一种鲁棒的子空间聚类算法

彭柳青，张军英
（西安电子科技大学计算机学院，710071，西安）

摘要：针对聚类分析常面临的维数灾难和噪声污染问题，将样本加权思想与子空间聚类算法相结合，提出了一种鲁棒的子空间聚类算法.该算法结合现有子空间聚类方法，为每个类簇计算一个反映各维度聚类贡献程度的权矢量，并利用该权矢量对各维度加权组合，得到各类簇所处的子空间.此外，算法还为每个样本分配一个反映离群程度的尺度参数，以区分正常样本和离群点在聚类过程中的地位，保证算法的鲁棒性.在二维数据集、高维数据集以及基因数据集上的对比实验结果表明，对于具有不同噪声比例的各种维度数据集，该算法均能取得较高的聚类精度，表现出较好的鲁棒性.

关键词：子空间聚类；鲁棒性；权参数；最优化
中图分类号：TP391.4 **文献标志码**：A **文章编号**：0253-987X(2011)06-0013-07

A Robust Subspace Clustering Algorithm

PENG Liuqing, ZHANG Junying
(School of Computer Science and Engineering, Xidian University, Xi′an 710071, China)

Abstract: A new algorithm is presented to simultaneously solve the problems that clustering suffers from the curse of dimensionality as well as noise contamination. Following some existing idea, the algorithm associates a weight vector to each cluster in the entire data space, and captures the contribution degrees of dimensions for identifying the cluster. Different subspaces for discovering clusters are obtained by combining dimensions via those weight vectors. Furthermore, the algorithm assigns a scalar value to each sample to discriminate the role of outliers from that of normal samples during the clustering process; therefore, the robustness of the algorithm is guaranteed. Experimental results show that the proposed algorithm gains high clustering accuracy on datasets of different dimensions with various noise ratios added.

Keywords: subspace clustering; robustness; weight; optimization

对于高维空间数据，聚类分析常面临维数灾难问题.高维数据的特征维数非常多，常有成千上万，但其中许多特征维与聚类结果并不相关，容易发生同类样本在不相关维上距离较远的情况.针对此类问题，人们提出了各种高维数据聚类方法.子空间聚类算法[1-6]作为高维数据聚类的主要方法，一直是人们关注的重点.此类方法在划分类簇的同时，还指出各类簇所处的子空间，即允许不同类簇分布在不同子空间内.此外，还存在一类基于特征选择或特征提取的高维聚类算法[7]，该类方法可在同一特征空间内识别各类簇.比较而言，子空间聚类算法更具推广意义.

子空间聚类按类型可分为投影子空间聚类[2-3]和软子空间聚类[4-6].投影子空间聚类作为一种启发式搜索方法，为各类簇搜索若干重要维度，各类簇所在的子空间由这些重要维度直接组合张成.软子空间聚类是一类基于最优化算法的聚类方法，它为各类簇生成一个体现各维度贡献大小的权矢量，因此，

收稿日期：2010-11-24. 作者简介：彭柳青(1984—)，女，博士生；张军英(联系人)，女，教授，博士生导师. 基金项目：国家自然科学基金资助项目(61070137，60933009)；陕西省科技攻关资助项目(2009K1-56).
网络出版时间：2011-02-21 网络出版地址：http://www.cnki.net/kcms/detail/61.1069.T.20110221.1605.009.html

各类簇所处的子空间由权矢量对各维度加权组合张成.

运用于实际数据的算法还应具备鲁棒性[8-9]. 目前,健壮算法主要分为健壮算子法[10]和样本加权函数法[11]. 文献[10]利用数据深度算子在估计类簇中心时所具备的健壮能力,针对高维数据,提出了二分k-空间均值聚类算法. 该算法具备一定的鲁棒性,但不能区分各维度的聚类贡献程度,因此只适合处理含正圆形状类的数据. 样本加权函数法的基本思想是为各样本分配一个尺度值,以区分正常样本和离群点,从而降低离群点对聚类结果的影响. 目前,该方法主要用于模糊聚类算法中[8,12-15],由于技术等方面的原因,其在硬划分聚类算法(如 K-Means 算法)中的运用较为罕见.

本文将样本加权思想与新近提出的基于硬划分的子空间聚类算法相结合,得到一种鲁棒的子空间聚类算法. 通过在二维数据集、高维数据集以及基因数据集上的对比实验,证实了所提算法的有效性. 特别是对于严重噪声污染的数据集,该算法仍能保持满意的聚类结果,表现出较好的鲁棒性.

