基于 ICA 与 ViSOM 的不完整数据处理*)

彭红毅[1] 蒋春福[2] 朱思铭[3]
(华南农业大学理学院 广州 510642)[1] (深圳大学数学与计算科学学院 深圳 518060)[2]
(中山大学数学与计算科学学院 广州 510275)[3]

摘 要 本文介绍了数据挖掘中不完整数据的研究现状及 ICA 与 ViSOM 的特点,提出了基于 ICA 与 ViSOM 的不完整数据的处理模型 IVIS-IDH,研究了数据之间存在相关关系且为非高斯分布时不完整数据的处理方法,对缺失数据值的估计方法及其估计的残差进行了详细的讨论和分析,并在 ViSOM 基础上取得了不完整数据集的可视化分析结果,从而克服了 S. Wang 提出的不完整数据处理方法的不足。

关键词 不完整数据,ICA,ViSOM,相关关系,高斯分布

Handling of Incomplete Data Sets Based on ICA and ViSOM

PENG Hong-Yi[1] JIANG Chun-Fu[2] ZHU Si-Ming[3]
(College of Science, South China Agricultural University, Guangzhou 510642)[1]
(Department of Mathematics, Shenzhen University, Shenzhen 518060)[2]
(Department of Mathematics, Sun Yat-sen University, Guangzhou 510275)[3]

Abstract The paper introduces the state of incomplete data as well as ICA's and ViSOM's characteristics, studies the method of incomplete data sets under the circumstances of that data remain dependent and non-Gaussian, discusses the method estimation of missing data, and analyzes the carried-out the residual analysis. And then based on ICA and ViSOM, a model named IVIS-IDH, is proposed in this paper. The proposed model can achieve the visualization of incomplete data sets based on ViSOM, so that it overcomes the remedy for handing of incomplete data proposed by S. Wang.

Keywords Incomplete data, ICA (independent component analysis), ViSOM(Visualization-Induced Self-Organizing Maps), Correlation, Gaussian distribution

1 引言

目前数据挖掘对于不完整数据的处理,其中一个比较简便的做法是从数据集中删除这些不完整的数据记录,但是这样做会丢掉很多有用的信息[1]。S. Oba 等[2]研究了用 PCAs 方法对缺失数据进行处理,由于其前提是假设各数据指标服从高斯分布,但现实生活中很多数据并不服从高斯分布,因而此方法不具有通用性。彭红毅等[3]提出了 ICA-MDH 方法对缺失数据值进行估计,并取得了良好的效果,但其对怎样实现缺失数据值的估计缺乏详细的讨论,且未对缺失数据值估计方法的残差进行分析,同时缺失数据值估计出来后,也缺少进一步的处理过程。S. Wang[4]对不完整数据处理进行了相关研究,但其在用 SOM 方法取得数据的可视化结果前,假定各数据指标互相独立,并用特征平均值替换数据集中的缺失数据,但实际上很多变量之间都存在一定的相关性,因此这种方法不能利用不完整记录中已知数据的信息,并且这样直接利用 SOM 方法取得数据的可视化结果具有某种程度上的不合理性。

自组织映射(Self-organizing maps, SOM)是一种竞争的无指导学习方法,可以将任意的高维数据映射到一维或二维的网络图,为数据挖掘提供了非常有用的高维数据可视化技术[1,5]。Yin H. 在 SOM 的基础上提出了一种可视化诱导自组织映射(Visualization-Induced SOM, ViSOM),它使用与 SOM 同样的网络结构[6]。与 SOM 相比较,ViSOM 对群聚数据具有更好的高维数据可视化分类功能[6~8]。但遗憾的是,标准的 ViSOM 方法假定各数据指标互相独立,而实际上数据之间往往存在某种相关性,因此该法不具有通用性,并且标准的 ViSOM 方法不能处理不完整数据集。

ICA 是近几年才发展起来的一种新的多用途统计方法。该方法的目的是将观察到的数据进行某种线性分解,使其分解成线性独立的成分。Y. Rai[9]提出了一种简化的 ICA 方法,A. Kocsor 和 J. Csirik[10]对 FastICA 进行了相关研究和应用并给出了代码。F. Theis[11]提出了一种基于几何方法的 ICA 学习算法。Z. Shi, H. Tang, Y. Tang 等[12]也对 Fast ICA 方法进行了相关研究。ICA 方法的兴起为将具有相关性的数据指标转换为互相独立的数据指标提供了强有力的基础。

