

基于 Voronoi 距离的鲁棒的双自组织特征映射网络

夏文文, 王士同

(江南大学 信息工程学院, 江苏 无锡 214122)

(xwwily@sohu.com)

摘要:提出了一种基于 Voronoi 距离的双自组织特征映射网络。该网络通过同时使用两个相关的映射网络扩展了原有的自组织神经网络。针对自组织特征映射网络容易受到高噪声的影响, 通过使用 Voronoi cell 的距离来取代网络中的欧式距离, 增强了网络的鲁棒性。将改进后的神经网络用于金融时间序列的预测, 实验结果表明改进后的神经网络具有较强的鲁棒性。

关键词:自组织映射网络; 鲁棒; Voronoi 图

中图分类号: TP183 **文献标识码:** A

Research of the improved robust twinned SOM based on Voronoi distance

XIA Wen-wen, WANG Shi-tong

(School of Information Engineering, Southern Yangtze University, Wuxi Jiangsu 214122, China)

Abstract: An improved twinned Self-Organizing Maps(SOM) based on Voronoi distance was presented in this paper. The traditional SOM is extended by using two related neuron networks simultaneously in order to enhance the robustness. Euclidean distance was replaced by the distance to the Voronoi cell in the proposed SOM. We illustrated the prediction power of the proposed SOM on a real financial time series and artificial data sets. Results demonstrate the effectiveness and robustness of the proposed SOM.

Key words: Self-Organizing Maps(SOM); robustness; Voronoi diagram

0 引言

自组织特征映射网络具有自稳定性, 能够识别向量空间中最有意义的特征。^[1]

在统计学中, 为了研究两组变量的相关性, 可以把两组变量的相关性转化为两个变量的相关性来考虑, 即考察第一组变量的线性组合与第二组变量的线性组合的相关性。通过选择线性系数使线性化后的变量有最大的相关系数, 形成第一对、第二对、第三对典型变量, 并使各对典型变量之间互不相关, 这样就将两组变量间的相关转化为几对典型变量间的相关。典型相关分析(Canonical Correlation Analysis, CCA)就是要找到这两组变量线性组合的系数, 使得这两个由线性组合生成的变量(和其他线性组合相比)之间的相关系数最大。现在已经有了一种执行 CCA 的神经网络模型^[2], 该神经网络模型也可以用来预测。

本文引入了成对的数据集的思想^[3], 即将本来用于单个数据集的算法, 用到成对的两个数据集中, 根据 CCA 的思想提出双自组织特征映射网络(twinned self-organizing maps)这个新颖的网络模型, 使用该网络预测一个金融时间序列, 由实验的结果得到, 该网络的缺点是其性能受到噪声的影响较大。

为了改进自组织特征映射网络模型, 使得其抗噪声能力增强, 提出用 Voronoi cell 距离取代 SOM 中使用的欧式距离, 使得在有噪声情况下, SOM 具有较好的鲁棒性。将改进的思想用于两个同时训练的并且相互关联的自组织特征映射网络, 进行预测实验, 其实验结果表明改进后的网络有更好的鲁棒性。

1 自组织特征映射网络和双自组织特征映射网络

1.1 自组织特征映射网络 SOM

SOM 神经网络是一个两层的神经网络, 如图 1 所示, 它包括一个输入层和一个输出层, 输入层所在空间称为样本空间, 输出层一般组织成网格形式, 称为输出网格, 是一个低维空间, 又称为拓扑空间, 如果输出层的宽度为 m , 则输出层一共有 $M = m \times m$ 个神经元。该网络可以把任意高维的输入映射到低维空间, 并且使得输入数据内部的某些相似性质表现为几何上邻近的特征映射, 在输出层映射为二维离散图形, 并且保持其拓扑结构不变。SOM 神经网络在很多领域获得了成功的应用, 它是聚类分析的强有力的工具, 但仍然存在着不少的缺陷, 如学习算法获胜邻域的调整, 网络的收敛速度和分类精度的改进, 以及能够更加清晰的可视化等问题。