1 鲁棒子空间聚类算法

1.1 子空间聚类算法

许多聚类算法(如标准 K-Means 算法)均基于类内散度最小的优化准则. 给定样本空间中的一个有限数据集 $S=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$,$\boldsymbol{x}_i\in\mathbf{R}^d$,$K$-Means 算法的目标函数为

$$F(\boldsymbol{C})=\sum_{k=1}^{K}\frac{1}{|S_k|}\sum_{\boldsymbol{x}_i\in S_k}|\boldsymbol{c}_k-\boldsymbol{x}_i|^2 \tag{1}$$

式中:$\boldsymbol{C}$ 为聚类中心矩阵;K 为类别数;$\{S_1,\cdots,S_K\}$ 是对 S 划分后形成的 K 个类簇;$\boldsymbol{c}_k$ 为类簇 S_k 的中心.

考虑到在识别高维数据各类簇时不同子维的贡献程度存在差异,文献[6]中为每个类簇引入一个反映各维聚类贡献程度的权矢量,将 K-Means 算法扩展到对高维数据集聚类,提出了 LAM(locally adaptive metrics)子空间聚类算法,其目标函数为

$$F(\boldsymbol{W},\boldsymbol{C})=\sum_{k=1}^{K}\frac{1}{|S_k|}\sum_{\boldsymbol{x}_i\in S_k}\sum_{j=1}^{d}w_{kj}(c_{kj}-x_{ij})^2+h\sum_{k=1}^{K}\sum_{j=1}^{d}w_{kj}\lg w_{kj} \tag{2}$$

式中:w_{kj} 满足 $\sum_{j=1}^{d}w_{kj}=w$,$\forall 1\leqslant k\leqslant K$,$w$ 为常量,记 $\boldsymbol{w}_k=[w_{k1},\cdots,w_{kd}]$,$\boldsymbol{w}_k$ 对应第 k 个类簇的权矢量;目标函数右边的第 2 项为关于权矢量的负熵项,加入该项的目的是为了让更多子维参与聚类过程[5];参数 h 为大于 0 的某个常数. LAM 算法在实现对数据划分的同时,利用所得的权矢量对空间维度加权组合,形成各类簇所处的子空间.

1.2 离群点问题

离群点即噪声样本,属于远离正常样本分布范围的一类数据,是由数据收集过程中各种误操作或异常情况造成的[12]. 根据目标函数式(2),高维数据中任一样本 $\boldsymbol{x}_i$ 到第 k 类中心的距离可由 $\boldsymbol{w}_k$ 加权表示为 $\sum_{j=1}^{d}w_{kj}(c_{kj}-x_{ij})^2$. 不难看出,离群点到各类中心的距离通常比较远,在目标函数式中贡献较大,因此极有可能对聚类结果造成严重影响. 为此,下面为每个样本点分配一个尺度参数,以解决离群点对聚类结果的影响问题.

1.3 鲁棒子空间聚类算法

为每个样本点分配一个尺度参数 λ_i 后,目标函数式(2)可变为

$$F(\boldsymbol{C},\boldsymbol{W},\boldsymbol{\Lambda})=\sum_{k=1}^{K}\frac{1}{|S_k|}\cdot\sum_{\boldsymbol{x}_i\in S_k}\left[\frac{1}{\lambda_i}\sum_{j=1}^{d}w_{kj}(c_{kj}-x_{ij})^2\right]+h\sum_{k=1}^{K}\sum_{j=1}^{d}w_{kj}\lg w_{kj}$$
$$\text{s.t.}\quad\sum_{j=1}^{d}w_{kj}=w,\quad 0\leqslant w_{kj}\leqslant w,\quad 1\leqslant k\leqslant K \tag{3}$$

式中:λ_i 满足 $\sum_{i=1}^{n}\lambda_i=\lambda$ 或 $\sum_{k=1}^{K}\sum_{\boldsymbol{x}_i\in S_k}\lambda_i=\lambda$,$\lambda$ 为常量,记 $\boldsymbol{\Lambda}=[\lambda_1,\cdots,\lambda_n]$.

