Operations Research and Fuzziology 运筹与模糊学, 2011, 1, 16-21
http://dx.doi.org/10.12677/orf.2011.11004 Published Online August 2011 (http://www.hanspub.org/journal/orf/)

Variable Selection for Soft-Sensing Model Based on False Nearest Neighbors in Self-Organizing Feature Mapping Feature Space*

Jie Hou[1], Taifu Li[2], Dejun Yu[3], Yang Cheng[4]
[1]College of Automation, Chongqing University, Chongqing
[2]Department of Electrical and Information Engineering, Chongqing University of Science and Technology, Chongqing
[3]Chongqing Electric Power College, Chongqing
[4]Chongqing Academy of Agricultural Sciences, Chongqing
Email: ynhj88311@163.com
Received: Jun. 20th, 2011; revised: Jul. 5th, 2011; accepted: Jul. 10th, 2011.

Abstract: A new variable selection method based on Self-Organizing Feature Mapping (SOM) is proposed for soft-sensing modeling to eliminate redundant information. In the proposed method, SOM is used to get new space from original variable space because of its simple and fast. The False Nearest Neighbor (FNN) is used to calculate the similarity of data in the new SOM space. The primary variable would be estimated to select secondary variables. The results show that the method is effective and suitable for variable selection. Therefore, a new method is provided for the variable selection of soft-sensing modeling.

Keywords: Variable Selection; Soft-Sensoring Modeling; Self-Organizing Feature Mapping (SOM); Feature Space; False Nearest Neighbor (FNN)

基于 SOM 特征映射空间相似度判别的软传感器建模变量选择*

侯　杰[1]，李太福[2]，余德君[3]，程　杨[4]
[1]重庆大学自动化学院，重庆；[2]重庆科技学院电气与信息学院，重庆；
[3]重庆电力高等专科学校，重庆；[4]重庆市农业科学院，重庆
Email: ynhj88311@163.com
收稿日期：2011 年 6 月 20 日；修回日期：2011 年 7 月 5 日；录用日期：2011 年 7 月 10 日

摘　要：针对软传感器建模中存在的信息冗余，提出一种基于自组织特征映射神经网络(Self-Organizing Feature Mapping，SOM)的变量选择方法。该方法借助 SOM 简单快速的特征映射能力对数据进行投影，采用虚假最近邻点法(False Nearest Neighbor，FNN)计算某变量删减前后数据在 SOM 投影空间的相似度，通过相似度来判断其对主导变量的解释能力，由此进行变量的选择。实验结果表明该方法能有效的进行变量选择，为软传感器建模变量选择提供了一种新思路。

关键词：变量选择；软传感器建模；SOM 神经网络；特征空间；虚假最近邻点法

1. 引言

软传感器建模(soft sensor modeling)[1]是利用工业生产过程中有关的过程变量间的关联，通过某些能够检测的过程变量和相应的数学模型，估计出过程中未知变量的技术。在建模时其复杂度随辅助变量增加而呈指数增长，易出现维度灾难。借助变量选择技术可剔除冗余特征参数，降低模型复杂性，研究软传感器建模中的变量选择方法有重要的科学意义和学术价

*国家自然科学基金资助项目(50905194)；重庆市自然科学基金资助项目(CSTC2008BB2356)。

值。

传统变量选择方法主要有前向选择法(Forward Selection)[2]，后向剔除法(Backward Elimination)[2]，逐步回归法(Stepwise Regression)[3]等。前向选择法，后向剔除法这两种方法是单向增加或者减少变量数目，可以快速的实现变量选择，它们的缺点是：容易陷入局部极值，且只适合于线性模型。逐步回归法也叫增 l 减 r 法，即搜索方向不再是单向加或者减，可以根据评估函数灵活的浮动，其问题在于 l 和 r 的大小难以确定且此方法也只适合于线性模型。

针对上述变量选择方法在变量选择中存在的不足和缺陷，在软传感器建模中普遍采用特征提取消除原始数据空间的信息冗余，然后在映射后的特征子空间作特征选择，从而解决判别模型的精度问题、并适当降低了模型复杂度。主要的特征提取方法有：因子分析法(Factor Analysis，FA)[4]，支持向量机法(Support Vector Machine，SVM)[5]，主成分分析法(Principal Component Analysis，PCA)[6]，以及它们的一些组合方法。这些方法在特征提取方面已取得显著的成绩。但是它们的不足和缺陷在于：FA 计算因子时，采用的是最小二乘法，此法有时可能会失效；SVM 算法对大规模训练样本难以实施；PCA 特征提取时间长。