本文在文[3]的基础上对缺失数据值的估计进行了更详细的讨论与分析,在缺失数据值的估计后进行进一步的处理,弥补文[4]的不足。文章第 2 节介绍了 ViSOM 模型和算法;第 3 节在 ICA 和 ViSOM 的基础上提出了不完整数据处理模型,称之为 IVIS-IDH 模型,并将与该模型相对应的缺失数据估计方法称为 IVIS-IDH 方法;第 4 节介绍了实验结果;最后是结束语。

*)本文得到国家自然科学基金资助项目(10371135)。**彭红毅** 博士,研究方向:数据挖掘、人工智能。**蒋春福** 博士,研究方向:金融统计。**朱思铭** 教授,博士生导师,研究方向:人工智能与计算机网络,动力系统,混沌理论。

2 ViSOM 模型及算法

ViSOM 网络的工作原理是将任意维输入模式在输出层映射成一维或二维离散图形，并保持其拓扑结构不变。一维阵列 ViSOM 网络模型如图 1 所示。

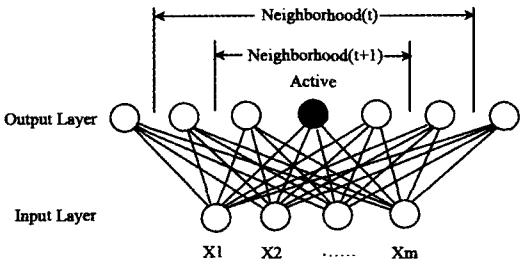

图 1　一维阵列 ViSOM 网络模型

定义输入向量 $X\in R^m$，节点索引为 $c(1\leqslant c\leqslant N)$，其中 N 表示输出节点数目，它的权值向量为 $w_c=[w_{c_1},w_{c_2},\cdots,w_{cm}]^T$。在时间步 t，输入数据为 $X(t)$，学习率为 $\eta(t)$。

ViSOM 网络算法可分两步进行：(1)对各观测数据指标进行标准化处理；(2)SOM 网络特征映射算法。前一步可以看成是对数据的预处理。

步骤(2)的具体算法如下：

Step1：初始化 ViSOM 训练过程。

Step1-1：置时间 $t=0$。

Step1-2：置 ViSOM 网络结构为 m 个输入节点，N 个输出节点。初始化权值向量 $w_k(0)=[w_{k1}(0),w_{k2}(0),\cdots,w_{km}(0)]^T(1\leqslant k\leqslant N)$，注意所有的权值向量应不同。

Step1-3：初始化邻域 $v(t=0)=q$，q 是一个任意的整数(例如本文中取 $q=N$)；

Step1-4：初始化学习率 $\eta(t=0)=p$，$0<p<1$；

Step2：训练 ViSOM。

Step2-1：在时间步 t，输入数据向量 $X(t)=[X_1(t),X_2(t),\cdots,X_m(t)]^T$。输入数据向量与权向量的匹配程度，用欧氏距离表示：$D_k=\|X(t)-w_k(t)\|$，其中 $w_k(t)$ 为 t 时刻的权值。

Step2-2：选择具有最小距离的节点 v 作为获胜节点，$d_v=\min\limits_k\{D_k\}$。

Step2-3：定义集合 $v_v(t)=\{k\mid\max[1,(v-v(t))]\leqslant k\leqslant\min[(v+v(t)),k]\}$。调节输出节点及其几何邻域内所连接的权值向量：如果 $k\in v_v(t)$，则

$$w_k(t+1)=w_k(t)+\eta(t)\times\left([X(t)-w_v(t)]+[w_v(t)-w_k(t)]\left(\frac{d_{vk}}{\Delta_{vk}\lambda}-1\right)\right);$$

否则 $w_j(t+1)=w_j(t)$。

Step2-4：如果学习率 $\eta(t)<\varepsilon$(其中 ε 为一个任意小的正数)，则转 Step2-1。否则 $\alpha(t)=\delta_1\alpha(t)$，$v(t)=\delta_2 v(t)$。其中，$0<\delta_1<1$，$0<\delta_2<1$，这样设置 δ_1,δ_2，以便当 $\eta(t)\xi\to\varepsilon$ 时，$v(t)\to1$。

