

# 改进型平均移位柱状图估算概率密度并对互信息作相关分析

韩 敏, 梁志平

(大连理工大学 电子信息与电气工程学部, 辽宁 大连 116023)

**摘要:** 将平均移位柱状图(averaged shifted histogram, ASH)概率密度估计中二次型平滑权值与均匀权值进行结合, 提出一种改进的概率密度估计方法: IASH(improved averaged shifted histogram). 通过相应区间内样本数目的方差确定原平滑权值与均匀权值之间的比例系数, 动态的改变平滑权值; 对ASH概率密度估计中边缘值部分的平滑权值按比例进行补偿, 改善过平滑的问题, 提高了IASH概率密度估计的精度. 在此基础上应用互信息进行变量间的相关性分析, 选择输入变量, 实现多元时间序列的预测. 采用人工数据和实际Housing数据进行仿真分析, 仿真结果验证了改进后方法的有效性.

**关键词:** 平均移位柱状图; 互信息; 相关性分析; 时间序列预测

**中图分类号:** TP138      **文献标识码:** A

## Correlation analysis of mutual information by probability density estimated from improved averaged-shifted-histogram

HAN Min, LIANG Zhi-ping

(Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian Liaoning 116023, China)

**Abstract:** We introduce the method of improved averaged-shifted-histogram(IASH) to estimate the probability density by combining the quadratic smooth weight with the uniform smooth weight. The ratio of the original smooth weight to the uniform smooth weight is dynamically adjusted according to the variance of the number of samples in the corresponding interval, thus the smooth weight for the edge part of the probability density obtained by the method of averaged-shifted-histogram(ASH) is proportionally compensated, mitigating the excessive smoothness and improving the precision in the estimation of probability density by the method of IASH. Using the estimated probability density, we perform the correlation analysis based on the mutual information between two variables, and select input variables to predict the multivariate time series. Simulations with the synthetic data and Housing data show the efficacy of the proposed method.

**Key words:** averaged-shifted-histogram; mutual information; correlation analysis; time series prediction

### 1 引言(Introduction)

多元时间序列预测已经在天气预报、经济预测、股市预测等方面得到了广泛的应用. 随着研究的深入人们发现, 预测的精度与输入变量的选择有很大关系. 对变量间的相关关系研究的主要内容正是分析多元时间序列中的相关性. 根据变量间的相关性, 选取合适的变量作为模型的输入, 对建立精确的预测模型具有重要的意义.

在多变量相关分析方面, 目前主要研究方法有统计学分析方法(主成分分析<sup>[1,2]</sup>、典型相关分析<sup>[3]</sup>、因子分析<sup>[4]</sup>等)、Granger因果分析法<sup>[5]</sup>、互信息分析法<sup>[6,7]</sup>等. 互信息的大小不仅能反映变量间的线性关系, 而且能够表征变量间的非线性关系, 因此互信息用于相关性分析取得了较为广泛的应用. 互信息的

计算需要估计数据的概率密度分布, 目前常用的概率密度估计方法有柱状图概率密度估计、核概率密度估计等. Scott<sup>[8]</sup>提出的平均移位柱状图(averaged shifted histogram, ASH)概率密度估计是一种介于核概率密度估计和柱状图概率密度估计之间的方法. 该方法较好的结合了柱状图方法计算速度快和核方法精度高的优点, 能够较好的对于样本数据点进行概率密度估计, 可有效的用于计算变量间的互信息, 并进行相关性分析.

但ASH概率密度估计对参数的选择比较敏感, 不适当的参数选择会使概率密度的估计值产生很大偏差, 而且由于ASH概率密度估计中平滑权值始终固定不变, 容易将非高斯型变量的概率密度估计倾向于高斯型; 在数据边缘值部分ASH概率密度估计会

产生过度平滑的现象,存在较大偏差,影响了概率密度估计的精度,从而相应的互信息用于相关性分析的结果会产生偏差.

针对ASH概率密度估计存在的问题,本文提出一种改进的IASH(improved averaged shifted histogram)概率密度估计方法.通过均匀权值与ASH密度估计中的Biweight权值相结合,并对过平滑部分对应的平滑权值进行相应的补偿,提高过平滑部分的估计值,从而提高了基于IASH概率密度估计的精度.最后利用改进后的IASH方法估计概率密度用于互信息的计算,并进行相关性分析,选择输入变量并进行预测,取得了较好的预测效果.