针对这些特征提取方法进行变量选择的不足和缺陷，一种具有很好的非线性映射投影能力和很好的非监督学习能力，可以对大样本进行简单，快速处理的 SOM 神经网络被一些研究者所注意，被用来进行变量选择。文献[7]使用 SOM 进行变量选择，该方法由于只是借助可视化 $\boldsymbol{U}$ 矩阵分析来进行变量选择，具有很大的模糊不确定性。文献[8]使用 SOM 神经网络对分析化学进行变量选择，该方法借助统计学方法进行计算，运算量大，且对选择变量没有很好的解释能力；文献[9]使用 SOM 神经网络对光谱分析进行变量选择，该方法借助统计学方法进行计算且运算量大，对选择变量没有很好的解释能力；文献[10]使用 SOM 神经网络进行多次重复建模计算进行变量选择，该方法过程复杂，使用多个不同的神经网络来建模，稳定性差。

本文提出一种结合自组织特征映射神经网络(Self-Organizing Feature Mapping，SOM)与虚假最近邻点法(False Nearest Neighbours，FNN)的变量选择新方法，简称 SOM-FNN 法。该方法结合利用了 SOM 神经网络 $\boldsymbol{U}$ 矩阵的强大可视化功能及 SOM 能简单，快速的对数据进行非监督非线性特征映射的优势以及 FNN 法的简单可行的计算能力，从而有效的克服只简单对 SOM 神经网络可视化分析时的模糊不确定，使用统计学方法计算量大，不便理解以及重复建模不稳定等问题。通过简单，方便的计算即可进行变量选择，从而有效解决上述文献存在的问题。本文使用 Matlab2009a 应用 somtoolbox 软件工具箱来对 SOM-FNN 进行了仿真研究。SOM 神经网络输出层节点数均采用 $5*\sqrt{\text{样本个数}}$ [11]。

2. 相关理论基础简介

2.1. SOM 神经网络和 U 矩阵

SOM 神经网络是芬兰赫尔辛基大学神经网络专家 Kohonen 于 1981 年提出的竞争式神经网络。它在训练中能无监督地进行自组织学习，可以把高维信息数据以有序方式映射到低维网络，形成一种拓扑意义上的有序图。SOM 神经网络由输出层和输入层 2 层构成。

可视化技术在数据结构分析中有着重要作用。SOM 神经网络的训练结果也必须借助其他方法才能实现可视化。SOM 神经网络常采用 $\boldsymbol{U}$ 矩阵，$\boldsymbol{D}$ 矩阵来对其进行可视化。$\boldsymbol{U}$ 矩阵是表示神经网络权矢量与其领域内的神经网络节点之间距离的度量的可视化技术，由 Ultscu 于 1998 年提出[12]。

2.2. 虚假最近邻点法(FNN)

虚假最近邻点法[13]是在高维相空间重构过程中，随着嵌入维数 m 增大，从混沌时间序列中恢复混沌运动的轨迹逐渐打开，相似度高的虚假邻点被逐步剔除，从而使混沌运动的轨迹得到恢复的一种特征筛选方法。

与计算相距离相比，计算相点的相关性能更全面地解释虚假最近邻点这一现象。假设一个 n 维变量组空间 Q，其中的一个样本点 $\mathbf{A}=(q_1,q_2,\cdots,q_i,\cdots,q_n)$，令 $q_i=0$，求出样本 $\mathbf{A}$ 在第 i 维的空间内的投影 $\mathbf{B}=(q_1,q_2,\cdots,0,\cdots,q_n)$。

计算 $\mathbf{A}$，$\mathbf{B}$ 的相关性：

$$\cos(\theta)=\frac{\mathbf{A}\cdot\mathbf{B}^{\mathrm{T}}}{\|\mathbf{A}\|\cdot\|\mathbf{B}\|} \quad (1)$$

若 $\cos(\theta)$ 值接近于 1，则说明 **A** 与 **B** 的相似度大，变量 q_i 对样本影响小，解释能力小；若 $\cos(\theta)$ 较大地偏离 1，说明 **A** 与 **B** 相似性小，变量 q_i 对样本解释能力较大，此时 **B** 即 **A** 的虚假最近邻点。