置 $t=1+1$，转 Step2-1。

Step3：训练结束。

算法中，d_{vk} 是神经元 v 与神经元 k 的权值向量的距离；Δ_{vk} 是得胜神经元 v 与神经元 k 在输出空间的距离；λ 是一个正实数，一般经验取值为 $\lambda=1\sim1.5\times\frac{4\times\sqrt{Var_{\max}}}{N}$，其中 $Var_{\max}$ 表示数据的最大方差。

3 IVIS-IDH 模型

IVIS-IDH 模型如图 2 所示。

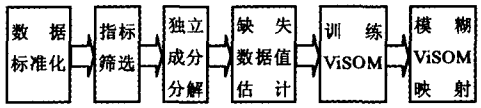

图 2　IVIS-IDH 模型

首先，我们需要从数据集中随机抽取适当的完整数据样本，对抽取的样本数据进行标准化处理并确定要进行分析的 m 个数据指标；然后，利用指标筛选算法从中筛选出 n 个线性无关的指标进行独立成分分解(详细的指标筛选算法参见 3.2 小节的 Case1)；接着进行独立成分分解，独立成分分解算法可参见文[3,10,12]；接着进行缺失数据值的估计(其详细过程参见 3.3 小节)；最后进行模糊 ViSOM 映射(其详细过程参见 3.5 小节)。3.4 小节为缺失数据值估计方法的残差分析。

3.1　指标之间线性关系的确定

假定 $X_i(i=1,\cdots,n)$，Y 为标准化的数据指标，$X_i,i=1,\cdots,n$ 之间线性无关，Y 与 X_i 的相关系数为 r_{Yi}，X_i 与 X_j 的相关系数为 r_{ij}，设 $Y=k_1X_1+k_2X_2+\cdots+k_nX_n$，下面讨论 $k_1,k_2,\cdots,k_n$ 的确定问题。

因为

$$\mathrm{cov}(Y,X_i)=k_1\,\mathrm{cov}(X_1,X_i)+k_2\,\mathrm{cov}(X_2,X_i)+\cdots+k_n\,\mathrm{cov}(X_n,X_i),$$

从而

$$r_{Yi}=k_1r_{1i}+k_2r_{2i}+\cdots+k_nr_{ni},i=1,\cdots,n。$$

写成矩阵的形式为：

$$R_{YX}=R_{XX}\cdot K$$

其中 $R_{YX}=[r_{Y1},\cdots,r_{Yn}]^T$，$R_{XX}=(r_{ij})_{n\times n}$，$K=[k_1,\cdots,k_n]^T$。

因为 $X_i,i=1,\cdots,n$ 之间线性无关，所以 R_{XX} 可逆，从而 $K=R_{XX}^{-1}\cdot R_{YX}$。

所以如果已知$(n+1)$个指标的相关系数矩阵，且其中有 n 个指标线性无关，另外有一个指标与这 n 个指标线性相关，我们就能确定另外一个指标与这 n 个指标对应的线性关系。

3.2　指标筛选

对于指标筛选，我们分两种不同的情况进行讨论：

Case 1：

设有 m 个指标，其相关协方差矩阵的秩为 n，则必可从中找到 n 个指标，这 n 个指标的相关系数矩阵的秩为 n。找的具体算法可参见文[13]。

Case2：

设某个观测对象中的指标 $X^{(c)}=[X_1,X_2,\cdots,X_c]^T$ 为缺失值，$X^{(m-c)}=[X_{c+1},X_{c+2},\cdots,X_m]^T$ 为已知值，$X=(X^{(c)},X^{(m-c)})$的相关系数矩阵的秩为 n，$Rank(X^{(m-c)})=k$，那么按照上面 Case 1 中介绍的方法可以从中找到 k 个指标，记为 $X^{(k)}$，使得 $Rank(X^{(k)})=k$，另外可以从 $X^{(c)}$ 中找到$(n-k)$个指标 $X^{(n-k)}$，使得 $Rank(X^{(k)},X^{(n-k)})=n$，设$(X^{(k)},X^{(c)})$的相关系数矩阵为 H，找的方法如下：

