# 数据挖掘中基于ICA的缺失数据值的估计*)

彭红毅　朱思铭　蒋春福

(中山大学数学与计算科学学院　广州 510275)

**摘　要**　本文简单介绍了数据挖掘中缺失数据的研究现状及ICA的特点与发展前景,提出了基于ICA的缺失数据估计模型——ICA-MDH模型。该模型研究了数据之间存在相关关系且为非高斯分布时缺失数据的处理方法,该方法能充分利用已知数据记录中的已知信息,且具有较好的通用性。实验通过对一些不完整经济数据进行了处理。结果表明,本文提出的缺失数据估计方法的精度明显优于平均值法和PCAs法,从而验证了本文所提模型的正确性与合理性。

**关键词**　缺失数据,ICA,相关关系,高斯分布

## Missing Value Estimation Based on ICA in Data Mining

PENG Hong-Yi　ZHU Si-Ming　JIANG Chun-Fu

(Department of Mathematics, Sun Yat-sen University, Guangzhou 510275)

**Abstract**　This paper introduces the study state of missing data as well as ICA's characteristics and foreground in brief, proposes a model, which is named as ICA-MDH model in this paper, based on ICA which is used abroad to dispose missing data under the circumstances of that data remain correlative and non-Gaussian distribution. This model takes full use of the known information of the given data to estimate missing data. By dealing with some economic data, the experiment verifies the results that ICA method is much better than PCAs and mean method, validating the correctness and reasonableness for the proposed model in this paper.

**Keywords**　Missing data, ICA(independent component analysis), Correlation, Gaussian distribution

## 1　前言

数据挖掘对于不完整数据中缺失数据的处理目前常用的方法主要有3种[1]:一是用一个全局量替换所有缺失值;二是用特征平均值替换缺失值;三是用属于给定种类的特征的平均值替换缺失值(此方法仅可用于样本预先分类的分类问题)。用特征平均值替换缺失值,实际上相当于假设各变量属性是互相独立的。但很多变量之间都存在某种程度上的相关性,因此这种方法并不能利用不完整数据记录中已知数据的信息。Shigeyuki Oba[8]等采用PCAs方法对缺失数据进行处理,这种方法考虑了数据之间的相关性,其前提是假设各数据指标服从高斯分布。但现实中有不少数据并不服从高斯分布,因而此方法并不具有通用性。Ad Feelders[2]研究了树结构中缺失数据处理方法,Jerzy W. Grzymala-Busse[3]研究了用粗略集方法对不完整数据进行处理,Bobby D. Gerardo[4]等讨论了在分布数据库中缺失数据的处理,Dan Li[5]、Zs. J. Viharos[6]、Rafal Latkowski[7]也对缺失数据进行了相关研究,但这些方法都未能充分考虑数据间的相关性。

独立成分分析(ICA)是近几年才发展起来的一种新的统计方法。它的目的是:为非高斯分布数据找到一种线性变换,这样成分与成分之间是统计独立或尽可能独立。独立成分分析方法是一种多用途的统计方法,在语音信号分离、生理学数据分析、金融数据分析、图像消噪、远程通讯、人脸识别等方面的应用成果充分显示了ICA的特点及非常重要的价值。ICA最早由Jutten C和Herault J[9]提出,后来Yogesh Singh·C. S. Rai[10]提出了一种简化的ICA方法,Andras Kocsor and Janos Csirik[11]对FastICA进行了相关研究和应用,Fabian J. Theis[12]提出一种基于几何方法的ICA学习算法。ICA方法的兴起为充分利用已知数据集中的信息进行缺失数据估计提供了可能。

本文主要研究非高斯分布数据之间存在相关关系时缺失数据的估计方法。文章第2节介绍了ICA模型及Fast ICA算法;第3节在ICA的基础上提出了缺失数据处理模型,称之为ICA-MDH模型,并将与该模型相对应的缺失数据估计方法称为ICA-MDH方法;第4节介绍了实验结果;最后对全文做了小结。

