一种改进的高维数据可视化模型*)

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 可视化诱导自组织映射(ViSOM)是一种人工神经网络模型,已经被成功应用于高维数据的可视化分析。但是,标准的ViSOM方法不仅没有考虑数据之间的相关性,而且当输出网络结点太多时,需要消耗大量运算开销;输出网络结点太少,又难以分析数据的可视化结果。为克服ViSOM的这两个弱点,本文首先在ViSOM的基础上提出了一个改进的映射算法MViSOM,接着在独立成分分析(ICA)与MViSOM的基础上提出了一个改进的高维数据可视化模型IMViSOM。论文最后通过实验说明了IMViSOM模型在对群聚数据的可视化分类效果及运算速度方面都优于ViSOM方法,从而验证了IMViSOM模型的正确性与合理性。

关键词 独立成分分析,可视化诱导自组织映射,相关性

A Modified Visualization-Model of High-dimensional Data

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 Visualization-Induced Self-Organizing Maps (ViSOM), as one of the artificial neural networks models, has been successfully applied in the analysis of visualization of high-dimensional data. However, it has two weaknesses. Firstly, it does not consider the correlation of data. Secondly, much memory will be used up if the output nodes are too large, and contrarily, the visibility results of data will be difficult to be analyzed if the output nodes are too small. In order to overcome the above two weaknesses of ViSOM, a modified algorithm named MViSOM, based on ViSOM, as well as a visualization-model of high-dimensional data, based on ICA (Independent Component Analysis) and MViSOM, are proposed in this paper. Finally, the experiments also show that IMViSOM method has advantages over ViSOM because of its excellent classified effect of swarm data and high calculating speed, confirming the correctness and reasonableness for the proposed model in this paper.

Keywords Independent component analysis, Visualization-induced self-organizing maps, Correlation

1 引言

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[1,3]。H. Yin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[2]。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[2~5]。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。另外,当ViSOM网络输出结点太多,则会消耗大量内存开销;输出网络结点太少,则难以分析数据的可视化结果。如何克服ViSOM的这两个缺点,是本文的主要研究工作。

独立成分分析(Independent Component Analysis, ICA)是近几年才发展起来的一种新的统计方法,它先由C. Jutten和J. Herault[7]提出,后由Yogesh Singh C.[8]对其进行了简化。此外,Y. Rai和Z. Shi等人[9,10]对Fast ICA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。

本文第2节介绍了ViSOM算法;第3节在ViSOM的基础上提出了一种改进的MViSOM算法;第4节在ICA和MViSOM的基础上提出了一种改进的高维数据可视化模型——IMViSOM模型,并称与其对应的方法为IMViSOM方法;第5节介绍了实验结果;最后对全文做了小结。

2 ViSOM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。

ViSOM通常使用与SOM同样的网络结构。它和SOM的最大差别在于ViSOM加入了得胜神经元以及输入输出空间距离关系来调整神经元的权值向量。定义输入向量 $x\in R^n$,节点索引为 $c(c=(i,j)\in\Omega$,其中 $i=1,2,\cdots,M;j=1,2,\cdots,N)$。它的权值向量为 $w_c=[w_{c1},w_{c2},\cdots,w_{cn}]^T$。在时间步 t,输入数据为 $x(t)$,学习率为 $\alpha(t)$,邻域函数为 $\eta(v,c,t)$,其中 v 表示获胜节点的索引。

*)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:数据挖掘、人工智能;**蒋春福** 博士生,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。

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)进行 ViSOM 网络特征映射算法。其中步骤(1)可以看成是对数据的预处理。步骤(2)的具体算法如下：