① 初始化 $d[j]=1,j=1,\cdots,k,d[j]=0,j=k+1,\cdots,k+c,tem=k+1$；

② 置 $d[tem]=1$，设 M 为从矩阵 H 中提取所有 $d[j]=$

$1, j=1,\cdots,k+c$ 的行与列的交叉元素组成的矩阵，如果 $\det(M)<\varepsilon_0$，则 $d[tem]=0$；其中 $\det(\cdot)$ 表示矩阵行列式值，ε_0 是一个很小的正数；

③ $tem=tem+1$，如果 $tem>k+c$，则转④，否则转②；

④ 如果 $d[j]=1, j=k+1,\cdots,k+c$，则 X_j 是我们要从 $X^{(c)}$ 找的线性无关的行向量，共有 $(n-k)$ 个，它们组成矩阵 $X^{(n-k)}$。

3.3 缺失数据值的估计

通过独立成分分析分解后，就可以由每个完整观测对象的数据值得出其对应的各独立成分的值。通过 SVM 的方法就可以近似估计每个独立成分的密度函数（详细的密度函数估计方法可参见文[1,2]），达到估计独立成分条件数学期望的目的，并最终达到估计缺失数据值的目的。

设 X^* 为原始数据，X 为标准化后的数据，$Rank(X)=n$，并设某个观测对象中的指标 $X^{(c)}=[X_1, X_2,\cdots, X_c]^T$ 为缺失值，$X^{(m-c)}=[X_{c+1}, X_{c+2}, X_m]^T$ 为已知值。

第一种情况：$m=n$。

设 $\tilde{X}$ 为去均值及白化处理后的数据指标，$S=[S_1, S_2,\cdots, S_n]^T$ 为独立成分分解后的指标，且有 $X=\tilde{B}\tilde{X}=\tilde{B}WS=\bar{B}S$，$\bar{B}=(\bar{b_{ij}})_{n\times n}$，按 3.2 小节中的指标筛选方法必能找到 c 个独立成分与已知的 $(m-c)$ 个指标线性无关，设这 c 个独立成分为 $S_i, i=1,\cdots,c$。

则根据密度函数的性质有

$$p(S_1,\cdots,S_c,X_{c+1},\cdots,X_n)=D\cdot p(S_1,\cdots,S_n)$$

其中 D 是坐标变换的雅可比行列式的绝对值。因此有

$$p(S_i\mid X_{c+1},\cdots,X_n)=\frac{\int\cdots\int p(S_1,\cdots,S_n)\mathrm{d}S_i\cdots\mathrm{d}S_{i-1}\mathrm{d}S_{i+1}\cdots\mathrm{d}S_c}{\int\cdots\int p(S_1,\cdots,S_n)\mathrm{d}S_1,\cdots\mathrm{d}S_c}$$

其中 $1\leqslant i\leqslant c$，$S_k(c<k\leqslant n)$ 按照 3.1 小节中介绍的方法可表示为 $S_1,\cdots,S_c,X_{c+1},\cdots,X_n$ 的线性函数。从而

$$\hat{S}_i=\int S_i p(S_i\mid X_{c+1},\cdots,X_m)\mathrm{d}S_i, 1\leqslant i\leqslant c$$

进一步可求出 $\hat{S}_i(c<i\leqslant n)$，因此可以得到标准化的缺失数据值的估计：

$$\hat{X}_i=\bar{b}_{i,1}\hat{S}_1+\bar{b}_{i,2}\hat{S}_2+\cdots+\bar{b}_{i,n}\hat{S}_n, i=1,\cdots,c。$$

再利用原始缺失数据指标的均值和方差就可以对原始缺失数据值进行估计。

第二种情况：$1\leqslant n<m$。

在这种情况下，缺失数据值估计的分析过程与文[3]类似，而指标之间线性关系的确定可仿照第一种情况由 3.1 小节与 3.2 小节中介绍的方法确定。

3.4 缺失数据值估计方法的残差分析

设 X 为标准化后的观测数据指标，X 各分量线性无关，并设某观测数据中 $X^{(2)}=[X_1, X_2,\cdots, X_c]^T$ 为缺失数据指标，$X^{(1)}$ 为已知数据指标，$X=\begin{bmatrix}X^{(1)}\\X^{(2)}\end{bmatrix}$，对 X 去均值及白化处理后的数据指标为 $\tilde{X}=\tilde{B}X$，其中 $\tilde{B}$ 为可逆正交矩阵，S 为对 $\tilde{X}$ 进行独立分解后的独立成分向量，且有 $S=B\tilde{X}=B\tilde{B}X=\bar{B}X$，我们进行如下分块：

$$S=[\bar{B}_1\ \bar{B}_2]\begin{bmatrix}X^{(1)}\\X^{(2)}\end{bmatrix}，\text{其中 } \bar{B}=[\bar{B}_1\ \bar{B}_2]。$$