## 2　ICA模型及Fast ICA算法

独立成分分析或者ICA模型的一般形式为:

$$X=AF \tag{1}$$

这里 $X=[X_1,X_2,\cdots,X_p]^T$ 为在输出层能观察到的向量,$A$ 为未知 $p\times m$ 矩阵,$F=[F_1,F_2,\cdots,F_m]^T$ 是独立成分向量。这个模型表示被观察到的数据是如何由独立成分混合而产生的。

为保证解的唯一性,一般要求:

①所有独立成分中至多只有一个是高斯的。

②观测到的线性混合数据数目 $p$ 必须不小于独立成分的个数 $m$,即 $1\leqslant m\leqslant p$。

*)本文得到国家自然科学基金资助(10371135)。彭红毅　博士生,研究方向:人工智能、数据挖掘;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;蒋春福　博士生,研究方向:金融统计。

③混合矩阵 $A$ 必须是列满秩的。

当前，ICA 算法有非高斯的最大化、互信息的最小化算法、最大似然估计(ML)等算法。ICA 的主要应用是特征提取、盲源信号分离、生理学数据分析、语音信号处理、图像处理及人脸识别等。本文介绍一种 Fast ICA 算法对不完整数据的处理。

Fast ICA 算法分三步实现：(1)对观测数据去均值；(2)对去均值后的观测数据白化处理；(3)独立分量提取算法及实现。前两步可以看成是对观测数据的预处理，通过去均值和白化可以简化 ICA 算法。

要提取独立分量，完成步骤(3)，需要使用如下算法：

①初始化 $g(0)$，令其模为 1，置 $k=1$；

②$g(k)=C^{-1}E\{X(g(k-1)^{\mathrm{T}}X)^3\}-3g(k-1)$，其中期望值可以由大量向量的采样点计算出来，$C$ 为数据 $X$ 的协方差；

③用 $\| g(k) \|$ 去除 $g(k)$；

④如果 $|g(k)^{\mathrm{T}}g(k-1)|$ 不是足够接近 1，那么令 $k=k+1$，返回第②步，否则输出 $g(k)$。

算法最后给出的向量 $g(k)$ 等于正交混合矩阵中的一列，在独立分离中意味着分离了其中一个非高斯独立成分。

为了估计 $m$ 个独立成分，必须运行上面算法 $m$ 次。为了保证每次估计的都是不同的独立成分，需要增加一个正交化投影操作。混合矩阵 $C$ 的列是正交的，这样就能对独立成分一个一个地进行估计，通过投影当前的 $g(k)$ 解到混合矩阵 $G$ 的列上。定义矩阵 $G$ 的列是目前已找到的混合矩阵 $G$ 的列，增加投影操作到第(3)步开始：

$g(k)=g(k)-GG^{\mathrm{T}}g(k)$，用 $\| g(k) \|$ 去除 $g(k)$。

初始的随机向量在开始递推前也执行这个投影操作。为了避免 $g(k)$ 估计的恶化，投影操作可以在迭代一些次数后取消。

## 3 ICA-MDH 模型

由已知的一系列独立同分布数据 $x_1,\cdots,x_l\cdots$，可以构造一个经验分布函数：

$$F_l(x)=\frac{1}{l}\sum_{i=1}^{l}\theta(x-x_i)$$

其中 $\theta(u)$ 是阶越函数，当 $u\geqslant 0$ 时取值为 1，而其他情况下取值为 0。由经验分布函数可近似估计出其密度函数 $p(x)$。

每个独立成分的密度函数都可以采用这种经验分布函数方法来估计。设 $F_1,F_2,\cdots,F_m$ 互相独立，其密度函数分别为 $p(F_1),p(F_2),\cdots,p(F_m)$，则其联合分布密度为 $p(F_1,F_2,\cdots,F_m)=\prod_{i=1}^{m}p(F_i)$。

