稀疏特征的高维数据预处理的效果。

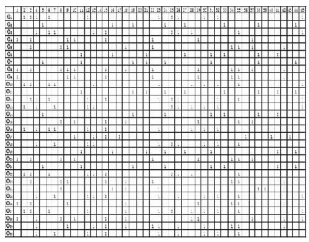

图8 对象—属性空间

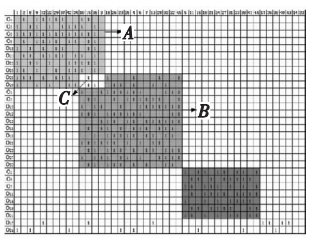

图9 对象—属性原子空间

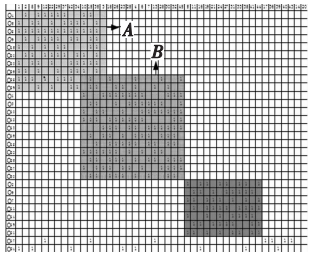

图10 优化后的对象—属性子空间

6 结束语

本文针对具有稀疏特征对象—属性子空间边缘存在重叠区域现象进行了研究，基于经典 K-means (下转第 113 页)

清晰地检测到两个正弦信号;从图 2(d)中可以看出,在窄高斯窗下很清晰地检测到了脉冲信号,但不能清晰地检测到两个正弦信号;由图 2(e)可以明显地看出,使用本文提出的多高斯窗进行检测,在多高斯窗下的 Gabor 能清晰地检测到脉冲信号,同时也能清晰地分辨出两个不同频率的正弦信号。图 2(f)是文献[2]中算法在宽高斯窗下的 Gabor 频谱图,虽然检测到了正弦信号,但时频平面上并不平坦;图 2(g)是文献[2]中算法在窄高斯窗下的 Gabor 频谱图,虽然检测到了脉冲信号,但是频谱两边粘滞不清;图 2(h)是文献[2]中算法在多高斯窗下的 Gabor 频谱图,虽然检测到正弦信号和脉冲信号,但是时频面上有其他频谱出现,而且不清晰。

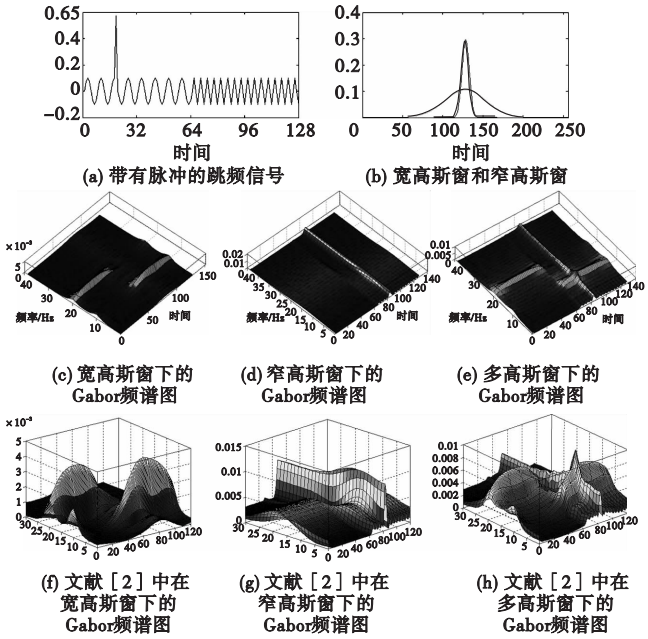

图2　带有脉冲的跳频信号变换分析

4　结束语

针对传统的单窗复值离散 Gabor 变换具有固定的时频分辨率,文献[2]给出了基于框架的多 Gabor 变换算法复杂度较高。本文在多 Gabor 变换的基础上提出了一种基于辅助双正交分析法的实值离散 Gabor 变换及其快速算法,该算法相比文献[2]算法减小了计算量。仿真实验证明提出的算法能明显改善传统 Gabor 变换的时频分辨率。

参考文献:

[1] GABOR D. Theory of communication[J]. Journal of the Institution of Electrical Engineers,1946,94(73):429-457.

[2] 陶亮,顾涓涓. 实值 Gabor 变换理论及应用[M]. 合肥:安徽科学技术出版社,2005.

[3] LI Shi-dong. Discrete multi-Gabor expansions[J]. IEEE Trans on Information Theory,1999,45(6):1954-1967.