3. SOM-FNN 变量选择方法

SOM-FNN变量选择方法即通过SOM神经网络将原始输入数据通过特征映射到新的投影空间，形成 $m\times n$ 维矩阵(其中m，n为神经网络输出节点数)，得到数据在SOM特征投影空间的投影向量。接着，使用FNN思想，将原变量组 $\mathbf{A}=(q_1,q_2,\cdots,q_i,\cdots,q_n)$ 中的任意变量 q_i 置为零，得一新的变量组 $\mathbf{B}=(q_1,q_2,\cdots,0,\cdots,q_n)$，然后将这两个点利用SOM神经网络进行特征投影将它们转换到SOM特征空间，再利用虚假最近邻点法(FNN)计算转换后两个点之间的相似度 $\cos(\theta)$。把所得到的 n 个 $\cos(\theta)$ 按从小到大的顺序排列，较大相似度对应的原始变量即没什么贡献，可以考虑删除。

根据以上分析得到 SOM-FNN 变量选择方法的算法和具体实现步骤如下：

SOM-FNN 算法实现步骤如下：

Step1：首先使用所有变量来对 SOM 神经网络进行训练。得到和记录相应的神经网络权值 **W**。

Step2：原来数据记为 $\mathbf{A}=(q_1,q_2,\cdots,q_i,\cdots,q_n)$，计算相应相似度值 $\mathbf{A}'$。

Table 1. Algorithm of SOM-FNN
表1. SOM-FNN算法

输入：$\mathbf{A}=(q_1,q_2,\cdots,q_i,\cdots,q_n)$ $\mathbf{B}=(q_1,q_2,\cdots,0,\cdots,q_n)$

Step1：实验 **A** 训练 SOM 神经网络得到权值向 **W**

Step2：虚假最近邻点法 FNN 计算相似度值

For i=1:S

For n=1:N

$\mathbf{A}'=\sum\mathbf{A}*\mathbf{W}^{\mathrm{T}}$

$\mathbf{B}'=\sum\mathbf{B}*\mathbf{W}^{\mathrm{T}}$

$\cos(\theta)=\frac{\mathbf{A}'\cdot\mathbf{B}'^{\mathrm{T}}}{\|\mathbf{A}'\|\cdot\|\mathbf{B}'\|}$

END

END

注释：S：变量数；N：输出层神经元个数；输出：变量平均相似度 $\cos(\theta)$。

Step3：将相应的待分析的变量置零的数据记为 $\mathbf{A}=(q_1,q_2,\cdots,q_i,\cdots,q_n)$，计算相应的 $\mathbf{B}'$。

Step4：计算平均相似度值 $\cos(\theta)=\frac{\mathbf{A}'\cdot\mathbf{B}'^{\mathrm{T}}}{\|\mathbf{A}'\|\cdot\|\mathbf{B}'\|}$ 确定各变量的重要程度，以进行变量选择。

Step5：把所得到的 n 个 $\cos(\theta)$ 按从小到大的顺序排列，较大 $\cos(\theta)$ 对应的原始变量没什么贡献，可以考虑剔除。

SOM-FNN变量选择方法如图1所示。

4. 仿真研究

4.1. 仿真数据

数据 1：使用论文[14]的数据来对 SOM-FNN 进行仿真研究。模型如下：

$$\mathbf{Y}=51+3\mathbf{X}_4+4\mathbf{X}_5+\boldsymbol{\varepsilon} \quad (2)$$

数据 2：使用论文[15]的数据对 SOM-FNN 进行仿真研究。模型如下：

$$\mathbf{Y}=10\sin\pi\mathbf{x}_1\mathbf{x}_2+20(\mathbf{x}_3-0.5)^2+10\mathbf{x}_4+5\mathbf{x}_5+\boldsymbol{\varepsilon} \quad (3)$$

数据 3：使用论文[15]的数据对 SOM-FNN 进行仿真研究。模型如下：

$$\mathbf{T}=\mathbf{X}_1\mathbf{X}_2+\overline{\mathbf{X}}_1\overline{\mathbf{X}}_2\overline{\mathbf{X}}_3\overline{\mathbf{X}}_4 \quad (4)$$

4.2. 仿真结果及分析

模型 1：使用全部数据训练 SOM 神经网络得到神经网络可视化 ***U*** 矩阵图如图 2 所示，SOM-FNN 计算得到的各变量平均相似度值如图 3 所示。