对于式(3)求极小值问题,本文构造如下优化函数

$$F_1(\boldsymbol{C},\boldsymbol{W},\boldsymbol{\Lambda})=\sum_{k=1}^{K}\frac{1}{|S_k|}\cdot\sum_{\boldsymbol{x}_i\in S_k}\left[\frac{1}{\lambda_i}\sum_{j=1}^{d}w_{kj}(c_{kj}-x_{ij})^2\right]+h\sum_{k=1}^{K}\sum_{j=1}^{d}w_{kj}\lg w_{kj}+$$
$$\delta\left(\sum_{k=1}^{K}\sum_{\boldsymbol{x}_i\in S_k}\lambda_i-\lambda\right)+\sum_{k=1}^{K}\beta_k\left(\sum_{j=1}^{d}w_{kj}-w\right) \tag{4}$$

式中:δ 和 $\beta_k(k=1,\cdots,K)$为拉格朗日乘子,它们将式(3)的约束条件引进式(4)中. 由此,通过令 $\frac{\partial F_1}{\partial\lambda_i}=0$ 以及 $\frac{\partial F_1}{\partial\delta}=0$,可得样本 $\boldsymbol{x}_i(i=1,\cdots,n)$的离群尺度值

$$\lambda_i=\frac{(X_{ki}/|S_k|)^{1/2}}{\sum_{m=1}^{K}\sum_{x_l\in S_m}(X_{ml}/|S_m|)^{1/2}}\lambda \tag{5}$$

式中：$X_{ki}=\sum_{j=1}^{d}w_{kj}(c_{kj}-x_{ij})^2$，且有 $\boldsymbol{x}_i\in S_k$，$\boldsymbol{x}_i=[x_{i1},\cdots,x_{id}]$. 式(5)为距离样本类心较近的样本分配一个较小的尺度值，而为距离所有类心较远的噪声点分配一个较大的尺度值.

与参数 $\boldsymbol{\Lambda}$ 的求解过程类似，参数 $\boldsymbol{W}=(w_{kj})_{K\times d}$ 可通过令 $\frac{\partial F_1}{\partial w_{kj}}=0$ 以及 $\frac{\partial F_1}{\partial \beta_k}=0$ 解得，类簇 $k(k=1,\cdots,K)$ 在第 $j(j=1,\cdots,d)$ 维的权值为

$$w_{kj}=\frac{\exp(-Y_{kj}/h)}{\sum_{t=1}^{d}\exp(-Y_{kt}/h)}w \tag{6}$$

式中：$Y_{kj}=\frac{1}{|S_k|}\sum_{\boldsymbol{x}_i\in S_k}\frac{1}{\lambda_i}(c_{kj}-x_{ij})^2$.

以上目标函数限制条件中涉及参数 λ 和 w 的选取，实验中选 $\lambda=n$，$w=d$，以保证 λ_i 和 w_{kj} 的均值都为 1.

给定 $\boldsymbol{W}$、$\boldsymbol{\Lambda}$ 和 $\boldsymbol{C}$，下面给出样本划分准则. 目标函数中与 $\{S_1,\cdots,S_K\}$ 相关的项为仅含 $\boldsymbol{W}$ 和 $\boldsymbol{\Lambda}$ 的类内散度项，距离测度可定义为

$$d^2(\boldsymbol{x}_i,\boldsymbol{c}_k,\boldsymbol{w}_k,\lambda_i)=\frac{1}{\lambda_i}\sum_{j=1}^{d}w_{kj}(c_{kj}-x_{ij})^2 \tag{7}$$

为了使目标函数值最小，需要将样本点分配到最邻近中心的类簇中，即

$$S_k=\{\boldsymbol{x}_i\mid(d^2(\boldsymbol{x}_i,\boldsymbol{c}_k,\boldsymbol{w}_k,\lambda_i)<d^2(\boldsymbol{x}_i,\boldsymbol{c}_t,\boldsymbol{w}_t,\lambda_i),\ \forall t\neq k\} \tag{8}$$

整理得

$$S_k=\Big\{\boldsymbol{x}_i\mid\Big(\sum_{j=1}^{d}w_{kj}(c_{kj}-x_{ij})^2\Big)^{1/2}<\Big(\sum_{j=1}^{d}w_{tj}(c_{tj}-x_{ij})^2\Big)^{1/2},\ \forall t\neq k\Big\} \tag{9}$$

最后，给定 $\boldsymbol{\Lambda}$ 和 $\{S_1,\cdots,S_K\}$，$\boldsymbol{C}$ 可通过 $\frac{\partial F}{\partial c_{kj}}=0$ 得到，c_{kj} 为

$$c_{kj}=\frac{\sum_{\boldsymbol{x}_i\in S_k}x_{ij}/\lambda_i}{\sum_{\boldsymbol{x}_i\in S_k}1/\lambda_i} \tag{10}$$

式中：$1\leqslant k\leqslant K$；$1\leqslant j\leqslant d$.

由式(5)、(6)、(9)和(10)，新的子空间聚类算法可总结如下.

输入为待聚类数据集 $S=\{x_1,\cdots,x_n\}$，$\boldsymbol{x}_i\in\mathbf{R}^d$，聚类数为 K，参数为 h.