Step1：初始化权值向量。将权值向量 $w_c(0), c\in\Omega$，以随机方式设定其值，但必须注意所有的权值向量的初始值都应不同；

Step2：在时间步 t，输入一个数据向量 $x(t)$，找出获胜神经元 v：$v=\arg\min_{c\in\Omega}\|x(t)-w_c\|$；

Step3：调整获胜神经元的权值向量

$w_v(t+1)=w_v(t)\alpha(t)[x(t)-w_c(t)]$；

Step4：调整获胜神经元邻域神经元的权值向量

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

Step5：重复 Step2 到 Step4，直到收敛为止。

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 $\lambda=1\sim1.5\times\frac{4\times\sqrt{Var_{\max}}}{\min\{M,N\}}$，其中 $Var_{\max}$ 表示数据的最大方差；$\eta(v,k,t)=exp\left(-\frac{\Delta_{vk}^2}{2\sigma^2(t)}\right)$，$0<\alpha(t)<1$，$\sigma(t)$ 为核宽函数，$\alpha(t)$ 和 $\sigma(t)$ 随时间 t 的增加而减少。ViSOM 可以将相同类的权值向量彼此靠近，因此 ViSOM 可以用来提供视觉化的群聚分析。

3 ViSOM 的改进算法——MViSOM 算法

假设经过训练 ViSOM 网络后，权值为 $w_c(c=(i,j),i=1,\cdots,M;j=1,\cdots,N)$，又设输入向量 $x(t)$，所对应的获胜神经元为 $v(v_i,v_j)$，但实际上 $x(t)$ 与获胜结点的权值 w_v 的距离可能并不等于零，此时将 $x(t)$ 映射到输出空间点 $v=(v_i,v_j)$ 并不能完全反映出其拓扑不变性，并且如果输出结点过多，则会消耗大量运算时间，输出结点过小，则会有大量输入向量都映射到同一结点 $v=(v_i,v_j)$ 上，这时在二维坐标平面上将很难分析数据的可视化结果。因此我们研究了一种 ViSOM 的改进算法——MViSOM 算法，其算法步骤如下：

(1)训练 ViSOM 网络；

(2)对数据点进行映射。

改进 MViSOM 算法步骤(1)与 ViSOM 网络算法是一致的，主要差别在于步骤(2)。下面对步骤(2)进行详细的讨论与说明。

设 $v=(v_i,v_j)$ 的 4 个邻域结点为 $v_1=(v_i+1,v_j)$，$v_2=(v_i,v_j+1)$，$v_3=(v_i-1,v_j)$，$v_4=(v_i,v_j-1)$，为了便于说明输入数据向量 $x(t)$ 在输出空间的映射，不妨以 $v=(v_i,v_j)$ 为原点建立直角坐标系，向量 $x(t)$ 在输出空间的映射的各种情形分别如图 1、图 2 和图 3 所示。

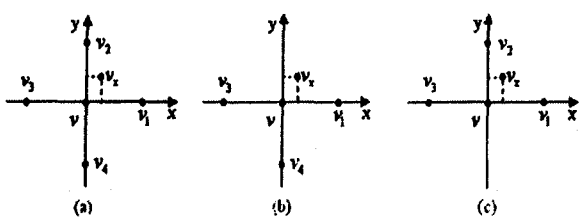

图 1　$1<v_i<M$：(a)$1<v_j<N$，(b)$v_j=N$，(c)$v_j=1$

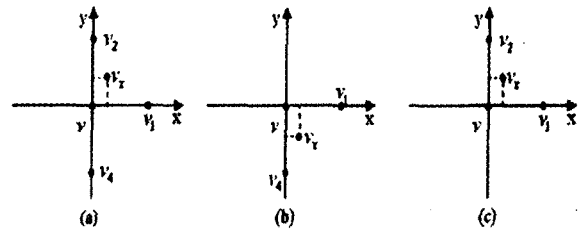

图 2　$v_i=1$：(a)$1<v_j<N$，(b)$v_j=N$，(c)$v_j=1$

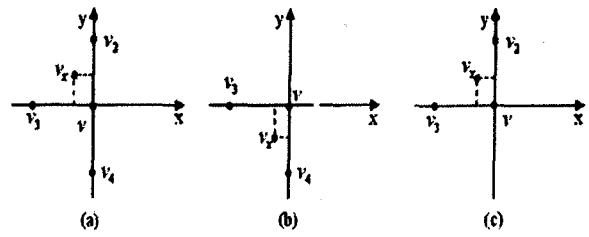

图 3　$v_i=M$：(a)$1<v_j<N$，(b)$v_j=N$，(c)$v_j=1$

用 d_{xv} 表示 $x(t)$ 与 w_v 的距离；用 $d_{xvi}(i=1,2,3,4)$ 表示 $x(t)$ 与 w_{v_i} 的距离；d_{xv1}，d_{xv3} 影响 x 轴方向的位移，d_{xv2}，d_{xv4} 影响 y 轴方向的位移，并设位移坐标为 (x,y)。