结果分析：

U 矩阵相关度不是太高。不是某一变量直接可以表示数据信息。而是某些变量混合作用的结果。

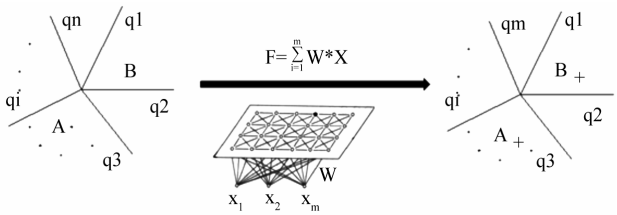

Figure 1. Schematic diagram of SOM-FNN
图1. SOM-FNN示意图

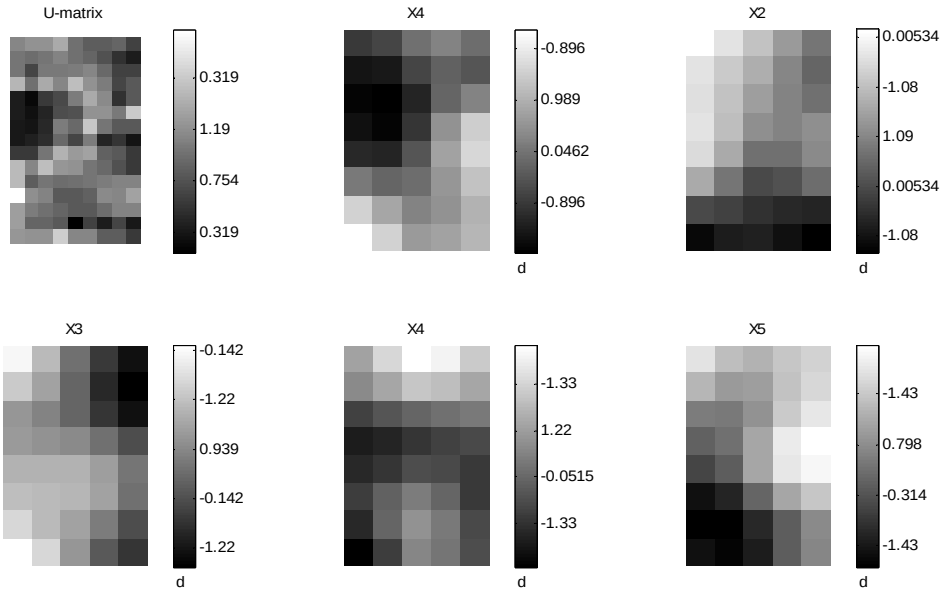

Figure 2. *U*-matrix map of data 1
图 2. 数据 1 的 *U* 矩阵图

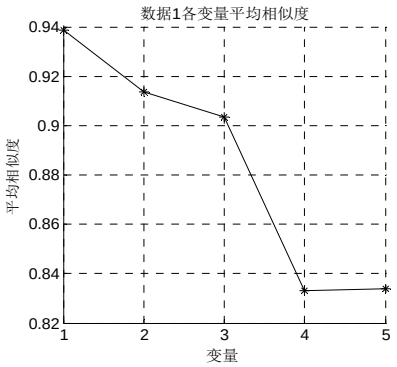

Figure 3. Average similarity of each variable of data 1
图 3. 数据 1 各变量平均相似度

使用 SOM-FNN 法对各变量进行相似度计算得到 5 个变量的平均相似度值和建模时各变量被选择的顺序如下表 2 所示，符合构造模型预期效果。

模型 2：使用全部数据训练 SOM 神经网络得到神经网络可视化 ***U*** 矩阵图如图 4 所示，SOM-FNN 计算得到的各变量平均相似度值如图 5 所示。

Table 2. Results of data 1
表 2. 数据 1 计算结果

变量系数	平均相似度	被选顺序
1	0.9386	5
2	0.9137	4
3	0.9034	3
4	0.8330	1
5	0.8338	2

结果分析：

U 矩阵可以看出变量 9,10；变量 4,5 一致，通过 SOM 神经网络 ***U*** 矩阵显示。***U*** 矩阵和具体某一个变量没有直接相关关系，而应该是某些变量的混合表示的结果。

使用 SOM-FNN 法对各变量进行相似度计算得到 10 个变量的平均相似度值和建模时各变量被选择的顺序如下表 3 所示，符合构造模型预期效果。

模型 3：使用全部数据训练 SOM 神经网络得到神经网络可视化 ***U*** 矩阵图如图 6 所示，SOM-FNN 计算得到的各变量平均相似度值如图 7 所示。