(1)初始化，随机生成 K 个聚类中心，将维度权重矩阵 $\boldsymbol{W}$ 的各项设为 w/d，将样本权重向量 $\boldsymbol{\Lambda}$ 的各项设为 λ/n；

(2)利用 $\boldsymbol{W}$ 和 $\boldsymbol{C}$ 的值，根据式(9)对数据集 S 进行划分，得 $\{S_1,\cdots,S_K\}$；

(3)利用当前划分的 $\{S_1,\cdots,S_K\}$ 和 $\boldsymbol{\Lambda}$，根据式(10)计算聚类中心矩阵 $\boldsymbol{C}$；

(4)利用当前划分的 $\{S_1,\cdots,S_K\}$、$\boldsymbol{\Lambda}$ 和 $\boldsymbol{C}$，根据式(6)计算矩阵 $\boldsymbol{W}$；

(5)利用当前划分的 $\{S_1,\cdots,S_K\}$、$\boldsymbol{C}$ 和 $\boldsymbol{W}$，根据式(5)计算向量 $\boldsymbol{\Lambda}$.

重复步骤(2)～(5)，直至收敛.

整个迭代过程中，离群点被赋予较大的尺度参数 λ_i. 由式(6)、(10)可知，在计算各类簇中心和权矢量时，离群点的地位明显被减弱，同时式(9)表明，数据集的划分不受其影响，因此理论上保证了算法的鲁棒性.

下面证明算法的收敛性. 在上述利用拉格朗日乘子法获得算法各参数的过程中，通过令目标函数对各参数的一阶偏导等于 0，仅能得到相应参数的驻点，该值可为极值点或拐点. 为了进一步验证所得的驻点为极值点，还需检验目标函数对各参数的二阶偏导情况. 不难得到

$$\frac{\partial^2 F_1}{\partial\lambda_i^2}=\frac{1}{|S_k|}\frac{2}{\lambda_i^3}\sum_{j=1}^{d}w_{kj}(c_{kj}-x_{ij})^2>0 \tag{11}$$

$$\frac{\partial^2 F_1}{\partial w_{kj}^2}=\frac{h}{w_{kj}}>0 \tag{12}$$

因此，关于参数 λ_i 和 w_{kj} 的驻点正是其各自的极小值点，故有

$$F(\boldsymbol{C}^{t+1},\boldsymbol{W}^{t+1},\boldsymbol{\Lambda}^{t+1})\leqslant F(\boldsymbol{C}^{t},\boldsymbol{W}^{t},\boldsymbol{\Lambda}^{t})$$

$F(\boldsymbol{C},\boldsymbol{W},\boldsymbol{\Lambda})$ 是迭代次数 t 的递减函数，算法将最终收敛.

下面分析算法的运行效率. 算法的运行过程主要包括样本集划分、权矢量计算、聚类中心更新和尺度参数估计等 4 部分. 在样本集划分时，需要计算并比较样本点到 K 个不同聚类中心的加权距离，因此其计算复杂度为 $O(ndK)$. 权矢量的计算包括计算各维度参与识别不同类簇的贡献程度，需要用到各样本点到其中心的尺度加权距离，因此该过程的计算复杂度为 $O(ndK)$. 同理，中心更新和尺度参数估计的复杂度也均为 $O(ndK)$. 若算法需迭代 l 次收敛，则总的计算复杂度为 $O(lndK)$. 因此，算法的计算复杂度与数据维数、样本点数以及聚类数呈线性增长关系.

2　实验评估

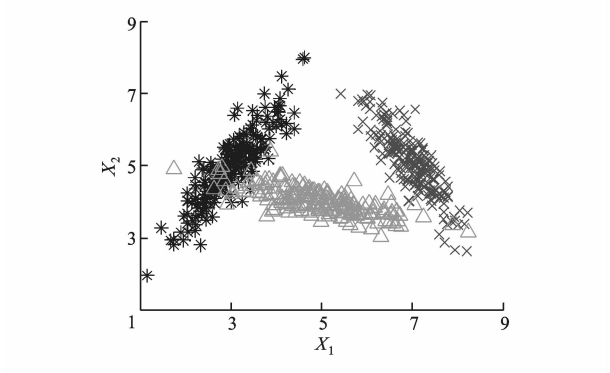
下面通过一系列数据实验的比较来说明本文算

法的性能. 首先测试二维数据集，包括无噪数据集 Data1，以及加噪数据集 Data1′和 Data1″[6]. 其次，测试高维仿真数据集 Data2[5]，以进一步研究本文算法的聚类性能和参数性质. 最后，在基因表达数据集 Lymphoma[16] 上比较不同算法的聚类效果.

