文章编号: 1001-0920 (2012) 06-0949-04

一种基于 k- 近邻互信息变化率的输入变量选择方法

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: 针对多变量时间序列建模中的输入变量选择问题, 提出一种基于 k- 近邻互信息变化率的变量选择方法. 根据多变量之间的相关关系, 以输入输出之间的 k- 近邻互信息变化率作为评价标准选择相关变量; 同时根据输入变量子集之间互信息值的大小判断变量是否为冗余变量; 通过设定合适的阈值系数, 可以有效地实现输入变量选择. Friedman, Lorenz 混沌时间序列以及 Housing 数据的变量选择仿真结果验证了所提出方法的有效性.

关键词: k- 近邻互信息；输入变量选择；相关分析

中图分类号: TP183 **文献标识码:** A

An input variables selection method based on k-nearest neighbors mutual information

HAN Min, *LIANG Zhi-ping*

(Faculty of Electronic Information and Electrical Engineering，Dalian University of Technology，Dalian 116023，China. Correspondent：HAN Min，E-mail：minhan@dlut.edu.cn)

Abstract: An input variables selection method is proposed based on the k-nearest neighbors mutual information. According to the mutual information between the multi-variables, the relevant variables which have great influence to the mutual information are selected. Meanwhile, the redundant variables are removed according to the mutual information between the input variables sets. Consequently, the input variables are selected with proper parameter settings. The results of the simulation based on the Friedman data, the Lorenz time series and the Housing data show the effectiveness of the proposed input variables selection method.

Key words: k-nearest neighbors mutual information；input variables selection；correlation analysis

1 引 言

多变量时间序列建模与预测已在天气预报、经济预测、电力负荷预测等方面得到了广泛的应用. 传统的预测方法很少考虑输入变量之间的关系, 如果输入变量选择不当, 则有可能产生较差的预测结果, 因此多变量间的相关分析及输入变量的选择对建立精确的预测模型具有重要意义[1-2].

互信息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. 如 Fernado 等人[3]提出利用偏互信息进行相关性分析, 但实际数据往往无法满足该方法所提假设条件. Peng 等人[4]提出了 mRMR (minimal-redundancy-maximal-relevance) 进行输入变量选择; Estévez 等人[5]提出了 NMIFS (normalized mutual information feature selection), 采用标准化互信息进行变量选择. 但是上述两种方法均无法实现变量选择的自动终止, 必须人为设定选择输入变量的个数.

Kraskov 等人[6]提出了一种基于 k- 近邻的互信息计算方法, 在变量选择方面取得了较好的应用[7]. 本文在 k- 近邻互信息的基础上, 提出一种前向式输入变量选择方法, 以输入输出间 k- 近邻互信息变化率为评价标准来判断是否为相关变量, 同时根据输入变量子集之间互信息的大小判断是否为冗余变量. 利用人工数据及实际数据进行了仿真, 所得结果验证了本文方法的有效性.

2 k- 近邻互信息估计

互信息理论来源于信息论中熵的概念. 熵可以作为信息不确定性的良好度量[8], 序列 X 的信息熵定义

收稿日期: 2010-11-10；**修回日期:** 2011-01-18.

基金项目: 国家自然科学基金项目(61074096).

作者简介: 韩敏(1959−), 女, 教授, 博士生导师, 从事复杂工业系统建模与控制、智能技术及优化算法等研究；梁志平(1986−), 男, 硕士生, 从事多变量时间序列相关性分析与预测的研究.

为

$$H(X) = -\int \mathrm{d}x\mu(x)\log\mu(x), \tag{1}$$

其中 $\mu(x)$ 为 X 的概率密度分布函数.

变量 X,Y 的联合分布 $\mu(x,y)$ 与其边缘分布乘积 $\mu_x(x)\mu_y(x)$ 之间的相对熵即为互信息, 其定义为

$$MI(X,Y) = \iint \mathrm{d}x\mathrm{d}y\mu(x,y)\log\frac{\mu(x,y)}{\mu_x(x)\mu_y(y)}. \tag{2}$$

根据熵的定义, 互信息的计算可表示为

$$MI(X,Y) = H(X) + H(Y) - H(X,Y). \tag{3}$$

变量间的互信息值的大小, 说明了变量间的相关性的大小.