Table 3. Results of data 2
表 3. 数据 2 计算结果

变量系数	平均相似度	被选顺序
1	0.9991	3
2	0.9994	4
3	0.9994	5
4	0.9991	2
5	0.9991	1
6	0.9995	6
7	0.9995	7
8	0.9995	8
9	0.9995	9
10	0.9995	10

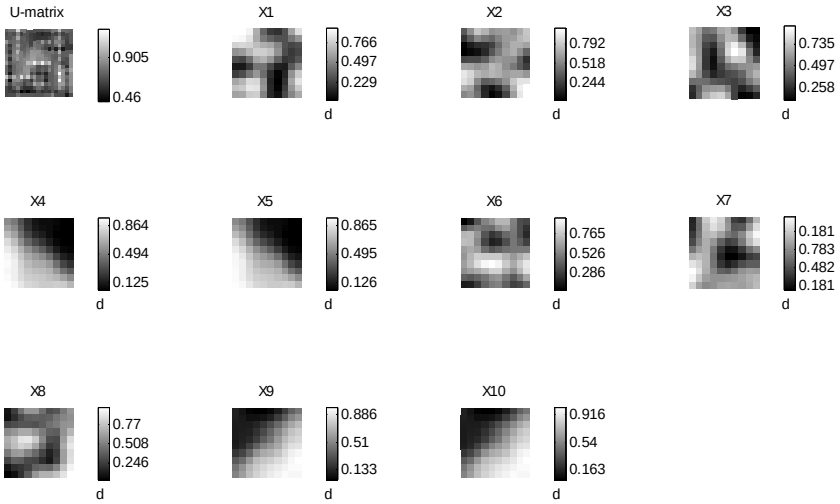

Figure 4. *U*-matrix map of data 2
图 4. 数据 2 的 *U* 矩阵图

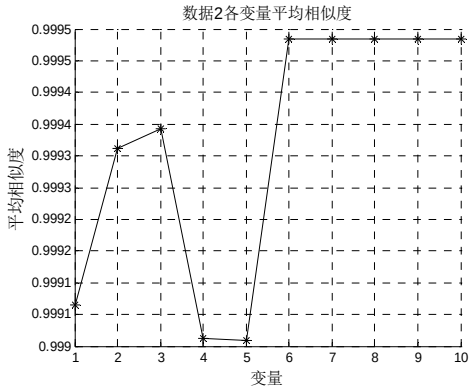

Figure 5. Average similarity of each variable of data 2
图 5. 数据 2 各变量平均相似度

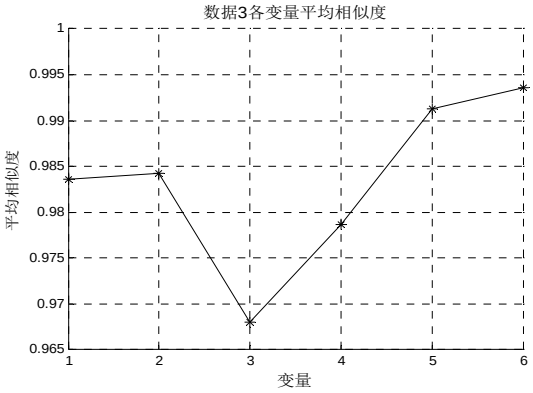

Figure 7. Average similarity of each variable of data 3
图 7. 数据 3 各变量平均相似度

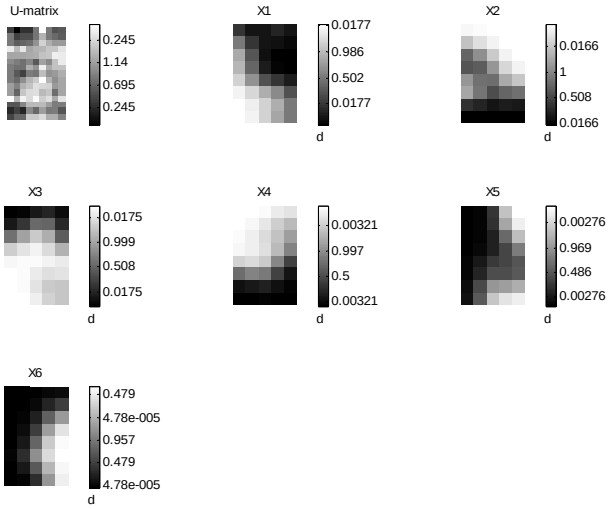

Figure 6. *U*-matrix map of data 3
图 6. 数据 3 的 *U* 矩阵图

结果分析：

U 矩阵相关度不是太高。不是某一变量直接可以表示数据信息。而是某些变量混合作用的结果。

使用 SOM-FNN 法对各变量进行相似度计算得到 6 个变量的平均相似度值和建模时各变量被选择的顺序如表 4 所示，符合构造模型预期效果。

5. 结语

1) 利用 SOM-FNN 法，通过计算某辅助变量选择前后的相似度，可以有效地剔除冗余信息，降低模型复杂度。

2) SOM-FNN 法通过计算变量选择前后的相似度，在建立模型时可以依次得到不同变量的被选顺序，依次加入变量来建立模型，可以得到理想的模型效果。

Table 4. Results of data 3
表 4. 数据 3 计算结果

变量系数	平均相似度	被选顺序
1	0.9835	3
2	0.9843	4
3	0.9680	1
4	0.9786	2
5	0.9913	5
6	0.9935	6

且计算量小，易于编程实现，可在实际中加以应用。

3) SOM-FNN 法可以借助 SOM 神经网络的可视化 $\boldsymbol{U}$ 矩阵对结果进行分析，同时，由于结合了 FNN 的计算可以简单，快速的计算得到结果。

参考文献 (References)

[1] 王孝红, 刘文光, 于宏亮. 工业过程软测量研究[J]. 济南大学学报: 自然科学版, 2009, 23(1): 80-86.

[2] M. Gevrey, I. Dimopoulos, and S. Lek. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecological Modelling, 2003, 160(3): 249-264.

[3] Q. R. Chen, C. H. Yang. Quasi-stepwise regression variable selection and its application in rural household net income forecasting. Systems Engineering: Theory & Practice, 2008, 28(11): 16-22.

[4] 贾洪飞, 隽志才, 王晓原等. 利用因子分析选取车辆跟驰模型输入变量[J]. 公路交通科技, 2004, 21(1): 81-84.

[5] K. O. Elish, M. O. Elish. Predicting defect-prone software modules using support vector machines. The Journal of Systems and Software, 2008, 81(5): 649-660.