于是

$$\begin{aligned}E(X^{(2)}\mid S)&=E\left(X^{(2)}\mid[\bar{B}_1\ \bar{B}_2]\begin{bmatrix}X^{(1)}\\X^{(2)}\end{bmatrix}\right)\\&=E(X^{(2)}\mid\bar{B}_1X^{(1)}+\bar{B}_2X^{(2)})\\&=E(X^{(2)}\mid\bar{B}_1X^{(1)})=E(X^{(2)}\mid X^{(1)})\end{aligned}$$

记 $M(X^{(1)})=E(X^{(2)}\mid X^{(1)})=[M_1(X^{(1)}), M_2(X^{(1)}),\cdots, M_c(X^{(1)})]^T$，根据数理统计的知识有

$$E[X_i-M_i(X^{(1)})]^2=\min_{L_i}E[X_i-L_i(X^{(1)})]^2, i=1,2,\cdots,c$$

其中 min 是对一切 x 的（可测）函数 $L_i(x)$ 取极小。

所以，3.3 小节中的缺失数据值的估计方法能使估计值与实际值的平均残差平方达到最小。

3.5 模糊 ViSOM 映射

缺失数据值估计后，我们可以得到一个模糊的数据集，在此基础上通过独立成分分析的方法就可得到一个相应的模糊的独立成分数据。假设我们已经训练了 ViSOM 网络，那么我们就可以产生一个模糊的 ViSOM 映射。其具体方法如下：

Step1：创建一个二维坐标系，横轴 J 有 N 节点（网络 ViSOM 的 N 个输出节点），纵轴 A 表示通过训练后的 ViSOM 网络，映射在节点上的样本数据个数度；

Step2：对每个观测对象 $X^{(i)}, i=1,\cdots,l$（其中 l 表示样本数目）的独立成分对象 $S^{(i)}$，执行下列子步骤：

Step2-1：将独立成分对象 $S^{(i)}$ 通过训练后的 ViSOM 网络映射到输出节点 $j(j=1,2,\cdots,N)$ 上；

Step2-2：将节点 j 定位在横轴 J 上。对于非缺失观测对象，在节点 j 的纵轴 A 方向上增加高度为 1 的黑条；对于缺失观测对象，在节点 j 的纵轴 A 方向增加高度为 1 的灰条。

4 实验结果

下面实验利用《北京统计年鉴》(2002,2000) 重点零售商业企业主要经济指标进行验证。这里选取含有六个经济指标共 290 条原始完整记录，这六个经济指标分别为：商品销售收入 $X1$，利税总额 $X2$，人均销售额 $X3$，人均创利税 $X4$，销售利税率 $X5$，存货周转率 $X6$。因为样本容量小于 2000，实验中先通过 Shapiro-Wilk W 检验，证实了 6 个数据指标都不服从高斯分布，然后从中抽取了 230 个完整记录进行独立成分分解，再从中分别随机抽取 60、70、80、100 条记录，使其中每条记录第二个指标 $X2$ 与第四个指标 $X4$ 成为空缺值，由每条记录中已知的 $X1$、$X3$、$X5$、$X6$ 来估计 $X2$ 与 $X4$。实验所用设备为一台 PC 机，所用系统环境为 Windows XP，运行工具为 SAS9.0 中文版软件。在同样条件下分别用平均值法、PCAs 法、本文提出的 IVIS-IDH 方法，并与原来的真实值作比较，这四次随机抽取每次实验结果都表明本文提出的 IVIS-IDH 方法的估计精度要明显优于平均值法与 PCAs 法。下面只列出随机抽取 100 条记录的实验结果比较。