Kraskov 等人提出的 k-近邻互信息计算方法较为简单, 算法的基本思路为: 在 X 和 Y 构成的空间 $Z = (X,Y)$ 中, 以 $\varepsilon_i/2$ 为点 $z_i = (x_i, y_i)$ 到其 k-近邻的距离, $\varepsilon_x(i)/2$ 为点 $z_i = (x_i, y_i)$ 到 X 轴上相应点的距离, 同理可得 $\varepsilon_y(i)/2$.

统计到点 x_i 的距离严格小于 $\varepsilon_i/2$ 的数目 $n_x(i)$, 同样对变量 Y 作相同的处理得到 $n_y(i)$. 变量 X 和 Y 之间的互信息可以通过下式计算:

$$MI(X,Y) = \psi(k) - \langle\psi(n_x+1) + \psi(n_y+1)\rangle + \psi(N). \tag{4}$$

其中: $\psi(x)$ 为双 Γ 函数, 且满足 $\psi(1) = -0.577\,251\,6$, $\psi(x+1) = \psi(x) + 1/x$; 符号 $\langle\cdots\rangle$ 表示对其中的所有变量 $i = 1, 2, \cdots, N$ 取平均.

多个变量 $(X_1, X_2, \cdots, X_m)$ 之间的互信息可由式 (4) 扩展得到, 即

$$\begin{aligned} &MI(X_1, X_2, \cdots, X_m) = \\ &\psi(k) + (m-1)\psi(N) - \\ &\langle\psi(n_{x_1}) + \psi(n_{x_2}) + \cdots + \psi(n_{x_m})\rangle, \end{aligned} \tag{5}$$

其中 m 为变量个数.

3 基于 k-近邻互信息变化率的变量选择

3.1 多变量 k-近邻互信息分析

为分析变量间的相关关系, 本文从多维互信息的定义出发, 分析不同输入变量对于互信息值的影响. 设集合 S 为输入变量子集, 共包含 m 个变量 (设 S 与 $(X_1, X_2, \cdots, X_m)$ 等价), 与输出变量 Y 之间的互信息值定义如下:

$$\begin{aligned} MI(S,Y) = {} & H(X_1) + H(X_2) + \cdots + H(X_m) + \\ & H(Y) - H(X_1, X_2, \cdots, X_m, Y). \end{aligned} \tag{6}$$

若增加一个变量 X_s, 则互信息值的计算变为

$$\begin{aligned} MI(S, X_s, Y) = {} & H(X_1) + H(X_2) + \cdots + H(X_m) + \\ & H(X_s) + H(Y) - H(S, X_s, Y). \end{aligned} \tag{7}$$

两式相减, 得

$$\Delta_{MI}(S, X_s, Y) = H(X_s) + H(S,Y) - H(S, X_s, Y). \tag{8}$$

根据信息熵的理论, 信息熵代表不确定性的度量. 若增加无关变量, 则必然对不确定性的改变较少; 若 X_s 为相关变量, 则 $\Delta_{MI}(S, X_s, Y)$ 将变化较大. 因此, 可以用 $\Delta_{MI}(S, X_s, Y)$ 的变化幅度来衡量新增加的变量是否为相关变量. 定义 $\Delta_{MI}(S, X_s, Y)$ 的变化幅度通过如下互信息变化率进行衡量:

$$R_{\Delta_{MI}(S,X_s,Y)} = \frac{\Delta_{MI}(S, X_s, Y)}{MI(S,Y)}. \tag{9}$$

设定阈值 α, 若 $R_{\Delta_{MI}(S,X_s,Y)} > \alpha$, 则说明在变量子集 S 中增加该变量对原互信息值的影响较大, 即与输出变量之间的相关性较大, 该变量为相关变量.

除考虑输入输出变量之间的相关关系外, 同时需要判断增加的变量是否为冗余变量. 设定阈值 β, 判断输入变量集 S 与 X_s 之间的互信息, 若满足 $MI(S, X_s) < \beta$ 且 $R_{\Delta_{MI}(S,X_s,Y)} > \alpha$, 则说明该变量为有效相关变量, 可将该变量选为输入变量.