[4] AKAN A,CHAPARRO L F. Multi-window Gabor expansion for evolutionary spectral analysis[J]. Signal Processing,1997,63(3):249-262.

[5] TAO Liang,KWAN H K. Fast parallel approach for 2-D DHT-based real-valued discrete Gabor transform[J]. IEEE Trans on Image Processing,2009,18(12):2790-2796.

[6] TAO Liang,KWAN H K. Novel DCT-based real-valued discrete Gabor transform and its fast algorithms [J]. IEEE Trans on Signal Processing,2009,57(6):2151-2164.

[7] MATUSIAK E, MICHAELI T, ELDAR Y C. Noninvertible Gabor transforms[J]. IEEE Trans on Signal Processing,2010,58(5):2597-2612.

[8] 徐婉莹,黄新生,刘育浩,等. 一种基于 Gabor 小波的局部特征尺度提取方法[J]. 中国图象图形学报,2011,16(1):72-78.

[9] 杨清山,郭成安,金明录. 基于 Gabor 多通道加权优化与稀疏表征的人脸识别方法[J]. 电子与信息学报,2011,33(7):1618-1624.

[10] LAGAE A,LEFEBVRE S,DUTRE P. Improving Gabor noise [J]. IEEE Trans on Visualization and Computer Graphics,2011,17(8): 1096-1107.

(上接第 102 页)算法提出了确定该重叠区域归属的判别 OASE-DA,并设计了归属判断的目标函数。由于对象—属性子空间边缘重叠区域与各个相邻子空间均有一定的从属关系,本文在确定其归属时,不仅考虑了最小化子空间的内紧凑度,而且还考虑了子空间之间的分离度。最后通过算例验证了该方法的有效性与可行性。

参考文献:

[1] HAN Jia-wei,KAMBER M. 数据挖掘概念与技术[M]. 范明,孟小峰,等译. 2 版. 北京:机械工业出版社,2007.

[2] 武森,高学东,巴斯蒂安 M. 数据仓库与数据挖掘[M]. 北京:冶金工业出版社,2003.

[3] VERLEYSEN M. Learning high-dimensional data [C]//Proc of NATO Advanced Research Workshop on Limitations and Future Trends in Neural Computation. 2001:141-162.

[4] YANG Qiang,WU Xin-dong. 10 challenging problems in data mining research[J]. International Journal of Information Technology and Decision Making,2006,5(4):597-604.

[5] 杨风召. 高维数据挖掘技术研究[M]. 南京:东南大学出版社,2007.

[6] 张燕萍,姜青山. K-means 型软子空间聚类算法[J]. 计算机科学与探索,2010,4(11):1019-1026.

[7] 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.

[8] 何虎翼,姚莉秀,沈红斌,等. 一种新的子空间聚类算法[J]. 上海交通大学学报,2007,41(5):557.

[9] 祝琴,高学东,武森,等. 高维稀疏数据对象—属性分割[J]. 数学的认识与实践,2011,41(7):184-189.

[10] 单世民,张宁,江贺,等. 基于网格和密度的簇边缘精度增强聚类算法[J]. 计算机工程与应用,2008,44(23):143-146.

[11] QIU Bao-zhi,LI Xiang-li,SHEN Jun-yi. Grid-based clustering algorithm based on intersecting partition and density estimation[C]//Proc of Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin:Springer,2007:368-377.

[12] 高亚鲁. 子空间聚类算法的研究及应用[D]. 江苏:江苏大学,2009.

[13] 余灿玲, 王丽珍,张元武. 基于网格密度方向的聚类簇边缘精度加强算法[J]. 计算机研究与发展,2010,7(5):815-823.

[14] 皋军,王士同. 具有特征排序功能的鲁棒性模糊聚类方法[J]. 自动化学报,2009,35(2):145-153.

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

[16] MacQUEEN J. Some methods for classification and analysis of multivariate observations[C]//Proc of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967:281-297.

[17] XU Lei,JORDAN M I. On convergence properties of the EM algorithm for Gaussian mixtures[J]. Neural Computation,1996,8(1):129-151.

[18] HUANG Zhe-xue,NG M K,RONG Hong-qiang,*et al*. Automated variable weighting in K-means type clustering[J]IEEE Trans on Pattern Analysis and Machine Intelligence,2005,27(5):657-668.