第 34 卷 第 5 期 计 算 机 工 程 2008 年 3 月
Vol.34 No.5 Computer Engineering March 2008
·博士论文· 文章编号：1000—3428(2008)05—0008—03 文献标识码：A 中图分类号：TP391.4

矢量聚类及其在稀疏分量分析中的应用

蔡荣太[1,2]，王延杰[1]
(1. 中国科学院长春光学精密机械与物理研究所图像室，长春 130033；2. 中国科学院研究生院，北京 100039)

摘 要：针对传统聚类分析不能有效处理矢量数据聚类的问题，提出矢量聚类算法。该算法以点到矢量的距离最小化为分类依据，所得类簇中心为一矢量。根据稀疏信号的分布特性，用矢量聚类方法估计系统的混合矩阵，再利用估计的混合矩阵分离混合信号，从而得到稀疏信源的估计，简化了传统的混合信号分离过程。实验结果表明该矢量聚类方法能比传统的标量聚类方法更有效地估计矢量数据的中心，能在稀疏的处理域中很好地分离出稀疏信源。

关键词：盲源分离；稀疏信号分析；矢量聚类

Vector Clustering and Its Application in Sparse Component Analysis

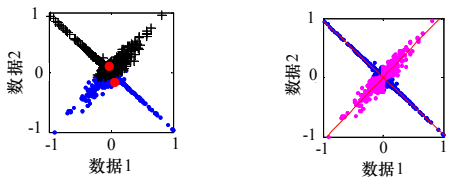
CAI Rong-tai[1,2], WANG Yan-jie[1]
(1. Image Lab, Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033;
2. Graduate School, Chinese Academy of Sciences, Beijing 100039)

【Abstract】 A vector clustering algorithm is proposed to cope with the inefficacy of traditional clustering algorithms to vector data. The algorithm classifies data into clusters by minimizeing the distance of a datum to a vector. The clustered centers are vectors. According to the distributing character of the sparse signal, a sparse signal separation algorithm is proposed which estimates the mixture matrix based on the vector clustering algorithm, and separates the source signal using the estimated mixture matrix. The algorithm is simple in computation comparing with traditional separation algorithms. Experimental results show that the algorithm is effective in vector data clustering and sparse signal separation.

【Key words】 Blind Source Separation(BSS); Sparse Component Analysis(SCA); vector clustering

1 概述

聚类是将物理或抽象的对象集合分成多个组的过程，聚类生成的组称为簇。聚类就是要让生成的簇内部的任意两个对象之间具有较高的相似度，不同簇的两个对象间具有较高的相异度。聚类分析是对数据建模，从而简化数据分析的一种方法，是多元统计分析的主要分支之一。聚类方法主要以距离和相似度为准则。常用的聚类方法有：基于划分的方法，基于层次的方法，基于密度的方法，基于网格的方法，基于模型的方法等[1]。传统的聚类分析都是针对标量数据的，不能解决矢量的聚类分析问题。如图 1 所示，图 1(a)为用传统的聚类方法对某一个数据集合聚类结果；图 1(b)为希望得到的聚类分析结果。图中的数据很明显是以 2 个矢量为中心分布的，图 1(a)传统的聚类结果显然是不合理的；图 1(b)将数据以两条直线为中心分为 2 类是比较合理的。针对该问题，文献[2]将数据转化到单位球上，将矢量聚类问题转化为标量聚类问题。当数据分布较为离散的时候，文献[2]的聚类算法效果并不理想。



(a)传统的聚类结果 (b)希望的聚类结果

图 1 数据散点图的聚类结果

2 矢量聚类分析

针对上述聚类算法的不足，提出一种以矢量数据为对象的矢量聚类算法。该算法将数据集合分成若干个簇，并且使得簇中内部的数据在矢量方向上具有较高的相似度，不同簇的数据之间在矢量方向上具有较高的相异度。该算法以空间数据点到矢量的聚类最小化为聚类依据，所得各簇的中心为一中心矢量，而非传统标量聚类方法的数据点。聚类结果将数据分为 n 簇，并求取各个簇的中心矢量。具体算法如下：