假设通过独立成分分析的方法能得出方程(1)，那么，我们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。通过经验分布函数近似估计每个独立成分的密度函数，最后达到估计独立成分条件数学期望的目的。

设 $X^*$ 为原始数据，$X^{**}$ 为标准化后的数据，$X$ 为去均值及白化处理后的数据。

$$X^{**}=BX \tag{2}$$

假设 $\mathrm{Rank}(A)=m$，并设某个观测对象中的指标 $X^{(c)}=[X_1,X_2,\cdots,X_c]^{\mathrm{T}}$ 为缺失值，$X^{(p-c)}=[X_{c+1},X_{c+2},\cdots,X_p]^{\mathrm{T}}$ 为已知值，为了说明怎样估计缺失指标 $X^{(c)}$，我们分情况讨论。

第一种情况：如果 $p=m$。

设 $A$ 的逆矩阵为 $G=(g_{ij})_{p\times p}$，从中另取 $c$ 个独立成分，设为 $F_{i_j}$，$j=1,\cdots,c$。

则根据密度函数的性质有：

$$p(F_{i_1},\cdots,F_{i_c},X_{c+1},\cdots,X_p)=D\cdot p(F_{i_1},\cdots,F_{i_p})$$

其中 $D$ 是坐标变换的雅可比行列式的绝对值。因此有

$$p(F_{i_j}\mid X_c,\cdots,X_p)=\frac{\int\cdots\int p(F_{i_1},\cdots,F_{i_p})\mathrm{d}F_{i_1}\cdots\mathrm{d}F_{i_{j-1}}\mathrm{d}F_{i_{j+1}}\cdots\mathrm{d}F_{i_p}}{\int\cdots\int p(F_{i_1},\cdots,F_{i_p})\mathrm{d}F_{i_1}\cdots\mathrm{d}F_{i_p}}$$

其中 $1\leqslant j\leqslant c$，$F_{i_k}$ $(c<k\leqslant p)$ 可表示为 $F_{i_1},\cdots F_{i_c},X_c,\cdots,X_p$ 的函数。从而

$$\hat{F}_{i_j}=\int F_{i_j}p(F_{i_{jk}}\mid X_c,\cdots,X_p)\mathrm{d}F_k,1\leqslant j\leqslant c$$

进一步可求出 $\hat{F}_{i_j}$ $(c<j\leqslant p)$，因此

$$\hat{X}_i=a_{i,i_1}\hat{F}_{i_1}+a_{i,i_2}\hat{F}_{i_2}+\cdots+a_{i,i_p}\hat{F}_{i_p},i=1,\cdots,c$$

那么，由方程(2)就可估计出标准化后的缺失数据，再与原始缺失数据指标的均值、方差一起就可估计出原始的缺失数据值。

第二种情况：如果 $1\leqslant m<p$，则有

$$X^{(c)}=A^{(c)}\cdot F,X^{(p-c)}=A^{(p-c)}\cdot F$$

其中 $A^{(c)}=(a_{ij})_{p\times m},i=1,\cdots,c,j=1,\cdots,m$

$A^{(p-c)}=(a_{ij})_{(p-c)\times m},i=c+1,\cdots,p,j=1,\cdots,m$

设 $\mathrm{Rank}(A^{(p-c)})=k$，$1\leqslant k\leqslant\min(p-c,m)$，记 $\mathcal{R}(X)$ 表示矩阵 $X$ 的行向量生成的线性空间。

(1)如果 $k=m$，那么必有 $\mathcal{R}(X^{(c)})\subseteq\mathcal{R}(X^{(p-c)})$。

(2)如果 $k<m$，那么向量 $X^{(p-c)}$ 中必有 $k$ 个行向量线性无关，设为 $X^{(k)}=[X_{c+1},X_{c+2},\cdots,X_{c+k}]^{\mathrm{T}}$。因为 $\mathrm{Rank}(A)=m$，所以必能从 $X^{(c)}$ 中找到 $m-k$ 个行向量，设为 $X^{(m-k)}=[X_1,X_2,\cdots,X_{m-k}]^{\mathrm{T}}$，记 $X^{(m)}=\begin{bmatrix}X^{(k)}\\X^{(m-k)}\end{bmatrix}$，则有 $\mathcal{R}(X^{(m)})=\mathcal{R}(X)$，因此