[6] F. Westad, M. Hersleth, P. Lea, et al. Variable selection in PCA in sensory descriptive and consumer data. Food Quality and Preference, 2003, 14(5-6): 463-472.

[7] W.-S. Lee, Y.-S. Kwon, J.-C. Yoo, et al. Multivariate analysis and self-organizing mapping applied to analysis of nest-site selectionin Black-tailed Gulls. Ecological Modelling, 2006, 193(3-4): 602-614.

[8] G. R. Lloyd, K. Wongravee, C. J. L. Silwood, et al. Self organising maps for variable selection: Application to human saliva analyzed by nuclear magnetic resonance spectroscopy to investigate the effect of an oral health care product. Chemometrics and Intelligent Laboratory Systems, 2009, 98(2): 149-161.

[9] F. Corona, E. Liitiainen, A. Lendasse, et al. A SOM-based approach to estimating product properties from spectroscopic measurements. Neurocomputing, 2009, 73(1-3): 71-79.

[10] K. M. Najman, K. Najman. Applying the Kohonen self-organizing map networks to select variables. Data Analysis, Machine Learning and Applications, 2008, 1: 45-54.

[11] Y.-S. Park, J. Tison, S. Lek, et al. Application of a self- organizing map to select representative species in multivariate analysis: A case study determining diatom distribution patterns across France. Ecological Informatics, 2006, 1(3): 247-257.

[12] 廖广兰, 陈勇辉, 史铁林. 自组织映射网络的可视化研究[J]. 计算机工程与应用, 2003, 39(9): 35-37.

[13] 王海燕, 盛昭瀚. 混沌时间序列相空间重构参数的选取方法[J]. 东南大学学报(自科版), 2000, 30(5): 113-117.

[14] R. Philips, I. Guttman. A new criterion for variable selection. Statistics & Probability Letters, 1998, 38(1): 11-19.

[15] A. Eleuteri, R. Tagliaferri, and L. Milano. A novel information geometric approach to variable selection in MLP networks. Neural Networks, 2005, 18(10): 1309-1318.