设网络的输入模式为 $P_k = (p_1^k, p_2^k, \dots, p_n^k)$, $k = 1, 2, \dots, q$, k 表示第几个输入模式, n 是输入向量的维数。竞争层神经元 j 与输入层神经元之间的连接权矢量为

$W_j = (w_{j1}, w_{j2}, \dots, w_{jn})$, $i = 1, 2, \dots, n$; $j = 1, 2, \dots, M$

输出层分布着网络的 M 个神经元。网络的工作和学习规则如下:

(1) 初始化: 将网络的连接权 $\{w_{ij}\}$ 赋予 $[0, 1]$ 区间内的随机值, 确定学习率 $\eta(t)$ 的初始值为 $\eta(0)$ ($0 < \eta(0) < 1$); 确定邻域 $N_g(t)$ 的初值 $N_g(0)$ 及总学习次数。邻域 $N_g(t)$ 是指以获胜神经元为中心, 且包括若干神经元的区域。

(2) 任选 q 个学习模式中的一个模式 P_k 提供给网络的输入层, 并进行归一化处理:

$$\bar{P}_k = \frac{P_k}{\|P_k\|} = \frac{(p_1^k, p_2^k, \dots, p_n^k)}{[(p_1^k)^2 + (p_2^k)^2 + \dots + (p_n^k)^2]^{1/2}} \quad (1)$$

收稿日期: 2006-11-27; 修订日期: 2007-01-27

基金项目: 国防应用基础研究基金资助项目(A1420061266); 教育部 05 年度科学研究重点基金资助项目(105087)

作者简介: 夏文文(1982-), 女, 安徽蚌埠人, 硕士研究生; 王士同(1964-), 男, 江苏扬州人, 教授, 博士生导师, 主要研究方向: 人工智能、模式识别、图像处理、生物信息学

(3) 对连接权矢量 $W_j = (w_{j1}, w_{j2}, \dots, w_{jn})$ 进行归一化处理, 计算 \overline{W}_j 与 P_k 之间的欧式距离:

$$\overline{W}_j = \frac{W_j}{\|W_j\|} = \frac{(w_{j1}, w_{j2}, \dots, w_{jn})}{[(w_{j1})^2 + (w_{j2})^2 + \dots + (w_{jn})^2]^{1/2}} \quad (2)$$

$$d_j = \left[\sum_{i=1}^n (\overline{p}_i^k - \overline{w}_{ji})^2 \right]^{1/2}, j = 1, 2, \dots, M \quad (3)$$

(4) 找出最小距离 d_c , 确定获胜神经元 c

$$d_c = \min_j [d_j], j = 1, 2, \dots, M \quad (4)$$

(5) 进行连接权的调整, 对竞争层邻域 $N_g(t)$ 内所有的神经元与输入层神经元之间的连接权进行修正:

$$\overline{w}_{ji}(t+1) = \overline{w}_{ji}(t) + \eta(t) \cdot [\overline{p}_i^k - \overline{w}_{ji}(t)], \\ j \in N_g(t), j = 1, 2, \dots, M \quad (5)$$