$$\mathcal{R}(X^{(c-k)})\subseteq\mathcal{R}(X^{(m)})$$

其中 $X^{(c-k)}=[X_{m-k+1},\cdots,X_c]^{\mathrm{T}}$。记 $A^{(m-k)}=(a_{ij})_{(m-k)\times m},i=1,\cdots,m-k,j=1,\cdots,m,A^{(k)}=(a_{ij})_{k\times m},i=c+1,\cdots,c+k,j=1,\cdots,m$，则有

$$\begin{bmatrix}X^{(m-k)}\\X^{(k)}\end{bmatrix}=\begin{bmatrix}A^{(m-k)}\\A^{(k)}\end{bmatrix}F$$

接下来缺失值 $X^{(m-k)}$ 的估计方法，可仿照第一种情况，从而也可估计出缺失值 $X^{(c-k)}$。

## 4 实验结果

本文采用 ICA-MDH 法进行缺失数据估计，实验中所用设备为一台 PIV1.7G 256M 的 PC 机，系统环境为 Windows XP，运行工具为 SAS9.0 中文版软件。

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数据验证本文提出的模型。在这里共选取 290 条原始完整记录、6 个原始经济指标，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，证实数据不服从高斯分布，然后从中抽取 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。在同样条件下分别用平均值

法、PCAs 法、ICA-MDH 法进行缺失数据的估计，并与原来的真实值做比较，这 4 次随机抽取中每次实验结果都表明 ICA-MDH 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。

表 1　随机抽取 100 条记录的结果比较

| 指标 | 所用方法 | 平均残差平方 |
|---|---|---|
| X2 | 平均值法 | 13789596.57 |
| | PCAs 法 | 5536853.42 |
| | ICA-MDH 法 | 9440.15 |
| X4 | 平均值法 | 37.3312510 |
| | PCAs 法 | 7.7899376 |
| | ICA-MDH 法 | 0.0016714 |

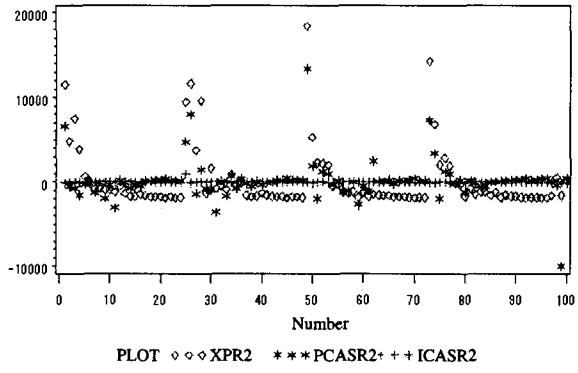


图 1　随机抽取 100 条记录指标 X2 的残差散点图

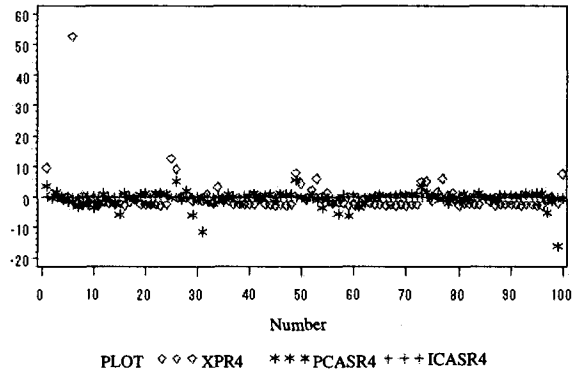


图 2　随机抽取 100 条记录指标 X4 的残差散点图

表 1 为用各种方法对数据指标 X2 和 X4 作出的平均残差平方结果比较，图 1、图 2 为实际值与用各种方法估计的相关指标值的残差散点图，图中 Number 表示缺失数据的序号，XPR2、PCASR2、ICASR2 分别表示用平均值法、PCAs 法、ICA-MDH 法对实际利税总额作出的估计值的残差。XPR4、PCASR4、ICASR4 分别表示用平均值法、PCAs 法、ICA-MDH 法对实际人均创利税作出的估计值的残差。