3.2 前向式变量选择过程

变量选择的过程采用前向式变量选择. 首先要确定变量集合中的第 1 个变量为与输出变量互信息值最大的变量, 设 M 为候选输入变量的个数, 有

$$X_{s_1} = \arg\max_{X_j}\{MI(X_j, Y)\},\ 1 \leqslant j \leqslant M. \tag{10}$$

其余的输入变量则根据互信息变化率及变量间的冗余特性进行选择.

综上所述, 基于 k-近邻的互信息变化率的变量选择方法的算法 (kNN_MI_VS) 设计步骤如下 (设 S 为输入变量集合, M 为输入变量个数):

1) 计算各输入变量与输出变量之间的互信息 $MI(X_i, Y)$.

2) 将互信息值进行排序, 设定初始变量集合 S 为选择的第 1 个输入变量 X_{s_1}.

3) 根据互信息的排序结果, 选择其余的输入变量 X_{s_2}, 计算互信息变化率 $R_{\Delta_{MI}(S,X_{s_2},Y)}$.

4) 若满足 $R_{\Delta_{MI}(S,X_{s_2},Y)} > \alpha$, 且 $MI(S, X_{s_2}) < \beta$, 则选择该变量 $S = \{S \bigcup X_{s_2}\}$; 否则, 集合 S 保持不变.

5) 重复步骤 3) 和 4), 直至选择完所有输入变量.

算法中参数 α, β 为人为设定的变量选择的阈值. 算法过程中步骤 3) 考虑输入变量与输出变量之间相关性最大, 步骤 4) 考虑使各输入变量之间冗余最小, 通过该算法能较好地进行变量选择.

4 仿真实例

为表明本文方法的有效性, 利用公用 Friedman 数据集, Lorenz 序列以及 Housing 数据进行变量选择

的仿真分析.

4.1 Friedman 数据变量选择结果分析

Friedman 的数学模型为

$$Y = 10\sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon. \tag{11}$$

其中: 变量 $X_i\,(1 \leqslant i \leqslant 10)$ 为 $[0,1]$ 间的均匀分布, ε 为方差是 1 的高斯噪声, 变量 $X_6 \sim X_{10}$ 是无关变量. 仿真选取的样本数目为 500.

仿真并与 mRMR 和 NMIFS 算法进行比较. 设定 mRMR 和 NMIFS 选择变量子集的个数为 5, 得到的变量选择结果如表 1 所示. 可以看出, 所提出方法能够有效选择全部输入变量, 而 mRMR 仅能选择全部有效输入变量中的 4 个.

表 1　Friedman 数据变量选择结果(无冗余变量)

变量选择方法	变量选择结果
mRMR	X_4, X_2, X_1, X_5, X_6
NMIFS	X_4, X_2, X_1, X_3, X_5
kNN_MI_VS	X_4, X_2, X_1, X_3, X_5

在上述仿真中并未考虑输入变量中包含冗余变量的情况, 因此在原 Friedman 数据集中增加变量 X_{11}, 并令 $X_{11} = 0.5X_1$. 这样变量 X_{11} 与变量 X_1 具有很强的相关关系, 为冗余变量, 在相同条件下进行仿真, 结果如表 2 所示.

表 2　Friedman 数据变量选择结果(含冗余变量)

变量选择方法	变量选择结果
mRMR	X_4, X_2, X_1, X_5, X_6
NMIFS	$X_4, X_1, X_2, X_{11}, X_3$
kNN_MI_VS	X_4, X_2, X_1, X_3, X_5

从表 2 中仿真结果可以看出, mRMR 仍未选择全部有效输入变量, 而 NMIFS 算法在输入变量子集中选择了冗余变量 X_{11}, 3 种变量选择方法中仅有本文方法选择了全部有效输入变量. 综合表 1 和表 2 的变量选择结果, 本文提出的基于 k-近邻互信息的变量选择方法能更有效地选择输入变量子集.