(6) 选取另一个学习模式提供给网络的输入层, 返回步骤(3)。

(7) 更新学习率 $\eta(t)$ 和邻域 $N_g(t)$ 。

$$\eta(t) = \eta(0) \left(1 - \frac{t}{T}\right)$$

t 为学习次数, T 为总的学习次数。

$$N_g(t) = \text{INT}[N_g(0) \left(1 - \frac{t}{T}\right)]$$

$\text{INT}[*]$ 为取整符号。

(8) 令 $t = t + 1$, 返回步骤(2), 直到 $t = T$ 为止。

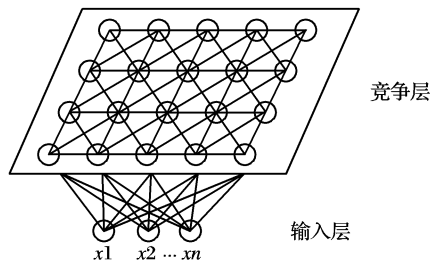


图1 SOM网络基本结构

1.2 双自组织特征映射网络的介绍

本文扩展了SOM神经网络模型, 使得两个SOM网络通过获胜神经元关联。设网络的输入模式为

$$P_k = (p_1^k, p_2^k, \dots, p_{n1}^k), k = 1, 2, \dots, q$$

和

$$R_k = (r_1^k, r_2^k, \dots, r_{n2}^k), k = 1, 2, \dots, q$$

其中一个网络的竞争层神经元与输入层神经元之间的连接权矢量为

$$W_j = (w_{j1}, w_{j2}, \dots, w_{jn1}), i = 1, 2, \dots, n1; j = 1, 2, \dots, M$$

同时另一个网络的连接权矢量为

$$V_i = (v_{i1}, v_{i2}, \dots, v_{in2}), i = 1, 2, \dots, n2; j = 1, 2, \dots, M$$

输出层分布着网络的 M 个神经元。网络的工作和学习规则如下:

(1) 初始化: 将网络的连接权 $\{w_{ji}\} \setminus \{v_{ij}\}$ 赋予 $[0, 1]$ 区间内的随机值, 确定学习率 $\eta(t)$ 的初始值 $\eta(0)$ ($0 < \eta(0) < 1$), 确定邻域 $N_g(t)$ 的初值为 $N_g(0)$ 及总学习次数。邻域 $N_g(t)$ 是指以获胜神经元为中心, 且包括若干神经元的区域。

(2) 任选 q 个学习模式中的两个模式 P_k 和 R_k 分别提供给两个网络的输入层。

(3) 计算欧式距离:

$$d_j = d_E(P_k, W_j) + d_E(R_k, V_j), j = 1, 2, \dots, M \quad (6)$$

(4) 找出最小距离 d_c , 确定获胜神经元 c

$$d_c = \min_j [d_j], j = 1, 2, \dots, M \quad (7)$$

(5) 进行连接权的调整, 对竞争层邻域 $N_g(t)$ 内所有的神经元与输入层神经元之间的连接权按下式进行修正 ($j \in N_g(t); j = 1, 2, \dots, M$):

$$w_{ji}(t+1) = w_{ji}(t) + \eta(t) [p_i^k - w_{ji}(t)] \quad (8)$$

$$v_{ji}(t+1) = v_{ji}(t) + \eta(t) [r_i^k - v_{ji}(t)] \quad (9)$$

(6) 选取另一个学习模式提供给网络的输入层, 返回步骤(3)。

(7) 更新学习率 $\eta(t)$ 和邻域 $N_g(t)$ 。

$$\eta(t) = \eta(0) \left(1 - \frac{t}{T}\right)$$

t 为学习次数, T 为总的学习次数。

$$N_g(t) = \text{INT}[N_g(0) \left(1 - \frac{t}{T}\right)]$$

$\text{INT}[*]$ 为取整符号。

(8) 令 $t = t + 1$, 返回步骤(2), 直到 $t = T$ 为止。

2 利用 Voronoi 距离改进双自组织特征映射网络

当SOM网络模型受到输入样本的刺激后, 可以得到网格响应的位置, 这个位置是所有神经元竞争的结果, 其竞争的基础是以欧式距离为依据的, 即欧式距离作为相似度度量的依据。响应的位置 r' 可按如下式子计算:

$$\|x - w_{r'}\| < \|x - w_r\|, r \in A$$

A 是输出空间。获胜神经元的感知域是一个 Voronoi 多边形^[4], 同一个多边形内的输入向量被映射到输出空间上同一点。该获胜神经元定义了一个输入空间中的子区域, 该子区域中所有的点离获胜神经元的向量的距离小于其他任何神经元。