**小结**　由以上实验结果可以看出，当数据之间存在相关性且为非高斯分布时，本文提出的缺失数据估计方法 ICA-MDH 法要明显优于平均值法和 PCAs 方法。实验结果说明，ICA-MDH 方法能较好地进行不完整数据的处理，并具有较好的通用性。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，而处理不完整数据记录中缺失数据的技术是数据挖掘中不可缺少的部分。本文提出的基于 ICA 缺失数据估计方法——ICA-MDH 法提供了一种有效的数据挖掘技术，这样数据挖掘工作者能够充分利用缺失数据的观测值得到和加强只用完整数据集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. Beijing: Tsing hua University Press, 2003
2 Feelders A D. Handling Missing Data in Trees: Surrogate Splits or Statistical Imputation. LNAI 1704, 1999. 329～334
3 Grzymala-Busse J W. Rough Set Approach to Incomplete Data. In: LNAI 3070, 2004. 50～55
4 Gerardo B D, et al. The Association Rule Algorithm with Missing Data in Data Mining. In: LNCS3043, 2004. 97～105
5 Li Dan, et al. Towards Missing Data Imputation- A Study of Fuzzy K-means Clustering Method. In: LNAI 3066, 2004, 573～579
6 Viharos Z J, et al. Training and Application of Artificial Neural Networks with Incomplete Data. In: LNAI 2358, 2002. 649～659
7 Latkowski R. Incomplete Data Decomposition for Classification. In: LNAI 2475, 2002. 413～420
8 Shigeyuki O, et al. Missing Value Estimation Using Mixture of PCAs. LNCS 2415, 2002. 492～497
9 Jutten C, Herault J. Independent component analysis versus PCA. In: Proceeding of European Signal Processing Conf, 1988. 287～314
10 Yogesh Singh. Rai C S. A simplified approach to independent component analysis. Neural Comput & Applic, 2003, 12: 173～177
11 Kocsor A, Csirik J. Fast Independent Component Analysis in Kernel Feature Spaces. In: LNCS 2234, 2001. 271～281
12 Theis F J, et al. Overcomplete ICA with a Geometric Algorithm, In: LNCS 2415, 2002. 1049～1054


(上接第 163 页)

算效率。性能测试结果表明，构建和存储块排序索引结构所需要的代价比后缀树要小得多，块排序索引是一种非常适合生物序列计算的数据结构。本文主要针对 DNA 序列局部比对查询处理。在以后的工作中，我们将利用块排序技术解决多序列比对等复杂的查询问题。

## 参 考 文 献


1 The Human Genome Project (HGP). http://www.nhgri.nih.gov/HGP/.
2 Smith T, Waterman. Identification of common molecular subsequences. Journal of Molecular Biology, 1981, 147: 195～197
3 Meek C, Jignesh M P, Shruti K. OASIS: An Online and Accurate Technique for Local-alignment Searches on Biological Sequences. In: VLDB, 2003
4 Kahveci T, Singh A K. An Efficient Index Structure for String Databases. In: VLDB, 2001. 351～360
5 Pearson, Lipman. Improved tools for biological sequence comparison. In: Proc. NatlAcad Sci, 1988. 2444～2488
6 Alschul S, gish W, Miller W, et al. Basic local alignment search tool. Journal of Molecular Biology, 1990, 215(3): 403～410
7 Manzini G. The Burrow-Wheeler Transform: Theory and Practice. Lecture Notes in Computer Science, 1999, 1672: 34～47
8 Kelly K, Labute P. The A* search and Applications to Sequence Alignment