IVIS-IDH、PCAs 及平均值估计方法的残差比较结果与文[3]相同，这里不再列出。图 3 为 IVIS-IDH 映射结果，图 3 中 type=1 表示有缺失数据的观测对象，type=2 表示没有缺失数据的观测对象。表 1 为每个类别的数字特征描述，表 2 为数据集的各指标的数字特征描述。

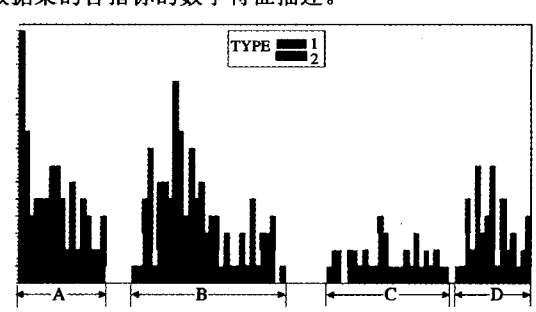

图 3 IVIS-IDH 映射结果

表 1 每个类别各指标的数字特征

类别 \ 指标		X1	X2	X3	X4	X5	X6
A	平均值	8860.8	286.5	15.2	0.48	3.24	1081.4
	标准差	6341.0	244.7	5.8	0.27	1.14	1755.1
B	平均值	14330.3	1303.8	34.8	3.16	10.37	4692.9
	标准差	17131.7	1840.9	29.0	3.98	15.77	7583.5
C	平均值	93725.8	8470.1	84.7	8.53	9.58	1385.7
	标准差	27765.1	4859.6	74.3	11.52	5.93	693.6
D	平均值	43067.5	1272.3	46.9	1.28	2.83	1581.9
	标准差	16232.2	964.6	20.8	0.96	1.51	1209.3

表 2 数据集各指标的数字特征

指标	X1	X2	X3	X4	X5	X6
平均值	27811.8	1925.4	37.3	2.73	6.83	2636.2
标准差	32594.7	3348.7	39.8	5.46	10.77	17440.8

从图 3 可以看出，数据集很明显地被分为四类，分别标记为类 A、B、C 与 D。四个类中，类 A、B 与 D 大部分由完整数据构成，而类 C 大部分由模糊数据构成。同表 1 中各指标的平均值相比较，表 1 中类 A 的各指标的平均值都很小；表 1 类 B 中指标 $X6$ 的平均值比表 2 中 $X6$ 的平均值大得多；表 1 中类 C 的指标 $X1$、$X2$、$X3$、$X4$ 的平均值比表 2 中相应指标的平均值大很多；表 1 中类 D 的指标 $X1$ 的平均值比表 2 中指标 $X1$ 的平均值大得多。通过表 1 与表 2 的结果比较，我们可以很容易地理解图 3 中的各个类别的主要特征。

结束语 ViSOM 是数据挖掘中一种很有用的高维数据可视化处理技术，然而标准的 ViSOM 方法不能处理不完整数据，而实际中很多数据集中每一条数据记录都是完整的情况很少见。文[6]在 SOM 的基础上为数据挖掘提供一种不完整数据处理方法，然而它不适用于数据指标之间存在相关性的情形。事实表明 ViSOM 是一种比 SOM 更有效的数据可视化方法[6~8]。本文在文[3]的基础上对缺失数据值的估计方法进行了更详细的讨论与分析，然后对该估计方法的残差进行了分析，并在缺失数据值的估计后，对数据集作了进一步的处理，提出了一种 IVIS-IDH 不完整数据处理模型，弥补了文[4]与文[3]的不足。

数据挖掘是信息时代发展很快的领域，最初的原始数据通常是不完整的。不完整数据处理技术是数据挖掘中不可缺少的部分。本文提出的 IVIS-IDH 方法为数据挖掘工作者提供了一种有效的数据挖掘技术，以至数据挖掘者能够充分利用有缺失数据的观测值来证实和加强只用完整数据才能得到的数据挖掘结果。