## 2 IASH概率密度估计(IASH probability density estimation)

ASH是一种较好的非参数概率密度估计方法,该方法将 $m$ 个估计值进行平滑处理得到概率密度估计值.但由于平滑权值的作用使得估计值产生了偏差,本文将从平滑权值和过平滑两方面进行改进,提高概率密度估计的精度.

### 2.1 ASH概率密度估计(ASH probability density estimation)

作为一种介于传统的柱状图概率密度估计和核概率密度估计之间的方法,ASH较好的结合了两者的优势,Fernando等<sup>[9]</sup>通过仿真验证了ASH方法在保证估计精度的同时,计算速度要优于核概率密度估计的方法.

对于一组序列 $\{x_1, \dots, x_N\}$ ( $N$ 为样本总数)来自分布 $f(x)$ ,ASH通过将每 $m$ ( $m$ 为区间划分参数)个变化的子区间上的估计值进行平滑作为 $f(x)$ 的估计值.将每个长度为 $h$ 的区间进行划分(一般称 $h$ 为带宽参数),令 $\delta = h/m$ , $t_k = k\delta$ ,得子区间 $I_k = [t_{k-1}, t_k]$ , $v_k$ 为在每个小区间内的样本点的个数,则每个区间上的概率分布为式(1),

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1}^{m-1} w_m(i) v_{k+i}, \quad (1)$$

其中 $w_m$ 为权重函数,通过式(2)进行计算.

$$w_m(i) = \frac{mK(i/m)}{\sum_{j=1}^{m-1} K(j/m)}. \quad (2)$$

$K(t)$ 为平滑核函数,可选取核函数中的三角形核函数,二次型权值核函数(biweight kernel)或三次型权值核函数(triweight kernel).本文选用二次型核函数,如式(3)所示.

$$K(t) = \frac{15}{16}(1-t^2)^2, \quad -1 \leq t \leq 1. \quad (3)$$

当变量扩展为二维时,则将二维变量对应的区间分别划分为 $\delta_x = h_x/m_x$ 和 $\delta_y = h_y/m_y$ ,二维数据的概率估计变为

$$\hat{f}(x, y) = \frac{1}{N\delta_x\delta_y} \sum_{i=1}^{m_x-1} \sum_{j=1}^{m_y-1} w_{m_x}(i)w_{m_y}(j)v_{k+i,l+j}. \quad (4)$$

### 2.2 平滑权值的改进(Improved smooth weight)

通过ASH的估计式(1)可以看出,在对每 $m$ 个概率密度估计值进行平滑处理时,平滑权值是固定的权值函数,即进行ASH概率密度估计时,对不同的样本分布,在参数 $m$ 固定的情况下,平滑计算的权值是始终不变的.这样会造成无论样本点为何种分布,通过该权值函数进行平滑后的结果都近似趋近于高斯型概率分布.当样本点的类型为其他分布时,由于平滑过程趋近于高斯型分布,就有可能使得互信息的分析结果产生偏差.

为简化问题,以一般的均匀分布为实际样本点的概率密度进行分析.当样本点为 $[0,1]$ 间的均匀分布时,样本落入每一小区间的概率为区间的大小 $\delta = h/m$ ,设一般情况下落入相应区间内的样本点的数目为 $v_{k+i} = N(\delta + \Delta_{k+i})$ ,满足 $\sum \Delta_{k+i} = 0$ ,因此可得

$$\begin{aligned} \hat{f}(x) &= \frac{1}{Nh} \sum_{i=1}^{m-1} w_m(i) v_{k+i} = \\ &= \frac{1}{Nh} \sum_{i=1}^{m-1} w_m(i) (N\delta + N\Delta_{k+i}) = \\ &= \frac{1}{m} \sum_{i=1}^{m-1} w_m(i) + \sum_{i=1}^{m-1} w_m(i) \frac{\Delta_{k+i}}{h} = \\ &= 1 + \sum_{i=1}^{m-1} w_m(i) \frac{\Delta_{k+i}}{h}. \end{aligned} \quad (5)$$

可以看出,平滑后的概率密度估计值产生了微小的偏差.为解决此问题,引入可变参数 $\alpha$ 将原权值与均匀权值相结合,一方面保证权值的平滑特性,另一方面又通过参数 $\alpha$ 将平滑权值平均化,使平滑后的估计值更接近于样本点的实际分布.通过调整 $\alpha$ 的大小动态的改变平滑权值的系数.改进后的权值如式(6)所示.