我们设计如下的算法以完成 MViSOM 算法中的步骤(2)：

① 在时间步 t，输入一个数据向量 $x(t)$，找出获胜神经元 v：$v=\arg\min_{c\in\Omega}\|x(t)-w_c\|$，$v=(v_i,v_j)$；

②如果 $d_{xv}=0$，则 $x=0$，$y=0$ 转⑧；否则如果 $d_{xv}\neq0$，给变量 max 赋一个很大的正数，转③；

③如果 $1<v_i<M$，$1<v_j<N$，则转⑥；否则如果 $1<v_i<M$，$v_j=N$，则令 $d_{xv2}=\max$，转⑥；否则如果 $1<v\vee i<M$，$v_j=1$，则令 $d_{xv4}=\max$；转⑥；

④如果 $v_i=1$，$1<v_j<N$，则令 $d_{xv3}=\max$，转⑥；否则如果 $v_i=1$，$v_j=N$，则令 $d_{xv2}=\max$，$d_{xv3}=\max$，转⑥；否则如果 $v_i=1$，$v_j=1$，则令 $d_{xv3}=\max$，$d_{xv4}=\max$，转⑥；

⑤如果 $v_i=M$，$1<v_j<N$，则令 $d_{xv1}=\max$，转⑥；否则如果 $v_i=M$，$v_j=N$，则令 $d_{xv1}=\max$，$d_{xv2}=\max$，转⑥；否则如果 $v_i=M$，$v_j=1$，则令 $d_{xv1}=\max$，$d_{xv4}=\max$，转⑥；

⑥求得位移 x 的值：

如果 $d_{xv1}<d_{xv3}$ 则 $x=\frac{d_{xv}}{d_{xv}+d_{xv1}}$；否则，如果 $d_{xv1}=d_{xv3}$，则 $x=0$；否则，如果 $d_{xv1}>d_{xv3}$，则 $x=-\frac{d_{xv}}{d_{xv}+d_{xv3}}$；

⑦求得位移 y 的值：

如果 $d_{xv2}<d_{xv4}$ 则 $y=\frac{d_{xv}}{d_{xv}+d_{xv2}}$；否则如果 $d_{xv2}=d_{xv4}$，则 $y=0$；否则如果 $d_{xv2}>d_{xv4}$，则 $y=-\frac{d_{xv}}{d_{xv}+d_{xv4}}$；