参 考 文 献

1 Kantardzic M. Data Mining Concepts, Models, Methods, and Algorithms. Tsing hua University Press, 2003
2 Oba S, et al. Missing Value Estimation Using Mixture of PCAs. LNCS, 2002, 2415:492～497
3 彭红毅，朱思铭，蒋春福. 数据挖掘中基于 ICA 的缺失数据值的估计. 计算机科学，2005，32(12)：203～205
4 Wang S. Application of self-organising maps for data mining with incomplete data sets. Neural Comput & Applic, 2003,12:42～48
5 Kohonen T. Self-organizing maps. 3rd ed. Berlin Heidelberg New York: Springer, 2001
6 Yin H. ViSOM——a novel method for multivariate data projection and structure visualization. IEEE TRANSACTION ON NEURAL NETWORKS, 2002, 1: 237～243
7 Yin H. Data visualization and manifold mapping using the ViSOM. Neural Networks, 2002, 15: 1005～1016
8 Sarvesvaran S, Yin H. Visualisation of Distributions and Clusters Using ViSOMs on Gene Expresssion Data. LNCS, 2004, 3177: 78～84
9 Rai Y. A simplified approach to independent component analysis. Neural Comput & Applic,2003(12):173～177
10 Kocsor A, Csirik J. Fast Independent Component Analysis in Kernel Feature Spaces. LNCS 2234,2001. 271～281
11 Theis F, et al, Overcomplete ICA with a Geometric Algorithm. LNCS 2415,2002. 1049～1054
12 Shi Z, Tang H, Tang Y. A fast fixed-point algorithm for complexity pursuit. Neurocomputing, 2005, 64 :529～536
13 彭红毅，蒋春福，朱思铭. 基于 ICA 与 SVM 的孤立点挖掘模型，计算机科学，2006(9)：175～177

(上接第 137 页)

图 6 所示是基于模拟的协同设计与制造工作流图。最上面的一层是工作流组合 GUI，以及工作流执行 Portal，第二层是基于网格的服务插件，将工作流中的任务翻译为服务需求，工作流仓库记录整个任务的操作信息，第三层基于语义的服务代理，根据翻译的服务需求，由 LSF，OpenPBS 等本地网格节点内部调度器进行作业调度。分配任务给高性能计算机或者使用高性能计算机上的模拟软件，完成模拟分析过程，并将结果反馈给工作流执行 portal。

结论 本文讨论基于模拟的虚拟产品开发涉及到建模和模拟技术，模拟主要是设计评估和验证，确保设计的性能，验证设计的缺陷。根据目前虚拟产品开发的特点，设计实现了网格环境下的基于模拟的协同设计与制造。

本文介绍了一种网格环境下的协同设计与制造技术，分析了松耦合的协同设计与制造，结合网格的最新研究进展，设计基于模拟的虚拟产品开发，更有效地利用各种有效的资源，缩短产品开发周期，实现网格环境下协同设计与制造。本文主要集中在基于模拟设计的协作网格环境的开发。

目前我们做到的网格环境下的协同设计与制造技术，是一种松耦合的协同设计与制造，任务的划分比较清晰，协作部门或者协作单位之间的任务，只是整个任务的一个环节。下一步，我们将研究紧耦合的协同设计与制造网格环境。

本文主要讨论了基于模拟的协同设计和制造的整个过程，结合网格的目前的发展趋势和最新的技术，充分利用现有资源，实现了基于网格的协同设计和制造的问题求解环境。

致谢 本文得到了国家自然科学基金委(90412010)的资助，得到了基金委项目组的成员的大力支持，在此表示感谢。

参 考 文 献

1 李国杰．关于超级计算与能力服务的战略思考. http://www.ncic.ac.cn/news/news_38.htm
2 Cutierrez A. e-business on demand: a developer roadmap. http://www.ibm.com/developerworks/grid/library/i-ebodov/index.html
3 Kesselman F C, Tuecke S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. International Journal of Supercomputer Applications, 2000,15(3)
4 Cockburn A. Goals and use cases. Journal of object-oriented programming, 1997. 35～40,Nov-Dec. 1997. 56～62
5 Tech Trend in PLM update. Collaborative Product Development Associates (www.cpd-associates.com), April 2005
6 Jeffrey E M. A BDI Agent Software Development Process, MS Thesis, (Advisor: Chang-Hyun Jo), University of North Dakota, USA, May 2002
7 Lewis C, Reiman J. Task-centered user interface design. Available via ftp.cs.colorado.edu/pub/cs/distribs/clewis/HCI-Design-Books.
8 SIMDAT. http://www.scai.fraunhofer.de/publications_simdat.0.html