$$w_m(i) = \alpha \frac{mK(i/m)}{\sum_{j=1}^{m-1} K(j/m)} + (1-\alpha) \frac{m}{2m+1}. \quad (6)$$

参数 $\alpha$ 的值的大小通过相对应的区间的方差决定.设每个相应区间 $I_k = [t_{k-1}, t_k]$ 及邻近的 $2m+1$ 个子区间的样本数目 $v_l (k-m \leq l \leq k+m)$ 的方差为 $\sigma(k)$ ,则对应的向量 $\alpha$ 由式(7)确定.

$$\alpha = \frac{\sigma - \min(\sigma)}{\max(\sigma) - \min(\sigma)}. \quad (7)$$

方差越大说明数据离散程度较高,则相应的原二次型的权值比例较大,方差越小说明数据的离散程度较低,分布较均匀,则均匀型的平滑权值占的比例较大.这样通过动态的改变平滑权值,使估计的结果更接近于实际的概率密度分布,有效的提高了概率密度估计的精度.

### 2.3 过平滑的改进(Improved over-smoothed weight)

由ASH的估计公式(1),在样本数据的边缘值部分,平滑会使该子区间的概率密度估计值产生偏差.设在一般情况下,样本点落入每一子区间的概率为 $p_{k+i}$ ( $\sum p_{k+i} = 1$ ),则相应的子区间的样本点的数目为 $v_{k+i} = N \cdot p_{k+i}$ ,所以可得

$$\hat{f}(x) = \frac{1}{Nh} \sum_{i=1-m}^{m-1} w_m(i)v_{k+i} = \frac{1}{Nh} \sum_{i=1-m}^{m-1} w_m(i)Np_{k+i} = \frac{1}{h} \sum_{i=1-m}^{m-1} w_m(i)p_{k+i}. \quad (8)$$

当区间在样本数据值的边缘时,存在 $v_{k+i}$ 等于0的情况,这样原估计式变为

$$\begin{aligned} \hat{f}(x) &= \frac{1}{Nh} \sum_{i=1-m}^{m-1} w_m(i)v_{k+i} = \\ &= \frac{1}{Nh} \sum_{i=1-m}^{m-1} w_m(i)Np_{k+i} = \\ &= \frac{1}{Nh} \sum_{i=1-m}^t w_m(i) \cdot 0 + \frac{1}{Nh} \sum_{i=t}^{m-1} w_m(i)Np_{k+i} = \\ &= \frac{1}{Nh} \sum_{i=t}^{m-1} w_m(i)Np_{k+i} < f(x). \end{aligned} \quad (9)$$

在该子区间的概率密度估计值低于实际值,为改进此问题,将 $v_{k+i}$ 为0部分对应的权值按比例补偿到非0部分对应的平滑权值上.设全部子区间数为 $n_{\text{bin}}$ ,当 $k < m$ 时共有 $t$ 个 $v_{k+i}$ 为0.调整后的平滑权值为式(10),其中 $L = 1 - m + t$ .

$$w_m(i) = \begin{cases} 0, & i \leq L, \\ w_m + \frac{w_m(i)}{\sum_{l=1}^{m-1} w_m(l)} \sum_{l=L}^0 w_m(i)(l), & i > L. \end{cases} \quad (10)$$

调整后的平滑权值同时满足式(6)的条件,避免了将一般分布估计成类高斯分布的情况,同时在边缘值部分对估计值进行相应的补偿,提高了IASH概率密度估计的精度.

ASH概率密度估计对参数 $m$ 和 $h$ 的比较敏感,为提高IASH用于概率密度估计的精度,本文采用Scott提出的UCV(unbiased cross-validation)<sup>[10]</sup>作为评价准则实现参数 $m$ 和 $h$ 的选择.