Voronoi 图是对欧式平面的一个分割, 欧式平面上存在一组点集 $p = \{p_1, p_2, p_3, \dots\}$, 分割之后的 Voronoi 图将每一个点分配到最近的区域。一个 Voronoi Cell (也称为 Voronoi 区域) 包含所有被分配到这个区域的点, 可以表示为:

$$V(p_i) = \{x: |x - p_i| \leq |x - p_j|, j \neq i\}$$

Voronoi 图的特点是在该图中, 一个区域中的所有的点 (包括边界) 到该区域的距离都比到该平面中其他区域的距离要近。

一个数据对象 x_j 到超平面的欧式距离可以表示为^[5]:

$$|(x_j - h_s)^T n_s|$$

那么这个距离便把类 p_i 和 p_s 分开, 其中 h_s 是超平面上的一个点, 例如, 可以令

$$h_s = (p_s + p_i)/2$$

n_s 是法向量:

$$n_s = \beta_s \cdot (p_i - p_s)$$

其中:

$$\beta_s = \frac{1}{\|p_i - p_s\|}, s \neq i$$

在不取绝对值的情况下, 距离 $(x_j - h_s)^T n_s$ 为有向距离, 并且当 x_j 和类中心在同一边时该距离为正值, 反之为负值。所有的有向距离绝对值的最小值为到单元边界的距离。如果 x_j 在类 i 的 Voronoi cell 之中, 则到单元的距离为零。可以很容易的形式化这种特殊的情况, 设 $\beta_s = 1$ 并且定义:

$$d_v(x_j, p_i) = \left| \min_{1 \leq s \leq c} (x - h_s)^T n_s \right| \quad (10)$$

如果不将法向量 n_s 衡量为最小整数长度, 而对所有 s 都假设 $\beta_s = 1$, 我们保存 d_v 的形状 (超平面的位置不变), 仅仅是不同超平面变化的斜率。对 (10) 式进行简单的变形:

$$\begin{aligned}
d_v(x, p_i) &= \left| \min_{1 \leq s \leq c} \left(x - \frac{p_s + p_i}{2} \right)^T (p_i - p_s) \right| \\
&= \left| \min_{1 \leq s \leq c} x^T (p_i - p_s) - \frac{1}{2} (p_i^T p_i - p_s^T p_s) \right| \\
&= \frac{1}{2} \left| \min_{1 \leq s \leq c} \|x - p_s\|^2 - \|x - p_i\|^2 \right| \\
&= \frac{1}{2} \left(\|x - p_i\|^2 - \min_{1 \leq s \leq c} \|x - p_s\|^2 \right)
\end{aligned}$$

在上式中,我们用到了一个事实,即任何 $d_E(x, p_i)$ 都大于或等于 $\min_{1 \leq s \leq c} d_E(x, p_s)$, 所以:

$$\|x - p_i\|^2 - \min_{1 \leq s \leq c} \|x - p_s\|^2 \geq 0$$

由此可以得到,给定一个由一组不同的点 $p_i (1 \leq i \leq c)$ 和点 x 引起的 Voronoi 图。对所有的 $S (1 \leq s \leq c)$, 令 $\beta_s = 1$, x 到 p_i 的 Voronoi cell 的距离给定为:

$$d_v(x, p_i) = \frac{1}{2} \left(d_E^2(x, p_i) - \min_{1 \leq s \leq c} d_E^2(x, p_s) \right) \quad (11)$$

因此,可以用一个 Voronoi cell 距离替代欧式距离。本文中,定义 SOM 网络的输入向量为 $X \in R^n$, 输出神经元权向量为 W_j , 输出层有 M 个神经元, 数据空间中 X 与 W_j 的距离度量为 $d(X, W_j)$, 可以得到如下的结果:

$$d_v(X, W_j) = \frac{1}{2} \left(d_E^2(X, W_j) - \min_{1 \leq s \leq M} d_E^2(X, W_s) \right) \quad (12)$$

我们知道,一般的广义 P 范式距离定义为:

$$d_p(X, W_j) = \left(\sum_{i=1}^n |x_i - w_{ji}|^p \right)^{\frac{1}{p}} \quad (13)$$

根据文献[6], 当时 $P = 1$, 这个式子的抗噪声能力最强, 但它因是绝对值的形式不易求导数, 而 $P = 2$ 时, 为欧式距离, 也最常用, 但其抗噪声能力却相对较弱。