步骤 1 初始化：输入簇数 n 和 n 个初始的中心矢量。

步骤 2 距离计算：计算各个数据点到各个中心矢量的距离。

步骤 3 数据分类：遍历每个数据点，将其归类到与其距离最近的中心矢量所在的类别。

步骤 4 中心更新：遍历每一类，按距离最小的原则重新确定各类的矢量中心。

步骤 5 求总距离：重新计算各点到更新后的中心矢量的距离，记录总的距离量。

步骤 6 终止判断：计算前后 2 次的总距离之差，如果总距离之差小于一定的阈值，或者达到一定的迭代次数，则迭代终止。否则回到步骤 2。

基金项目：国家“863”计划基金资助项目(2006AA703405F)
作者简介：蔡荣太(1979－)，男，博士研究生，主研方向：信号分析与图像处理；王延杰，研究员、博士生导师
收稿日期：2007-03-30 **E-mail**：gjrtcai@163.com

在 R^n 空间中，点 p 到 n 维向量 $\boldsymbol{v}$ 的距离为由该点和空间原点构成的矢量 $\boldsymbol{v}_1(p,0)$和该矢量在 $\boldsymbol{v}$ 上的投影之差构成

$$d(p,\boldsymbol{v})=\boldsymbol{v}_1(p,0)-<\boldsymbol{v}_1(p,0),\boldsymbol{v}>$$

其中，$<\boldsymbol{v}_1(p,0),\boldsymbol{v}>$为 $\boldsymbol{v}_1(p,0)$在 $\boldsymbol{v}$ 上的投影。

在二维和三维空间中，点到矢量之间的距离就是通常的点到空间直线的距离和点到平面上直线的距离。后者为

$$d(p,l)=\frac{|ax_0+by_0+c|}{\sqrt{a^2+b^2}}$$

其中，点 p 为(x_0,y_0)；直线 l 为 $ax+by+c=0$。

3　基于矢量聚类的稀疏分量分析

3.1　稀疏分量分析

稀疏分量分析[3]是独立分量分析[4-5]的一个前沿研究领域。独立分量分析以独立性或非高斯性构造代价函数，通过一定的优化算法求解代价函数。其中非高斯性表现为超高斯或者表现为亚高斯。而稀疏性就是超高斯性。独立分量分析需要同时估计混合矩阵和原信号，算法复杂。作为独立分量分析的特例，稀疏分量分析可以将稀疏性作为分解的先验知识，而不必构造目标函数，从而简化分析分解过程。

设系统的观测模型为

$$\boldsymbol{A}s=x \tag{1}$$

其中，$\boldsymbol{A}$ 为 $(n\times m)$ 混合矩阵；s 为 $(m\times q)$ 的稀疏信源；x 为$(n\times q)$ 观测信号（混合信号）。

本文将稀疏分量分解分为 2 步：(1)根据稀疏性的先验知识，估计出混合矩阵 $\tilde{\boldsymbol{A}}$；(2)通过估计的混合矩阵 $\tilde{\boldsymbol{A}}$ 和观测信号 x 求取稀疏信源。

3.2　基于矢量聚类的混合矩阵估计

对某一时刻 t，由式(1)得

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}\cdot\begin{pmatrix} s_{1t} \\ s_{2t} \\ \vdots \\ s_{mt} \end{pmatrix}=\begin{pmatrix} x_{1t} \\ x_{2t} \\ \vdots \\ x_{nt} \end{pmatrix} \tag{2}$$

其中，a_{ij} 为 $\boldsymbol{A}$ 的第 ij 个元素；s_{ij} 为 s 的第 ij 个元素；x_{ij} 为 x 的第 ij 个元素。