## 3 互信息估计及变量选择(Estimation of mutual information and variable selection)

### 3.1 互信息估计(Estimation of mutual information)

互信息理论来源于信息论中熵的概念.熵是对实验结果不确定性的良好度量.如果事件是完全确定的,那么熵值为零,换言之,如果对实验结果的输出是确定的,那么事件发生的概率就很大,而熵值就会很小<sup>[11]</sup>.熵的基本定义为式(11).

$$H(X) = - \sum_{i=1}^N P(x_i) \log_2 P(x_i). \quad (11)$$

由熵的概念得到互信息的计算公式为式(12).互信息的大小反映了两变量间的统计依赖程度,变量间的互信息值的大小,说明了变量间的相关性的大小.

$$I(X, Y) = H(X) + H(Y) - H(X, Y). \quad (12)$$

根据选择好的IASH概率密度估计的参数,对数据进行概率密度估计,然后计算变量间的互信息.这样利用二维IASH概率密度估计进行互信息计算分析的步骤如下:

- 1) 分别将样本点按照所选的参数进行网格划分,统计落入每个网格内的样本点数目;
- 2) 分别计算各变量在 $I_k$ 相对应区间内样本数目的方差,作为调整权值的参数 $\alpha$ ,并分别计算二次型权值和均匀权值,根据公式(6)计算调整后的平滑权值;
- 3) 根据公式(10)调整相应的平滑权值;
- 4) 由新的平滑权值分别计算相应的概率密度分布;
- 5) 利用得到的概率密度值计算二维变量的互信息.

这样可以通过IASH计算各输入变量与输出变量间的互信息值,并分析变量间的相关关系,从而实现模型的变量选择.

### 3.2 变量选择(Variable selection)

变量间的互信息值越大,说明相关性越好.通过互信息相关分析,可以有效的进行输入变量选择.本文采取逐步回归向前法实现变量选择.设 $\varphi$ 为输入变量集合, $M$ 为输入变量个数,基于互信息的变量选择过程描述如下:

- 1) 利用IASH概率密度估计方法计算各输入变量与输出变量之间的互信息大小;
- 2) 根据互信息值的大小删除互信息极小的变量,并设定初始变量集合 $\varphi$ 为选择的第一个变量

$$\varphi = X_{s_1} = \arg \max_{X_j} \{I(X_j, Y)\}, 1 \leq j \leq M; \quad (13)$$

3) 根据预测模型计算预测误差 $\text{Error}(\varphi)$ ;

4) 选择其余的输入变量 $X_{s_2}$ , 结合输入变量集合 $\varphi$ 建立预测模型, 计算预测误差 $\text{Error}(\{\varphi \cup X_{s_2}\})$ ,

$$X_{s_2} = \arg \max_{X_j} \{I(X_j, Y)\}, 1 \leq j \leq M, j \neq s_1; \quad (14)$$

5) 若 $\text{Error}(\{\varphi \cup X_{s_2}\}) < \text{Error}(\varphi)$ , 则选择该变量 $\varphi = \{\varphi \cup X_{s_2}\}$ , 否则输入变量集合 $\varphi$ 保持不变;

6) 重复步骤3)~5), 直至选择完所有输入变量.

根据上述变量选择方法, 可以较好的实现输入变量的选择, 结合模型预测误差反映不相关变量和冗余自变量的引入对预测模型的影响, 在保证预测精度的同时, 也一定程度上避免了冗余输入变量的引入, 从而简化了预测模型, 提高了预测精度.

### 4 仿真结果与分析(Simulation results and analysis)

#### 4.1 人工数据的概率密度估计(Probability density estimation of synthetic data)

为验证IASH用于概率密度估计的改进效果, 仿真对(0,1)间Arcsin型分布的数据进行概率密度估计. 该分布的密度分布形状为U形, 即在两边缘处的概率密度值较高, 中间部分的密度值较低, 密度函数如式(15)所示. 仿真将IASH及ASH方法的概率密度估计结果同实际的概率密度值比较, 仿真结果如图1所示.

$$p(x) = \frac{1}{\pi\sqrt{x(1-x)}}, 0 < x < 1. \quad (15)$$

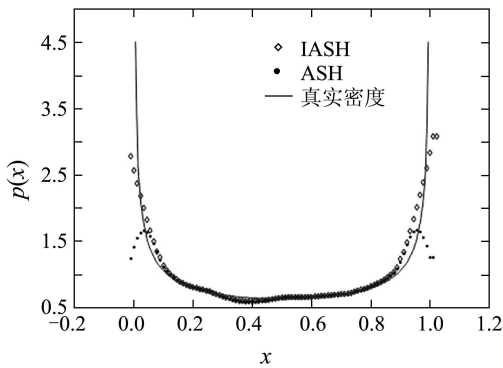


图1 Arcsin型分布的概率密度估计结果  
Fig. 1 Density estimation of Arcsin distribution

从仿真结果可以看出在数据点的边缘值部分, ASH方法的估计值开始呈现出下降的趋势, 这与式(9)的推导结果是一致的, 而改进后的IASH密度估计值与实际的密度值的差异较小, 能够较好的反映实际的概率密度估计结果, 说明了IASH方法的有效性.