这里用 Voronoi cell 距离替代欧式距离, 通过减去一个最小值, 相当于使 P 的范围缩小到 1 到 2 之间, 既解决了不容易求导数的问题, 又使得网络对噪声和离群数据有较少的敏感度。

3 改进后的双自组织特征映射网络

将 Voronoi 距离用于双自组织特征映射网络中, 即用 Voronoi cell 距离替代欧式距离, 得到新的双 SOM 网络模型, 网络的训练过程如下:

- (1) 初始化: 使用小的随机数对连接权值 $\{w_{ij}\}$ 、 $\{v_{ij}\}$ 进行初始化, 确定学习率 $\eta(t)$ 和邻域 $N_g(t)$ 的初始值。
- (2) 从样本集中选择任意的模式提供给网络的输入层。
- (3) 计算获胜节点, 输出层中和输入模式最接近的节点, 这里利用 Voronoi 距离代替欧式距离:

$$d_c = \arg \min_j \{d_v(X_1, W_j) + d_v(X_2, V_j)\}$$

- (4) 对网络上获胜神经元 c 拓扑邻域内的神经元进行权值向量的更新。

$$W_j(t+1) = W_j(t) + \eta(t)(X_1 - W_j(t))$$

$$V_j(t+1) = V_j(t) + \eta(t)(X_2 - V_j(t))$$

$$t = 0, 1, 2, \dots, T$$

更新学习速率以及拓扑邻域。

- (5) 判断迭代次数 t 是否超过 T , 如果 $t < T$, 就转到第(2)步, 否则结束迭代过程。为了检验网络的精确度, 这里只用第一个数据集 X_1 得到获胜神经元

$$d_c = \arg \min_j \{d_v(X_1, W_j)\}$$

确定出获胜神经元 c , 然后使用 $|X_2 - V_c|$ 作为这个预测的误

差的度量值。

4 双自组织特征映射网络的预测实验

将 Voronoi 距离用于双自组织特征映射网络中, 并且将改进后的网络应用于对金融时间序列数据的预测, 发现其具有较好的鲁棒性。

我们用平均绝对百分误差 (MAPE) 作为预测结果的评价值, 平均绝对百分误差的计算公式如下:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{A_i - F_i}{A_i} \right|$$

其中 A_i 为预测对象实际值, F_i 为预测值。

实验 1 真实数据的实验。用美元兑换人民币的汇率值做一个预测的实验, 金融预测的数据集一般具有: 高噪声、不稳定、非线性等特点^[7]。

假设前几天的汇率已经知道, 我们能够精确的预测出接下来几天的汇率, 这里用 2006 年 2 月份到 2006 年 11 月份的汇率率做为训练集, 表 1 只列出了部分样本。样本集来自于中国人民银行网站 <http://www.pbc.gov.cn/huobizhengce>。

表 1 人民币汇率中间价图表 (2006 年)

日期	美国	日期	美国	日期	美国
11-16	7.8733	11-2	7.8750	10-19	7.9090
11-15	7.8715	11-1	7.8720	10-18	7.9085
11-14	7.8703	10-31	7.8792	10-17	7.9110
11-13	7.8644	10-30	7.8781	10-16	7.9148
11-10	7.8667	10-27	7.8871	10-13	7.9116
11-9	7.8697	10-26	7.8940	10-12	7.9174
11-8	7.8719	10-25	7.9007	10-11	7.9164
11-7	7.8829	10-24	7.9049	10-10	7.9128
11-6	7.8804	10-23	7.8998	10-9	7.9103
11-3	7.8729	10-20	7.8995	9-29	7.9087

如果已经知道前 10 天的汇率, 可以预测出接下来 5 天的汇率。表 2 是将双自组织映射网络和改进后的网络用于预测后得出的平均绝对百分误差值的比较。从表中可以看出, 将新网络用于存在噪声的汇率率数据集中, 预测的精度较高, 说明新网络具有很好的抗噪声能力。