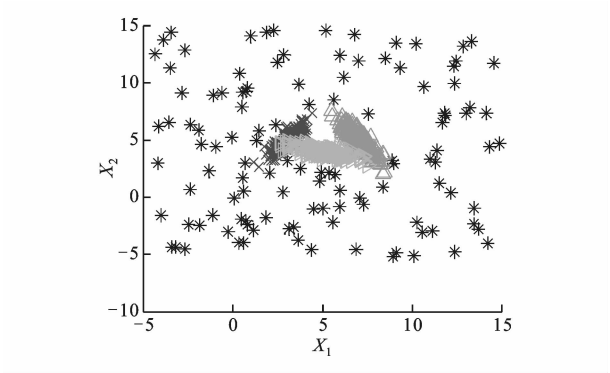
对于本文算法中的参数 h，实验中若无特别说明，统一取值为 100. 此外，由于参与比较的多数算法受初始点影响较大，故本文采用性能较优的 Kaufman 算法[17]提供初始化.

2.1 二维数据集实验

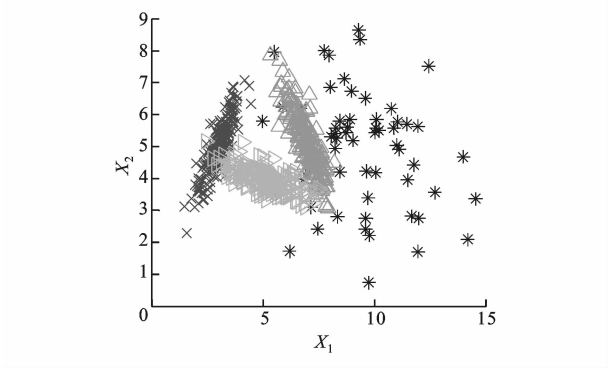
图 1 显示了二维数据集 Data1[6]、Data1′和 Data1″，其中 Data1′和 Data1″分别为在 Data1 中加入均匀噪声和高斯噪声后得到的数据集. 图 2～图 4 分别给出了不同算法对以上数据集的聚类结果.



(a)Data1



(b)Data1′

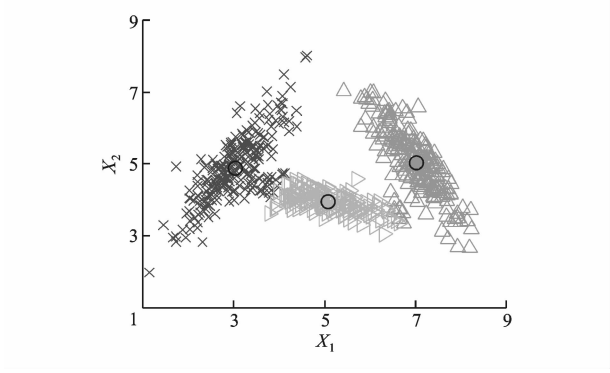


(c)Data1″

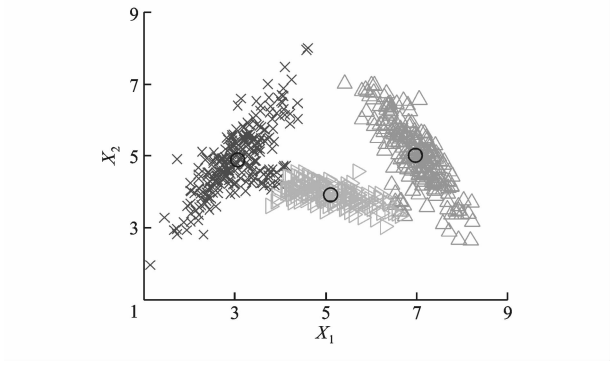
图 1 二维仿真数据集

由图 2 可见：本文算法较好地保持了 LAM 算法的聚类特点，能够区分各维对识别不同类簇的贡献程度；BK-spatialM 算法聚类效果不够理想，这主要是由于 BK-spatialM 算法基于欧式距离识别各类簇，并不区分各维的聚类贡献程度.

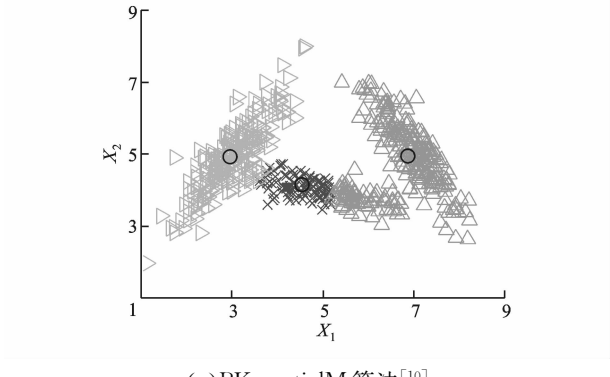
由图 3 和图 4 可见，对于 2 组噪声污染数据集，本文算法仍能够识别出各个类簇，其聚类效果明显好于其他 2 种算法. 此外，本文算法还获得了更准确的聚类中心.