⑧求得输入的数据向量 $x(t)$ 在输出空间中的映射点 $v_x=(v_i+x,v_j+y)$。

4 IMViSOM 模型

基于 ICA 与 MViSOM 的高维数据可视化模型 IMViSOM 如图 4。

IMViSOM 算法实现可以总结为如下步骤：

步骤 1：从数据集中抽样适量样本；

步骤 2：对每一个维度的数据进行标准化处理；

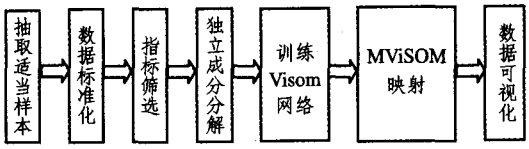

图 4　IMViSOM 模型

步骤 3：用指标筛选的方法筛选出 k 个线性无关的指标（假定 $k\geqslant 2$），详细的指标筛选算法可以参见文[11]；

步骤 4：采用 ICA 算法对筛选出 k 个线性无关的指标进行独立成分分解，获取一个独立成分数据集，详细的 ICA 算法可参见文[9,10,12]；

步骤 5：用获取的独立成分数据集训练 ViSOM 网络；

步骤 6：用改进的方法 MViSOM 对数据点进行映射；

步骤 7：二维数据可视化。

5　实验结果

实验中采用鸢尾花资料数据集，数据集中包括 3 类，第 1 类为刚毛鸢尾花，第 2 类为变色鸢尾花，第 3 类为弗吉尼亚鸢尾花；每个类由 50 组数据组成，每组数据包含四个数据指标：花瓣长度、花瓣宽度、花萼长度、花萼宽度；并且第 2,3 类是线性不可分的，通过检验，四个指标数据都存在一定的相关性。由于数据集属于高维数据资料，我们无法观察数据的分布情况，因此我们采用 ViSOM 及本文提出的 IViSOM 映射进行数据可视化分析。由于实验数据量比较小，故未做抽样操作。需要注意的是，当数据量比较大时，为节省运算开销，抽样操作这一步不能省略。实验所用设备为一台 PC 机，所用系统为 Windows XP，运行工具 SAS9.0 中文版。我们用 IViSOM 方法表示先对数据进行独立成分分解后，直接用 ViSOM 方法取得数据可视化。