表 2 双自组织映射网络与原 SOM 网络预测结果比较表

方法	第 1 天	第 2 天	第 3 天	第 4 天	第 5 天
SOM	0.3305	0.3322	0.3270	0.3652	0.3241
改进后的双自组织映射网络	0.3169	0.2304	0.3107	0.3454	0.3376

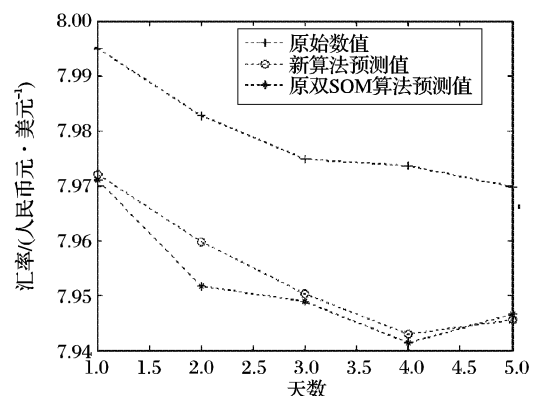


图 2 新网络与原 SOM 预测性能比较

图 2 是新网络与原 SOM 预测性能对比, 从图中也可以看

出,改进后的网络较原先的网络具有更好的鲁棒性。

实验2 噪声数据的实验。SOM神经网络不仅具有分类能力,同时还具有很强的自联想功能。SOM网络的自联想功能可以描述如下: 设含 n 维输入节点, $m \times m$ 个输出神经元的SOM网络已经通过样本进行训练。如果输入向量

$$X = (x_1, x_2, \dots, x_{n-m-1}, \dots, x_n)$$

已知 $n-m$ 维分量,还有 m 维分量未知。欲确定未知分量,先用 $n-m$ 维已知分量作为网络输入,按式(13)确定对应的优胜者 C ,优胜者与未知分量节点间的权值即为该未知分量。

$$d_c = \min[d_{Eucli}]$$

$$\min \left[\sqrt{\sum_{i=1}^{n-m} (x_i - y_i)^2} \right] \quad (14)$$

本文中使用的双自组织特征映射网络,同样也可以实现这样的功能。本实验即是利用双自组织特征映射网络来实现类似功能的预测实验。

实验中使用90个学生的6门模拟考试分数作为样本集,并对数据集添加均值为0,方差为0.1的高斯噪声,考试结果被分成两个数据集,其中的3门考试是闭卷考试,另外的3门是开卷考试,把这6门成绩分成2个数据集,由于学生的闭卷考试能力和开卷考试能力有相关性,从实验可以看出,新的网络具有很好的预测性能。

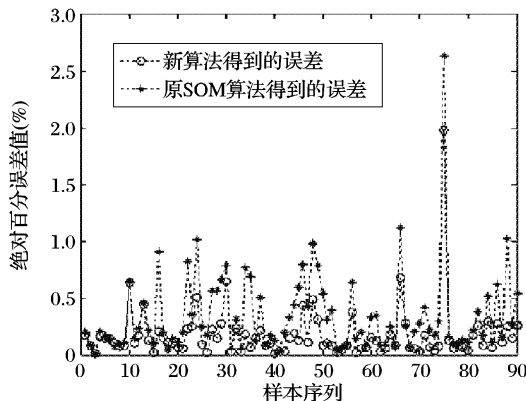


图3 原始考试成绩预测误差比较

图3是数据集没有加噪声前,分别使用双SOM和改进后的网络,通过90个学生的闭卷考试成绩来预测他们的开卷考试能力所得到的平均绝对百分误差值的对比,最后计算出改进前网络的误差的平均值为0.3445,改进后网络的误差平均值为0.1762。从图中可以看出改进的网络具有更高的预测精度。