(a)本文算法

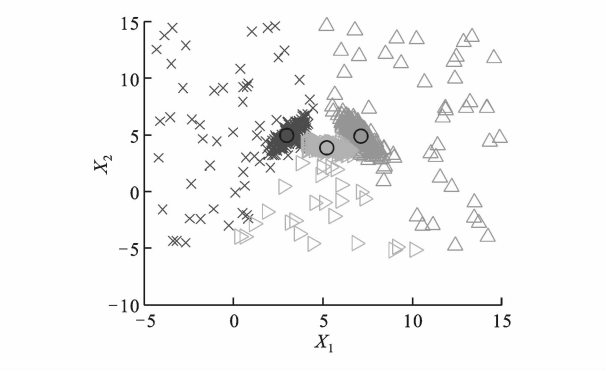


(b)LAM 算法[6]

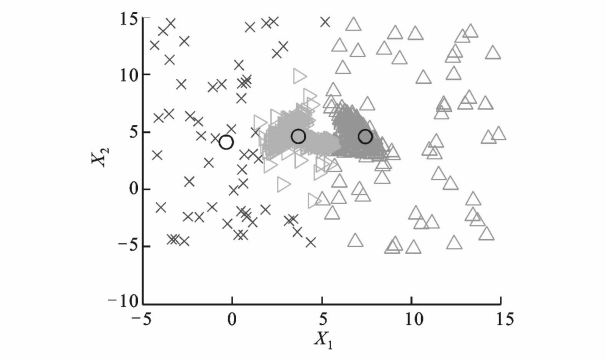


(c)BK-spatialM 算法[10]

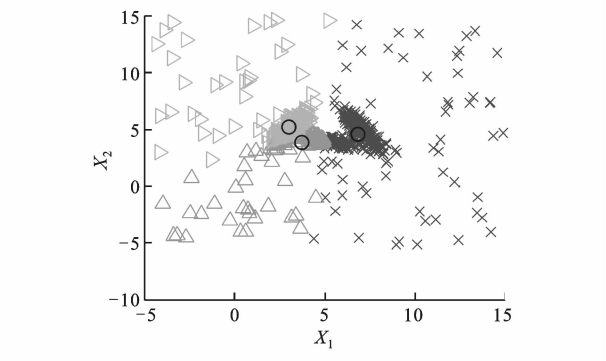
图 2 不同算法对 Data1 的聚类结果



(a)本文算法

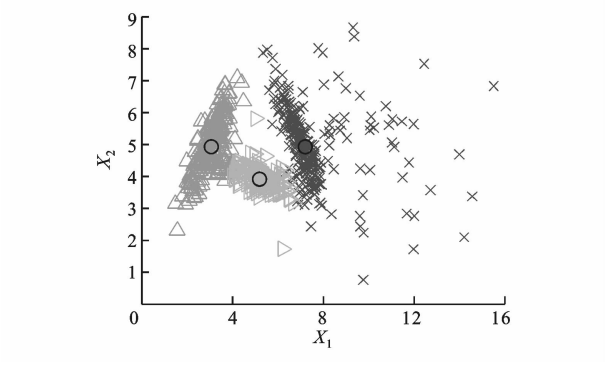


(b)LAM 算法[6]

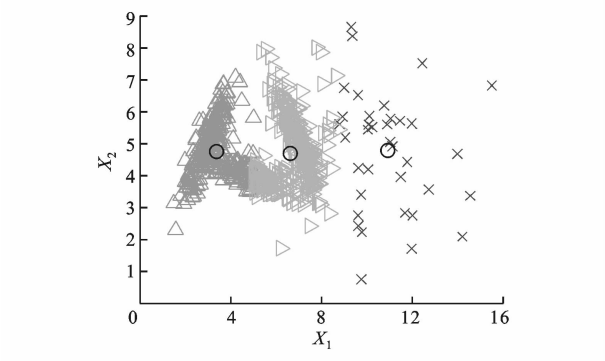


(c)BK-spatialM 算法[10]