4.2 Lorenz 混沌时间序列仿真

为进一步说明本文所提出变量选择方法的有效性, 对 Lorenz 混沌时间序列进行仿真. Lorenz 序列的方程如下所示:

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = a(-x + y), \\ \dfrac{\mathrm{d}y}{\mathrm{d}t} = bx - y - xz, \\ \dfrac{\mathrm{d}z}{\mathrm{d}t} = xy - cz. \end{cases} \tag{12}$$

其中: $a = 10, b = 28, c = 8/3$, 初始值 $x(0) = 12, y(0) = 2, z(0) = 9$.

利用 4 阶 Runge-Kutta 法求解时间序列 $x(t)$, $y(t)$, $z(t)$, 步长 0.02, 取 2 000 组作为输入输出样本对, 其中 1 500 组用于训练, 其余 500 组进行测试. 仿真中对 $x(t)$, $y(t)$, $z(t)$ 序列进行相空间重构, 根据文献 [9] 延迟时间分别设定为 8, 7, 8, 嵌入维数均为 6, 得到输入样本为 18 维.

$$X(t) = [x(t), x(t-8), \cdots, x(t-5\times 8), y(t), y(t-7), \cdots, y(t-5\times 7), z(t), \cdots, z(t-5\times 8)]. \tag{13}$$

单步预测的输出为

$$Y(t) = [x(t+1), y(t+1), z(t+1)]. \tag{14}$$

预测模型采用广义回归神经网络 (GRNN). 采用下式所示均方根误差评价预测精度:

$$E_{\mathrm{RMSE}} = \left(\frac{1}{N_{\mathrm{test}} - 1}\sum_{k=1}^{N_{\mathrm{test}}}[\hat{y}(k) - y(k)]^2\right)^{1/2}. \tag{15}$$

其中: $y(k)$ 为预测变量的实际值, $\hat{y}(t)$ 为该变量的预测值, N_{test} 为对应测试样本数目.

表 3 对预测均方根误差进行了比较, 从表中结果可以看出, $x(t)$, $y(t)$, $z(t)$ 序列经过变量选择后的预测误差明显低于选择全部变量作为输入变量的情况, 说明了本文方法能够实现对 Lorenz 混沌时间序列特征变量的选择及建模预测.

表 3　变量选择后的预测误差比较

序列	全部变量	变量选择结果	选择变量
$x(t)$	0.973 1	$x(t), y(t), z(t-8)$	0.158 1
$y(t)$	0.534 3	$y(t), x(t-8), x(t-16)$	0.279 2
$z(t)$	1.042 8	$z(t), z(t-16), z(t-8)$	0.272 0

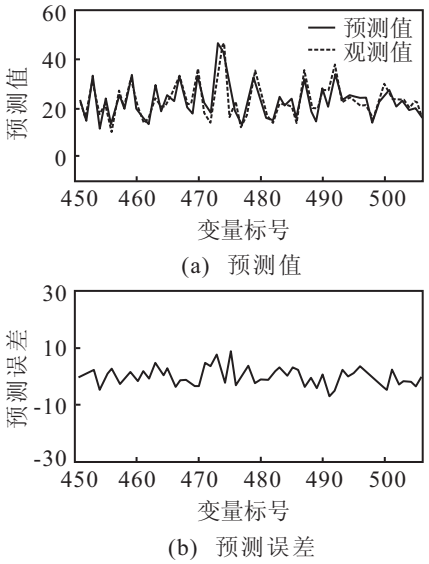
4.3 Housing 数据建模预测

为验证本文所提出变量选择方法的有效性, 以 UCI 数据中的 Housing data 进行仿真分析[7]. 通过每栋房子周围代表 13 个属性的统计值 (定义为输入变量 $X_1, \cdots, X_{13}$) 来预测每栋房子的房价 (定义为输出变量 Y).

利用本文算法对 Housing 数据进行输入变量选择, 并通过 GRNN 网络进行回归预测. 仿真并与文献 [7] 中变量选择方法, mRMR 以及 NMIFS 变量选择算法进行比较. 在预测模型中, 随机选取 Housing 数据中 338 组作为训练数据, 其余 168 组作为测试数据, 表 4 比较了 50 次仿真均方根误差的均值与方差. 图 1 和图 2 为一次运行的仿真结果, 为便于说明, 图中仅选择其中 56 个预测点的结果进行分析说明.