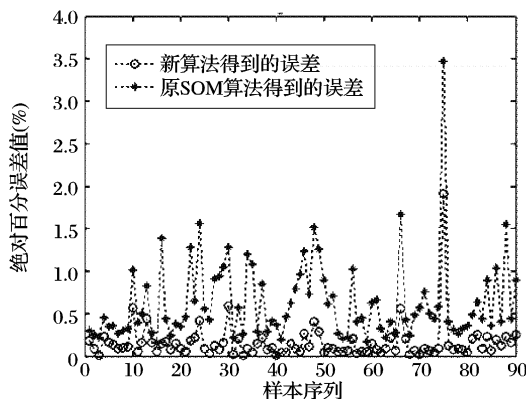


图4 考试成绩加噪后新算法和原算法预测结果比较

图4是对数据集添加噪声后,分别使用双SOM和改进后的网络,发现原网络的精确度明显降低,而新网络的预测结果

受噪声影响很小,由此可以看出新的网络抗噪声能力更强。

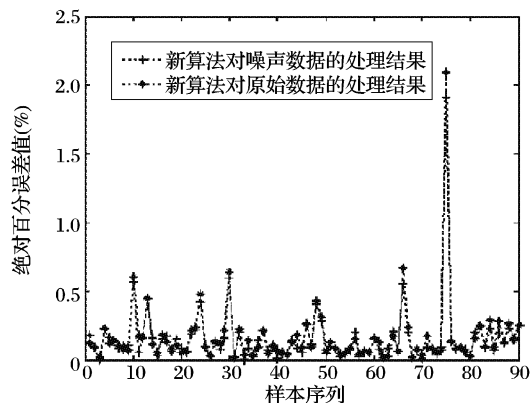


图5 新算法对数据的处理结果比较

图5是新网络用于原始数据和噪声数据的预测结果比较图,实验结果表明,两种情况下的预测精度几乎相等,更进一步证明了改进后的双自组织特征网络具有较强的鲁棒性。

5 结语

本文对双自组织特征映射网络进行改进,将本来用于单个数据集的算法,用到成对的两个数据集中,并且使用Voronoi cell距离替代欧式距离,这样使得双自组织特征映射网络具有更强的鲁棒性。将改进后的双自组织映射网络用于对金融数据的预测实验,发现改进的网络较原始的网络具有更强的抗噪声能力。最后,考虑到SOM的任务不仅是可以用来预测,它还有一个主要的功能是对高维数据集进行可视化。因此,未来研究的另一个主题就是将本文的思想用于可视化^[8]等领域。

参考文献:

- [1] KOHONEN T. Self-organizing maps[M]. Springer, Berlin, 1995.
- [2] GOU Z, FYFE C. A family of networks which perform canonical correlation analysis[J]. International Journal of Knowledge-based Intelligent Engineering Systems, 2001, 5(2): 76-82.
- [3] HAN Y, CORCHADO E, FYFE C. Forecasting using twinned principal curves and twinned self-organising maps[J]. Neurocomputing, 2004, (57): 37-47.
- [4] 谭晓阳. 单训练样本条件下基于自组织神经网络的鲁棒人脸识别技术研究[D]. 南京: 南京大学, 2005.
- [5] HOPNER F, KLAUONN F. Improved fuzzy partitions for fuzzy regression models[J]. International Journal of Approximate Reasoning, 200, 32(2): 85-102.
- [6] 阎平凡, 张长水. 人工神经网络与模拟进化计算[M]. 北京: 清华大学出版社, 2000.
- [7] GILES CL, LAWRENCE S, TSOI AC. Noisy time series prediction using recurrent neural networks and grammatical inference[J]. Machine Learning, 2001, 44(1-2): 161-183.
- [8] FLEXER A. On the use of self-organizing maps for clustering and visualization[J]. Intelligent Data Analysis, 2001, 5(5): 373-384.
- [9] MULIER F, CHERKASSKY V. Self-organisation as an iterative kernel smoothing process[J]. Neural Computer, 1995, 6(6): 1165-1177.
- [10] HAN Y, FYFE C. Finding underlying factors in time series[J]. Cybernetics and Systems: An International Journal, 2002, 33(297): 297-323.
- [11] LENDASSE A, VERLEYSEN M, DE BODT E, et al. Forecasting Time-Series by Kohonen Classification[A]. ESANN'1998 proceedings[C]. 1998.