由于 s 是稀疏的，将有较大的概率出现 $s_{it}>>s_{jt},j=1,2,\cdots,m,$ 且 $j\neq i$。此时，由式(2)得

$$\begin{pmatrix} x_{1t} \\ x_{2t} \\ \vdots \\ x_{nt} \end{pmatrix}=\frac{\begin{matrix} \sum_{k=1}^{m}a_{1k}s_{kt} \\ \sum_{k=1}^{m}a_{2k}s_{kt} \\ \vdots \\ \sum_{k=1}^{m}a_{nk}s_{kt} \end{matrix}}{}\approx\begin{pmatrix} a_{1i}s_{it} \\ a_{2i}s_{it} \\ \vdots \\ a_{ni}s_{it} \end{pmatrix}=\begin{pmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{ni} \end{pmatrix}=\boldsymbol{a}_i \tag{3}$$

其中，$\boldsymbol{a}_i$ 为 $\boldsymbol{A}$ 的第 i 个列矢量。也就是说当 s 为稀疏信源时，它的各变量在不同的时刻单独出现的概率较大，观测信号 x 的元素将以 $\boldsymbol{A}$ 中各列矢量为中心聚集分布。因此，通过矢量聚类算法求取聚类中心可以得到 $\boldsymbol{A}$ 矩阵的列矢量估计。由式(3)，当 i 取遍 1 到 m，可求取 $\boldsymbol{A}$ 的估计。

结合本文提出的矢量聚类算法，混合矩阵 $\boldsymbol{A}$ 的估计算法如下：

步骤 1　数据预处理：去除噪声点，并对观测信号 x 的每一行$(x_{i1},x_{i2},\cdots,x_{ip})$做归一化处理，减小噪声影响和尺度模糊。

步骤 2　做散点图：将预处理后的数据表示在 m 维空间中。

步骤 3　聚类分析：采用本文提出的矢量聚类算法对步骤 2 中的散点图进行聚类分析，求取各个聚类的中心矢量，该矢量就是 $\boldsymbol{A}$ 的某个列矢量的估计。

步骤 4　混合阵估计：由步骤 3 中的各个列矢量构成混合矩阵 $\boldsymbol{A}$ 的估计 $\tilde{\boldsymbol{A}}$。

在有些情况下，s 本身不够稀疏，但是通过一定得变换可以得到稀疏的表示[6]。这种情况下，在变换域中可以得到更好的混合矩阵估计。如对式(2)做傅里叶变换，根据傅里叶变换的关系可以得到和式(3)类似的结论。因此，对变换后的数据，可以用上述的算法估计出混合矩阵 $\boldsymbol{A}$。

3.3　信源的分离

设 $\boldsymbol{A}$ 非奇异，当 $n=m$ 时，由式(1)可得

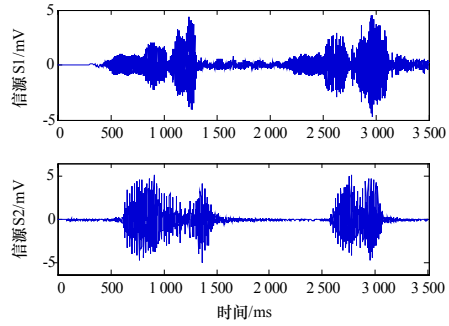
$$s=\boldsymbol{A}^{-1}x \tag{4}$$

将求得的混合矩阵的估计值 $\tilde{\boldsymbol{A}}$ 代入式(4)就可以估计出稀疏信源 s。当 $n>m$ 时，系统是超定的，可用最小二乘法求解；当 $n<m$，系统是欠定的，增加信源稀疏的约束条件，可化为线性规划问题求解。