表 4　数据变量选择结果及预测精度

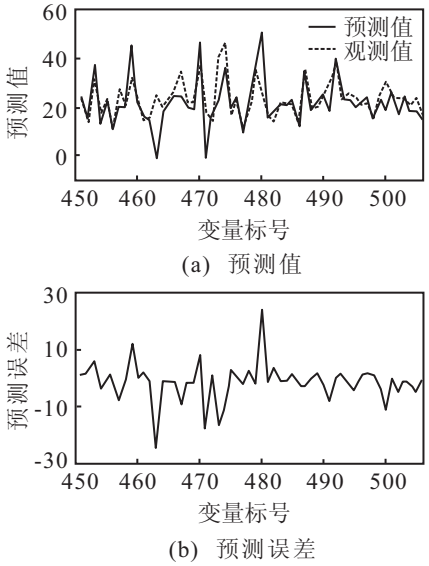
变量选择方法	选择变量结果	E_{RMSE}
全部变量	$X_1, X_2, \cdots, X_{13}$	16.543 1±1.149 2
文献[7]	X_6, X_{13}, X_1, X_4	4.729 8±0.539 9
mRMR	$X_7, X_6, X_{12}X_{13}$	8.002 4±0.715 0
NMIFS	X_{13}, X_6, X_{11}, X_7	6.321 5±0.460 0
kNN_MI_VS	X_6, X_{13}, X_1, X_3	**4.488 9±0.465 1**



(a) 预测值

(b) 预测误差

图 1 变量选择后预测结果及误差曲线



(a) 预测值

(b) 预测误差

图 2 全部变量的预测结果及误差曲线

从仿真结果可以看出, 本文方法的预测精度最高. 与其他变量选择方法相比, 本文方法选择的变量能够建立更为有效的预测模型.

5 结 论

本文通过对输入输出变量之间的互信息相关分析, 提出在变量选择过程中若在变量子集中增加相关变量, 则输入输出互信息的变化率较大. 根据该原则, 设计了前向式变量选择算法, 通过对所有变量完成遍历, 实现了输入变量的选择. 通过仿真研究验证了本文所提出变量选择方法不仅能够有效地选择相关变量, 而且还能去除掉冗余变量, 从而建立了较为有效的预测模型. Lorenz 混沌时间序列及 Housing 数据的预测结果均表明了本文方法的有效性.

参考文献(References)

[1] Roux E, Hernandez A I, Graindorge L, et al. Multivariate analysis of follow-up physiological data recorded by cardiac implantable devices[C]. Computers in Cardiology. Piscataway: IEEE, 2006: 765-768.

[2] 任海军, 张晓星, 孙才新, 等. 短期负荷多变量混沌时间序列正则化回归局域预测方法[J]. 计算机科学, 2010, 37(7): 123-140.
(Ren H J, Zhang X X, Sun C X, et al. Regulation regression local forecasting method of multivariable chaotic time series in short-term electrical load[J]. Computer Science, 2010, 37(7): 220-224.)

[3] Fernando 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]. J of Hydrology, 2009, 367(3/4): 165-176.

[4] Peng Hanchuan, Long Fuhui, Ding Chris. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27(8): 1226-1238.

[5] Pablo A Estévez, Michel Tesmer, Claudio A Perez, et al. Normalized mutual information feature selection[J]. IEEE Trans on Neural Networks, 2009, 20(2): 189-201.

[6] Alexander Kraskov, Harald Stogbauer, Peter Grassberger. Estimating mutual information[J]. Physical Review E, 2004, 69(6): 066138.

[7] 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): 1276-1288.

[8] Cover T M, Thomas J A. Elements of information theory[M]. New York: Wiley, 2006: 19-22.

[9] 韩敏, 魏茹. 基于改进典型相关分析的混沌时间序列预测[J]. 大连理工大学学报, 2008, 48(2): 292-297.
(Han M, Wei R. Chaotic time series prediction based on modified canonical correlation analysis[J]. J of Dalian University of Technology, 2008, 48(2): 292-297.)