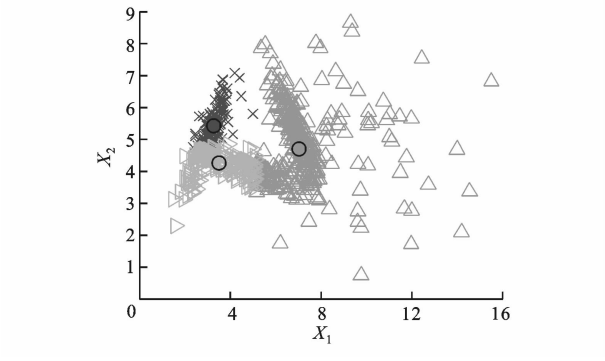
图 3　不同算法对 Data1′的聚类结果



(a)本文算法



(b)LAM 算法[6]



(c)BK-spatialM 算法[10]

图 4　不同算法对 Data1″的聚类结果

2.2　高维仿真数据实验

对于高维数据集 Data1，本文采用文献[5]中列出的数据生成方法获得. 该方法得到的数据集具备如下特点：①各成员类簇可具有不同形状和不同样本数；②属于某一类簇的样本，在该类的相关子维中服从高斯分布，而在该类的不相关子维中，大多数样本取值为 0，仅少数样本取随机正值；③不同类簇的相关子空间允许存在重叠. 实验生成 Data1 时，该数据生成方法中的主要参数设置如表 1 所示. 最后，还为 Data1 添加了不同比例的均匀分布噪声，噪声比例用 η 表示.

实验中，对于每种噪声比例，重复生成 100 组实验数据，并由下式计算不同算法的平均聚类精度

$$A = \sum_{k=1}^{K} D_k / n$$

式中：D_k 表示算法为第 k 类正确识别出的样本点数.

图 5 显示了不同算法的平均聚类精度随数据集中噪声比例变化的曲线. 可以看到，本文算法取得了比其他算法更好的聚类性能，尤其是在噪声污染程

度加重的情况下，其平均聚类精度仍很高，表现出较强的鲁棒性. BK-spatialM 算法被认为是对高维数据的鲁棒聚类算法，从图 5 可以看到，在噪声比例较小时，该算法能够保持稳定的聚类性能，但由于它不区分各维的聚类贡献程度，因此聚类性能总体上不如本文算法，而子空间聚类算法 LAC[4] 和 LAM 仅能够在数据集不含噪声时给出理想的聚类效果.

表 1 生成 Data1 的主要参数设置

参数	取值
类簇数 K	9
样本维数 D	16
正常样本数 n	375
平均每类的相关子维数 s	6
相关子维上的样本均值范围 μ	[0,20]
相关子维上的方差范围 σ	[0.1,0.2],[1,2],[5,6]
不相关子维上的样本取值范围 V	[0,5]
相关子维重叠率 ρ	0.2

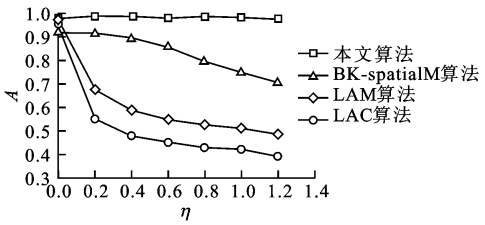
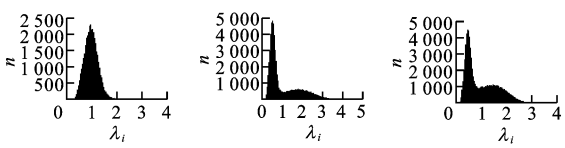


图 5 不同算法的聚类精度比较

图 6 显示了 3 种不同噪声比例(0、0.6 和 1.2)条件下，本文算法得到的样本尺度参数的直方图分布情况，各子图对应 100 次实验共 $100\times375\times(1+\eta)$ 个样本尺度值. 可以看到，在没有添加噪声的条件下，图 6a 中仅出现了 1 个波峰，而加入噪声后，图 6b 和图 6c 中各出现了 2 个波峰. 显然，第 2 个波峰及其周围对应着噪声点的尺度值分布. 对比图 6b 和图 6c 还可看出，第 2 个波峰下面积的变化恰恰对应着 2 个数据集中噪声比例的差异，因此尺度参数较好地定位了数据集中的噪声点.



(a)$\eta=0$ (b)$\eta=0.6$ (c)$\eta=1.2$

图 6 不同噪声比例条件下样本尺度参数值的分布情况

2.3 Lymphoma 数据集实验

基因表达数据集 Lymphoma[16] 含有 3 类 62 个样本，每个样本有 4 026 个基因表达值. 为了测试算法的鲁棒性，本文对该数据集添入不同比例的均匀噪声. 对于每种噪声比例 η，共随机生成 10 组含噪数据集. 表 2 给出了不同算法在 10 组含噪 Lymphoma 数据集上的平均聚类精度. 可以看到，随着数据集中噪声比例的增加，LAC 算法的聚类性能有所下降，LAM 算法的聚类性能在 $\eta=0.3$ 时有所提高，但其总体趋势为随着噪声比例增加而下降，而本文算法随噪声比例的增加仍保持着较高的识别性能，表现出较好的鲁棒性.