为说明所提对其他数据的适用性, 采用高斯与拉普拉斯一维混合密度进行仿真, 实际概率密度的分布模型如式(16)所示. 训练样本数为1000, 测试样本数为2000. 采用 $L_1$ 测试误差进行比较仿真结果.

$$p(x) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(x-2)^2}{2}} + \frac{0.7}{4} e^{-0.7|x+2|}, \quad (16)$$

$$L_1 = \frac{1}{N_{\text{test}}} \sum_{k=1}^{N_{\text{test}}} |p(x_k) - \hat{p}(x_k)|. \quad (17)$$

图2为一次仿真的密度估计结果, 从图中可以看出所提方法能够很好的反映真实密度分布的情况, 而且在边缘处能够取得较好的估计结果. 表1比较了改进ASH密度估计方法, ASH密度估计及经典kernel密度估计方法, 比较项目为100次重复实验得到的 $L_1$ 测试误差对应的均值与方差. 从表中可以看出所提方法在精度方面要略优于原ASH密度估计方法及经典的kernel密度估计方法.

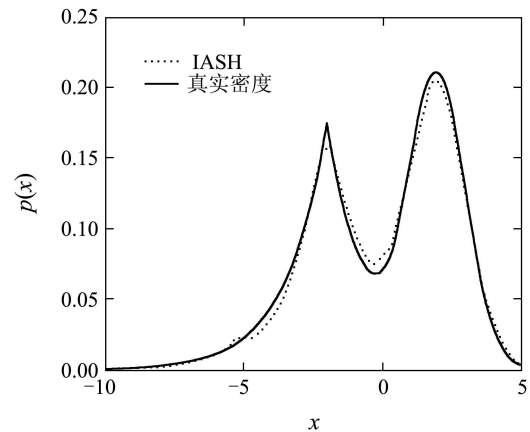


图2 高斯与拉普拉斯混合分布的概率密度估计结果  
Fig. 2 Density estimation of Gaussian Laplacian mixer

表1  $L_1$ 测试误差比较

Table 1 Results of $L_1$ test error	
方法	$L_1$ 测试误差
IASH	$(5.3751 \pm 1.1480) \times 10^{-3}$
ASH	$(5.7540 \pm 1.0826) \times 10^{-3}$
kernel	$(5.5322 \pm 1.1124) \times 10^{-3}$

#### 4.2 人工数据的互信息相关性分析(Mutual information based correlation analysis of synthetic data)

为分析改进后的IASH概率密度估计的互信息用于相关性分析的性能, 采用一组人工构造的数据进行输入变量与输出变量间互信息的分析. 数据的模型为10个输入变量 $X_i$ , 一个输出变量 $Y$ , 变量间的关系为

$$Y = 10 \sin(X_1 X_2) + 20 X_3 (X_3 - 0.5) + 10 X_4 + 5 X_5 + \varepsilon, \quad (18)$$

变量 $X_i (1 \leq i \leq 10)$ 为[0,1]间的均匀分布,  $\varepsilon$ 为方差为1的高斯噪声, 变量 $X_6$ 至 $X_{10}$ 是无关变量, 数据点的数目为1000组. IASH和ASH概率密度估计的互信息计算结果如图3所示.

可以看出, 利用ASH概率密度估计得到的互信息值各个变量间的差异不大, 不能有效的反映输入变量与输出变量间的相关性; 而IASH概率密度估计的互信息结果中, 前5个变量与输出变量间的互信息值较大, 有效的反映了输入变量与输出变量间的相关性. 在变量 $X_3, X_4$ 处的互信息值最大, 这与式(18)中的相关性是一致的. 仿真结果说明本文提出的IASH方法估计互信息能够有效的分析变量间的相关关系, 从而可以较好的选择预测模型的输入变量.