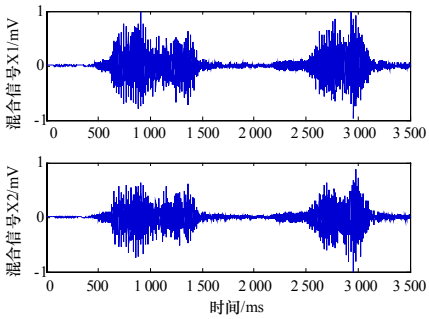
由于估计 $\boldsymbol{A}$ 时，不能保证各个列矢量和原来的次序一致，分离后信源的次序可能发生改变。同时，由于估计 $\boldsymbol{A}$ 时，只能保证各个列矢量内部分量比值为原混合矩阵的估计，不能保证各个列矢量之间的比例保持不变，因此分离后的估计信源在幅度上可能和原信源不一致。由于携带信息的是信号的波形，因此次序的不一致和幅度的变化是次要的。

4　实验结果与分析

图 2(a)是 2 个稀疏的语音信号源，图 2(b)为为混合后的 2 路观测信号。其中，混合矩阵 $\boldsymbol{A}=\begin{pmatrix} 0.623\,2 & 0.940\,4 \\ 0.799\,0 & -0.992\,1 \end{pmatrix}$。



(a)稀疏信源



(b)混合信号

图 2　原信号和混合信号

4.1　与基于标量聚类分析算法的比较

由 $\boldsymbol{A}$ 可得原混合矩阵的两个列矢量为

$$\boldsymbol{a}_1=\frac{0.780\,0}{1},\quad \boldsymbol{a}_2=-\frac{0.948\,4}{1}$$

本文实验采用的初始矢量为

$$\boldsymbol{a}_1^{\text{initial}}=-\frac{2}{1},\quad \boldsymbol{a}_2^{\text{initial}}=\frac{1}{1}$$

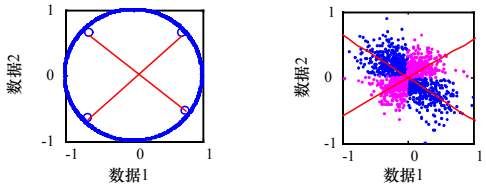
本文提出的矢量聚类方法估计得到的列矢量为

$$\boldsymbol{a}_1^{\text{vector}} = -\frac{0.581\,8}{1},\quad \boldsymbol{a}_2^{\text{vector}} = \frac{0.649\,0}{1}$$

文献[2]用归一化的标量聚类方法估计的列矢量为

$$\boldsymbol{a}_1^{\text{scalar}} = \frac{1.049\,4}{1},\quad \boldsymbol{a}_2^{\text{scalar}} = -\frac{1.180\,1}{1}$$

矢量聚类和标量聚类的比较见图 3。2 种聚类方法的结果对照见表 1。



(a)标量聚类结果　　(b)矢量聚类结果

图 3　矢量聚类和标量聚类比较

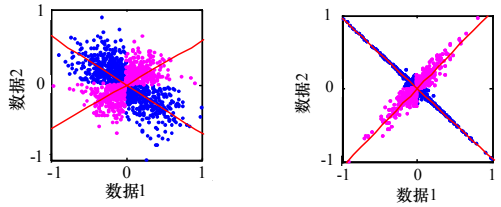
表 1　本文方法和文献[2]方法聚类效果比较

方法	$\boldsymbol{A}$ 的绝对估计误差均值	总距离	迭代次数
本文方法	0.248 8	165.301 5	7
文献[2]方法	0.250 4	474.424 0	9

实验结果表明本文提出的矢量聚类方法无论是在估计误差，内敛性(总距离)和迭代次数都优于文献[2]提出的标量聚类方法。

4.2　时域和频域的聚类实验比较

图 4(a)、图 4(b)分别为采用本文的聚类方法在时域和频域对混合信号的聚类结果。



(a)时域聚类结果　　(b)频域聚类结果

图 4　不同处理域中的聚类比较

从中可以看出，混合信号在时域中不够稀疏，但是变换后的频域中聚集分布在两个矢量方向上的。原混合矩阵的 2 个列矢量为

$$\boldsymbol{a}_1 = \frac{0.780\,0}{1},\quad \boldsymbol{a}_2 = -\frac{0.948\,4}{1}$$