为了保证实验的可行性和有效性，对 SOM、ViSOM 及 IViSOM 方法的 100×100 二维神经映射网络的一些参数作了如下规定：

(1)排序阶段：学习的最大次数为 2000；学习率的初始值为 0.9；学习率的最终值为 0.05；邻域的初始值为 70；邻域的最终值为 4；

(2)收敛阶段：学习的最大次数为 1500；学习率的初始值为 0.05；学习率的最终值为 0.01；邻域的初始值为 4；邻域的最终值为 1；

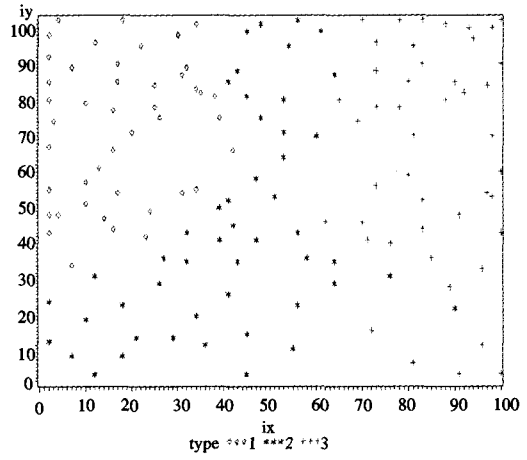

图 5　SOM 100×100 网络的实验结果

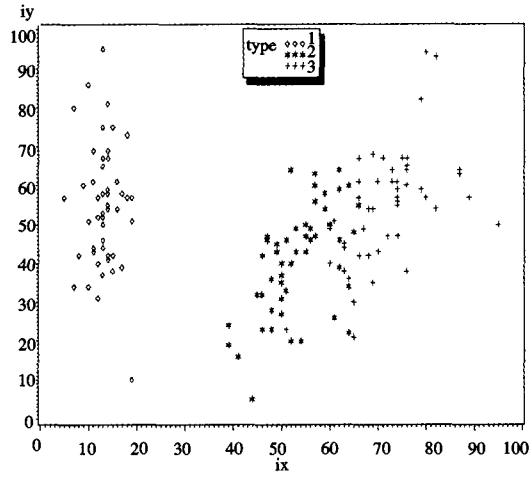

图 6　ViSOM 100×100 网络的实验结果

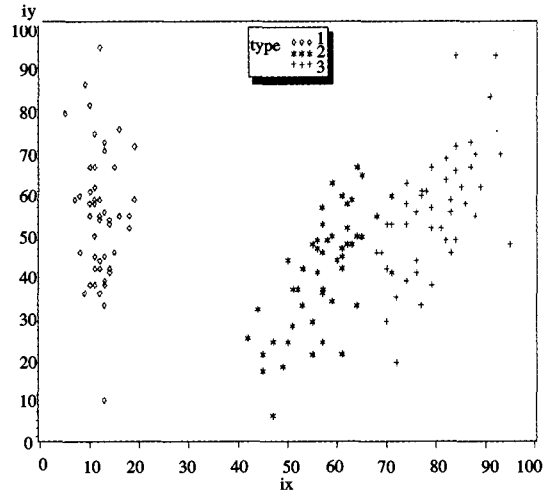

图 7　IViSOM 100×100 网络的实验结果

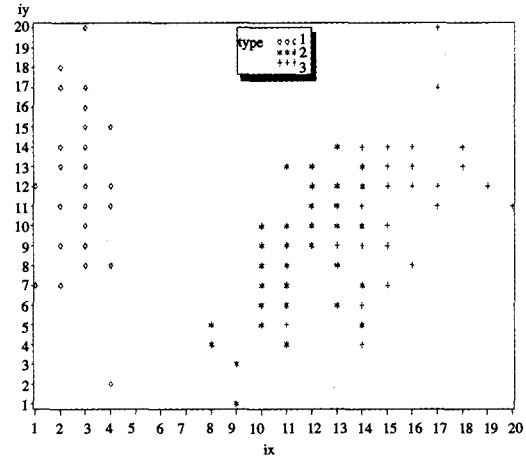

图 8　ViSOM 20×20 网络的实验结果

(3)$\lambda=0.05$(SOM 中不含有此参数)。

对 ViSOM、IViSOM 方法及 IMViSOM 方法的 20×20 二维神经映射网络的一些参数作了如下规定：

(1)排序阶段：学习的最大次数为 2000；学习率的初始值为 0.9；学习率的最终值为 0.05；邻域的初始值为 14；邻域的最终值为 2；

(2)收敛阶段：学习的最大次数为 1500；学习率的初始值

为 0.05;学习率的最终值为 0.01;邻域的初始值为 2;邻域的最终值为 0.5;

(3)$\lambda=0.25$。

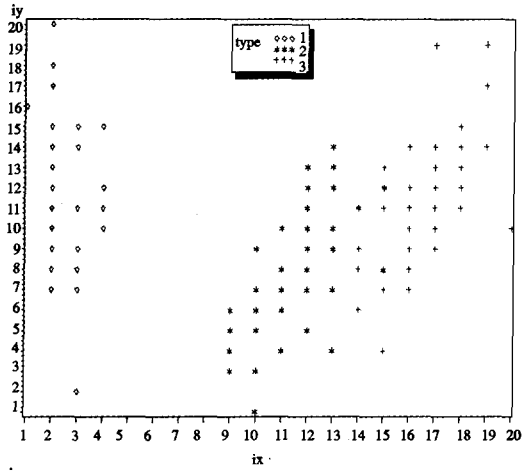

图 9 IViSOM 20×20 网络的实验结果

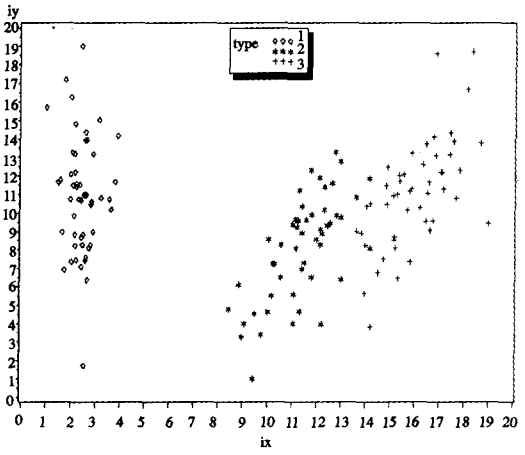

图 10 IMViSOM 20×20 网络的实验结果

表 1 各种可视化算法运行时间对比

方法	ICA	网络训练与映射	总时间
ViSOM100×100	——	15 分 44.5 秒	15 分 44.5 秒
IViSOM100×100	0.15 秒	15 分 44.5 秒	15 分 44.65 秒
IMViSOM20×20	0.15 秒	49.83 秒	49.98 秒