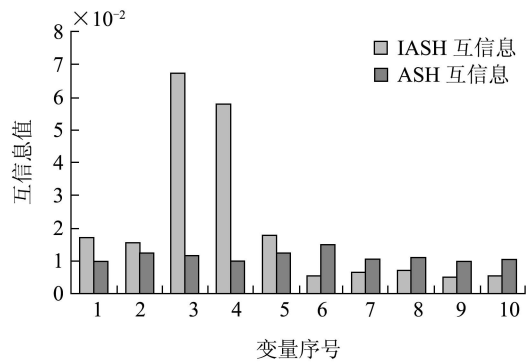


图 3 人工数据的互信息仿真结果

Fig. 3 Estimation of mutual information with synthetic data

### 4.3 实际数据互信息分析及预测(Mutual information analysis and prediction)

本文以公共数据Housing data进行实际的仿真分析<sup>[12]</sup>. 通过每个房子周围代表13个属性的统计值(定义为输入变量 $X_1, \dots, X_{13}$ )来预测每栋房子的房价(定义为输出变量 $Y$ ). 预测模型采用广义回归神经网络(GRNN), GRNN在结构上与RBF网络较为相似, 不同之处在于GRNN不需要设定模型的具体形式, 且对所有隐层单元的核函数采用同一的光滑因子, 训练极为方便. 在训练样本和光滑因子固定的情况下, 网络的结构和收敛性均保持不变, 关于GRNN的详细介绍见参考文献[13]. 在本文中, 设定仿真实验的光滑因子为0.8326. Housing数据共506组, 随机选取其中338组作为训练数据, 其余168组数据作为测试数据, 仿真重复进行50次.

为分析数值试验结果的优劣, 采用均方根误差 $E_{RMSE}$ 作为性能指标, 对预测结果进行比较, 定量说明各模型的预测效果. 设 $y(k)$ 为预测变量的实际观测值,  $\hat{y}(t)$ 为该变量的网络预测值,  $N_{test}$ 为对应测试样本的数目, 则均方根误差定义为

$$E_{RMSE} = \left( \frac{1}{N_{test} - 1} \sum_{k=1}^{N_{test}} (\hat{y}(k) - y(k))^2 \right)^{\frac{1}{2}}. \quad (19)$$

均方根误差反映了预测值相对于观测值的偏离程度, 取值大于或等于零, 预测无误差时等于零. 均方根误差值越小, 表明预测效果越好.

利用IASH概率密度估计进行互信息的分析. 分别计算每个变量 $X_i (1 \leq i \leq 13)$ 与输出变量 $Y$ 间的互信息值, 互信息的计算结果如图4所示. 可以看出与输出变量互信息值较大的变量为 $X_{13}$ , 说明与输出变量之间的相关性较强, 变量 $X_4, X_5$ 与输出变量间互信息的大小接近0, 说明 $X_4, X_5$ 与输出变量间几乎不存在任何相关关系, 可将其视为无关变量, 结合模型预测进行变量选择. 采用均方根误差 $E_{RMSE}$ 作为性能指标, 对所选变量的预测结果进行比较, 分析结果为50次运行得到的均值与方差如表2所示.

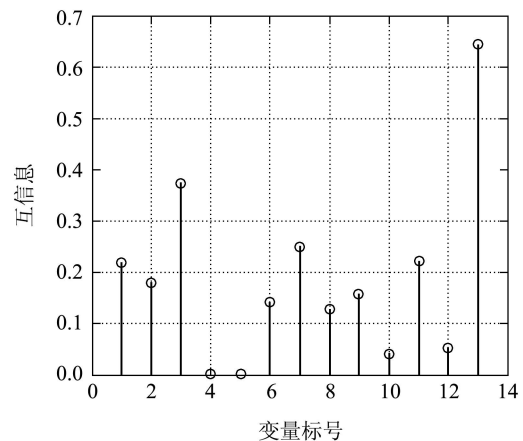


图 4 IASH概率密度估计的互信息结果

Fig. 4 Estimation of mutual information based on IASH

表 2 预测性能指标

Table 2 Results of  $E_{RMSE}$

	输入变量	$E_{RMSE}$
全部变量	$X_1, X_2, \dots, X_{13}$	$16.1317 \pm 1.1480$
文献[12]	$X_6, X_{13}, X_1, X_4$	$5.7966 \pm 0.4806$
IASH互信息	$X_{13}, X_3$	$5.2634 \pm 0.2754$

由仿真结果可以看出, 通过对输入与输出变量进行相关性分析, 并选择相关性较大的变量作为预测模型的输入, 可以有效提高模型的预测精度. IASH概率密度估计的互信息分析方法所选择变量的预测精度要优于FRANCOISA等<sup>[12]</sup>所选变量的预测精度. 而得到的最优变量也恰好为IASH互信息最大的两个, 说明了本文提出的IASH用于互信息分析的有效性.