表 2 不同算法在含噪 Lymphoma 数据集上的聚类精度

η	A		
	LAC 算法	LAM 算法	本文算法
0.0	0.968	0.952	0.984
0.3	0.721	0.953	0.984
0.6	0.679	0.832	0.984
0.9	0.674	0.719	0.984

3 结 语

本文将样本加权函数法运用到基于硬划分的高维数据聚类中，得到了一种鲁棒的子空间聚类算法，实验结果证明了该算法的有效性. 样本加权函数法种类较多，如何将它们成功地应用于高维数据聚类中，并对所表现出的鲁棒性进行比较分析，将是一项非常有意义的课题.

参考文献：

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

[2] AGRAWAL R, GEHRKE J, GUNOPULOS D, et al. Automatic subspace clustering of high dimensional data for data mining applications [C] // Proceedings of ACM SIGMOD International Conference on Management of Data. New York, USA: ACM, 1998: 94-105.

[3] AGGARWAL C, PROCOPIUC C, WOLF J L, et al. Fast algorithms for projected clustering [C] // Proceedings of ACM SIGMOD International Conference on Management of Data. New York, USA: ACM, 1999:

61-72.

[4] DOMENICONI C, PAPADOPOULOS D, GUNOPULOS D, et al. Subspace clustering of high dimensional data [C]// Proceedings of SIAM International Conference on Data Mining. Philadelphia, PA, USA: SIAM, 2004: 517-521.

[5] JING Liping, NG M K, HUANG J Z. An entropy weighting *k*-means algorithm for subspace clustering of high-dimensional sparse data [J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(8): 1026-1041.

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

[7] DING C, HE Xiaofeng, ZHA Hongyuan, et al. Adaptive dimension reduction for clustering high dimensional data[C]// Proceedings of IEEE International Conference on Data Mining. Piscataway, NJ, USA: IEEE, 2002: 147-154.

[8] DAVE R N, KRISHNAPURAM R. Robust clustering models: a unified view [J]. IEEE Transactions on Fuzzy Systems, 1997, 5(2): 270-293.

[9] 穆向阳，张太镒，周亚同. 一种鲁棒的概率主成分分析方法[J]. 西安交通大学学报，2008，42(10)：1217-1220.
MU Xiangyang, ZHANG Taiyi, ZHOU Yatong. A robust probability principle component analysis method [J]. Journal of Xi′an Jiaotong University, 2008, 42(10): 1217-1220.

[10] DING Yuanyuan, DANG Xin, PENG Hanxiang, et al. Robust clustering in high dimensional data using statistical depths [J]. BMC Bioinformatics, 2007, 8(S7): S8.

[11] LAM B S Y, YAN Hong. Robust clustering algorithm for high dimensional data classification based on multiple supports [C]// IEEE International Joint Conference on Neural Networks. Piscataway, NJ, USA: IEEE, 2008: 1969-1976.

[12] 沈红斌，王士同，吴小俊. 离群模糊核聚类算法[J]. 软件学报，2004，15(7)：1021-1029.
SHEN Hongbin, WANG Shitong, WU Xiaojun. Fuzzy kernel clustering with outliers [J]. Journal of Software, 2004, 15(7): 1021-1029.

[13] 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.
GAO Jun, WANG Shitong. Fuzzy clustering algorithm with ranking features and identifying noise simultaneously [J]. Acta Automatica Sinica, 2009, 35(2): 145-153.

[14] HADJAHMADI A H, HOMAYOUNPOUR M M, AHADI S M. Robust weighted fuzzy C-means clustering [C]// IEEE International Conference on Fuzzy Systems. Piscataway, NJ, USA: IEEE, 2008: 305-311.

[15] KELLER A. Fuzzy clustering with outliers [C]// Proceedings of the North American Fuzzy Information Processing Society. Piscataway, NJ, USA: IEEE, 2000: 143-147.

[16] ALIZADEH A A, EISEN M B, DAVIS R E, et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling[J]. Nature, 2000, 403(6769): 503-511.

[17] KAUFMAN L, ROUSSEEUW P. Finding groups in data: an introduction to cluster analysis [M]. New York, USA: Wiley, 1990.

（编辑　葛赵青　武红江）