初始矢量为

$$\boldsymbol{a}_1^{\text{initial}} = -\frac{2}{1},\quad \boldsymbol{a}_2^{\text{initial}} = \frac{1}{1}$$

本文提出的矢量聚类方法在时域中估计得到的列矢量为

$$\boldsymbol{a}_1^{\text{time}} = -\frac{0.581\,8}{1},\quad \boldsymbol{a}_2^{\text{time}} = \frac{0.649\,0}{1}$$

在频域中估计得到的列矢量为

$$\boldsymbol{a}_1^{\text{frequency}} = -\frac{0.956\,6}{1},\quad \boldsymbol{a}_2^{\text{frequency}} = \frac{1.062\,5}{1}$$

本文提出的矢量聚类算法在时域和频域中的聚类结果比较如表 2 所示。

表 2　时域和频域中的聚类分析结果比较

域	$\boldsymbol{A}$ 的绝对估计误差均值	总距离	迭代次数
时域	0.248 8	165.301 5	7
频域	0.145 4	57.908 7	3

实验结果表明本文提出的方法在稀疏性较好的频域中的聚类结果无论是在估计精度、内敛性和迭代次数方面都要优于稀疏性较差的时域中的聚类结果。

4.3　基于矢量聚类的稀疏信号分离实验

图 5 为采用本文提出的矢量聚类算法在频域估计的混合矩阵分离出来的原信号估计。可以看出，除了次序和幅度上的差别，估计的信号很好地逼近了原信号。可见提出的基于矢量聚类的稀疏信号分离效果是很好的。

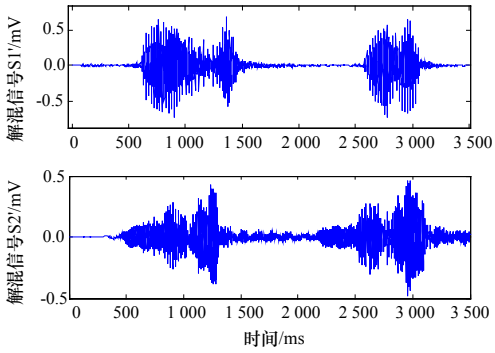


图 5　基于矢量聚类的稀疏信号分离结果

5　结束语

针对传统聚类方法不能有效处理矢量数据聚类的问题，提出一种新的矢量聚类算法，并将该算法成功应用于稀疏信号的分离中。

与传统的混合信号分离算法相比，提出的算法将混合矩阵的估计和信源的估计分为前后两步，算法简单，计算量小，可应用于实时性要求较高的信号处理系统中。

参考文献

[1] 刘远超，王晓龙，徐志明. 文档聚类综述[J]. 中文信息学报, 2006, 20(3): 55-62.

[2] Zibulevsk M. Blind Source Separation by Sparse Decomposition in a Signal Dictionary[J]. Neural Computation, 2001,13(4): 863-882.

[3] Goergiev P, Theis F, Cichocki A, et al. Sparse Component Analysis: A New Tool for Data Mining[C]//Proceedings of the Conference on Data Mining in Biomedicine. Gainesville, Florida, USA: [s. n.], 2004.

[4] Comon P. Independent Component Analysis, A New Concept?[J]. Signal Processing, 1994, 36(3): 287-314.

[5] 李加文，李从心. 基于ICA新算法的图像盲分离[J]. 计算机工程, 2006, 32(3): 186-190.

[6] Pearlmutter B A, Potluru V K. Sparse Separation: Principles and Tricks[C]//Proceedings of SPIE. San Jose, CA: [s. n.], 2003, 5102: 1-4.