### 5 结论(Conclusions)

针对ASH概率密度估计方法平滑权值始终固定的问题, 本文采用将原二次型平滑权值与均匀权值相结合的方法进行改进. 通过分析区间上数据的离散程度, 确定原平滑权值与均匀权值之间的比例系数, 来动态的改变平滑权值, 准确的反映样本数据的实际特性. 另一方面, 对于ASH概率密度估计中边缘部分存在过平滑的问题, 本文提出将相应的平

滑权值进行补偿, 来提高概率密度估计的精度. 通过IASH概率密度估计进行互信息的计算, 并选择相应的输入变量进行预测. 利用人工数据和Housing数据进行仿真分析, 实际仿真的结果验证了本文所提方法的有效性.

### 参考文献(References):

- [1] RADAN H, LUCIE P. Simultaneous analysis of climatic trends in multiple variables: an example of application of multivariate statistical methods[J]. *International Journal of Climatology*, 2005, 25(4): 469 – 484.
- [2] 史文利, 高天宝, 王树恩. 基于主成分分析与聚类分析的城市化水平综合评价[J]. *工业工程*, 2008, 11(3): 112 – 115.  
(SHI Wenli, GAO Tianbao, WANG Shuen. Evaluation of urbanization level using principal component analysis and cluster analysis[J]. *Industrial Engineering Journal*, 2008, 11(3): 112 – 115.)
- [3] HIROYUKI Y, HIDEKE Y, EIICHIRO F, et al. Canonical correlation analysis for multivariate regression and its application to metabolic fingerprinting[J]. *Biochemical Engineering Journal*, 2008, 40(2): 199 – 204.
- [4] LOPEZ J M, BORRAJO J L, GARCIA E D M, et al. Multivariate analysis of contamination in the mining district of Linares[J]. *Applied Geochemistry*, 2008, 23(8): 2324 – 2336.
- [5] CHEN Y H, RANGARNJAN G, FENG J F, et al. Analyzing multiple nonlinear time series with extended Granger causality[J]. *Physics Letters A*, 2004, 324(1): 26 – 35.
- [6] 高伟, 田铮. 基于条件互信息的多维时间序列图模型[J]. *控制理论与应用*, 2008, 25(2): 257 – 260, 267.  
(GAO Wei, TIAN Zheng. Graphical models for multivariate time series based on conditional mutual information[J]. *Control Theory & Applications*, 2008, 25(2): 257 – 260, 267.)
- [7] 张佃中. 非线性时间序列互信息与Lempel-Ziv复杂度的相关性研究[J]. *物理学报*, 2007, 56(6): 3152 – 3157.  
(ZHANG Dianzhong. Research on the correlation between the mutual information and Lempel-Ziv complexity of nonlinear time series[J]. *Acta Physica Sinica*, 2007, 56(6): 3152 – 3157.)
- [8] SCOTT D W. Averaged shifted histograms: effective nonparametric estimators in several dimensions[J]. *The Annals of Statistics*, 1985, 13(3): 1024 – 1040.
- [9] FERNADO T M K G, MAIER H R, DANDY G C. Selection of input variables for data driven models: an average shifted histogram partial mutual information estimator approach[J]. *Journal of Hydrology*, 2009, 367(3/4): 165 – 176.
- [10] SCOTT D W, TERRELL G R. Biased and unbiased cross-validation in density estimation[J]. *Journal of the American Statistical Association*, 1987, 82(400): 1131 – 1146.
- [11] CELLUCCI C J, ALBANO A M, RAPP P E. Statistical validation of mutual information calculations comparison of alternative numerical algorithms[J]. *Physical Review E*, 2005, 71(6): 066208.
- [12] FRANCOISA D, ROSSIB F, WERTZA V, et al. Resampling methods for parameter-free and robust feature selection with mutual information[J]. *Neurocomputing*, 2007, 70(7/9): 1276 – 1288.
- [13] SPECHT D F. A general regression neural network[J]. *IEEE Transactions on Neural Networks*, 1991, 2(6): 568 – 576.

### 作者简介:

韩敏 (1959—), 女, 教授, 博士生导师, 研究方向为神经网络、3S系统及混沌序列分析, E-mail: minhan@dlut.edu.cn;

梁志平 (1986—), 男, 硕士研究生, 研究方向为多变量时间序列相关性分析及预测, E-mail: liangzhp@mail.dlut.edu.cn.