图 5 为 SOM 100×100 网络的实验结果,图 6 为 ViSOM 100×100 网络的实验结果,图 7 为 IViSOM 100×100 网络的实验结果;图 8 为 ViSOM 20×20 网络的实验结果,图 9 为 IViSOM 20×20 网络的实验结果,图 10 为 IMViSOM 20×20 网络的实验结果;表 1 为各种可视化算法运行时间的对比结果。从图 5 可以发现 SOM100×100 网络虽然将数据集切割成三类,却广泛地分布于二维空间中,分类效果不明显。从图 6、图 7 可以看出,ViSOM100×100 网络及 IViSOM100×100 网络方法都可以将数据集分为三类,且其分类效果都明显优于 SOM 方法。虽然第二类与第三类数据都不能用一条直线将其完全分开,但从图 6、图 7 可以看出,IViSOM100×100 网络的分类效果要略好于 ViSOM100×100 网络的分类效果,这可能是因为各数据指标之间存在一定的相关性,而经过 ICA 处理后,相关性的数据指标转化为互相独立的数据指标。通过统计计算发现原始的四个数据指标确实存在一定的相关性。从图 8 与图 9 可以看出 ViSOM 20×20 网络与 IViSOM20×20 网络由于输出结点不多,使得很多输入数据在输出空间的映射点重叠,而不便于从二维图象中看出数据集的分布效果。从图 10 可以看出 IMViSOM 20×20 的对数据的分类效果与 IViSOM 100×100 网络的分类效果差别不大,比 ViSOM 分类效果略好,但从表 1 我们可以发现,IMViSOM 20×20 网络的时间远远小于 IViSOM 100×100 与 ViSOM 100×100 网络的运行时间。

小结 由以上实验结果可以看出,与标准的 ViSOM 方法和 SOM 方法相比,本文提出的 IMViSOM 模型能够实现高维数据的可视化,不仅适用数据指标互相独立的情况,而且适用于数据指标存在相关性的情形,因此具有较好的通用性;当数据具有群聚现象,并且数据之间存在相关性时,IMViSOM 对高维数据的分类效果要优于 SOM 与 ViSOM;在不明显影响数据可视化效果的前提下,可以适当降低 IMViSOM 的网络输出结点数,从而提高网络算法的运行速度。

数据挖掘是信息时代发展很快的一个领域,高维数据的可视化是其中一个很重要的方面,原始的数据通常存在一定的相关性,本文提出的 IMViSOM 模型对存在相关性的数据作了很好的处理,并在不明显影响数据可视化效果的前提下,可以适当降低网络输出结点数,以期提高网络算法运行速度,从而为数据挖掘提供了一种有效的高维数据可视化处理技术。

参 考 文 献

1 Kohonen T. Self-organizing maps. 3rd ed. Berlin Heidelberg New York: Springer, 2001

2 Yin H. ViSOM——a novel method for multivariate data projection and structure visualization. IEEE Transaction on Neural Networks, 2002, 1: 237~243

3 Yin H. Data visualization and manifold mapping using the ViSOM. Neural Networks, 2002, 15: 1005~1016

4 Sarvesvaran S, Yin H. Visualisation of Distributions and Clusters Using ViSOMs on Gene Expresssion Data. Lecture Notes in Computer Science, 2004, 3177: 78~84

5 Wu S, et al. PRSOM: A New Visualization Method by Hybridizing Multidimensional Scaling and Self-Organizing Map. IEEE Transactions on Neural Networks, 2005, 5: 1362~1380

6 Kantardzic M. Data Mining Concepts, Models, Methods, and Algorithms. Beijing: Tsing hua University Press, 2003

7 Jutten C, Herault J. Independent component analysis versus PCA. In: Proceeding of European Symposium on Signal Processing, 1988, 2: 287~314

8 Rai Y. A simplified approach to independent component analysis. Neural Comput & Applic, 2003, 12: 173~177

9 Kocsor A, Csirik J. Fast Independent Component Analysis in Kernel Feature Spaces [J]. Lecture Notes in Computer Science, 2001, 2234: 271~281

10 Shi Z, Tang H, Tang Y. A fast fixed-point algorithm for complexity pursuit. Neurocomputing, 2005, 64 :529~536

11 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)

12 彭红毅,朱思铭,蒋春福. 数据挖掘中基于 ICA 的缺失数据值估计. 计算机科学,2005,12:203~205