基于 Copula 熵的神经网络径流预报模型预报因子选择

陈 璐[1]，叶 磊[1]，卢韦伟[1]，周建中[1]，郭生练[2]，肖 舸[3]，陈健国[3]
(1. 华中科技大学水电与数字化工程学院，武汉 430074；2. 武汉大学水资源与水电工程科学国家重点实验室，武汉 430072； 3. 三峡梯调通信中心，湖北 宜昌 443133)

摘 要：采用神经网络进行水文预报的关键问题之一是预报因子（输入变量）的选择，目前国内尚缺有效、系统的理论方法，国外主要是采用偏互信息(Patial mutual information，PMI)法。本文针对偏互信息计算方法的缺陷，引入 Copula 熵的概念，推导 Copula 熵与互信息的关系，提出采用 Copula 熵计算 PMI；并借助模拟试验检验了所提方法的合理性；最后，将该方法应用到三峡水库的水文预报中，并与现行方法进行了比较分析。结果表明，本文所提方法不仅具有理论基础，而且结果合理可信。

关键词：水文学及水资源；神经网络；水文预报；预报因子选择；Copula 熵；偏互信息
中图分类号：TV213 **文献标识码：**A

Determination of input variabes for artificial neural networks for flood forecasting using Copula entropy method

CHEN Lu[1], YE Lei[1], LU Weiwei[1], ZHOU Jianzhong[1], GUO Shenglian[2], XIAO Ke[3], CHEN Jianguo[3]
(1. *College of Hydropower & Information Engineering*, *Huazhong University of Science & Technology, Wuhan* 430074; 2. *State Key Laboratory of Water Resources and Hydropower Engineering Science*, *Wuhan University*, *Wuhan* 430072; 3. *Three Gorges Cascade Dispatching & Communication Center*, *Yichang*, *Hubei* 443133）

Abstract： One of the key steps in artificial neural networks (ANN) forecasting is the determination of significant input variables. A partial mutual information (PMI) method was used to characterize the dependence of a potential model between its input and output variables. We also developed a copula entropy method for effective calculation of mutual information (MI) and PMI, and verified its accuracy and performance using numerical tests. This forecasting technique has been applied to a real-world case study of the Three Gorges reservoir (TGR), and results show that the proposed method is useful and effective for identification of suitable inputs of flood forecasting model.

Key words： hydrology and water resources; artificial neural networks (ANN); flood forecasting; input setection; Copula entropy; partial mutual information

0 引言

目前，神经网络已普遍应用于水文预报中。采用神经网络进行水文预报的关键问题之一是预报因子（输入变量）的选择。其中最为常用的方法是线性相关系数法。采用线性相关系数，需要满足两个假设：一是两变量的相关性必须是线性的；二是变量须服从多元正态分布。因此，这种方法在实际应用中具有很强的局限性。一种替代的方法是互信息（Mutual information, MI）法，它表示一个随机变量中包含的关于另一个随机变量的信息量。朱永英等（2009）结合预报因子与预报对象的相关性分析，利用粗集理论的属性重要性概念对预报因子进行优化和选择[1]。赵铜铁钢和杨大文（2011）采用互信息方法，选择了基于神经网络水文预报的预报因子[2]。Sharma（2000）对互信息的方法进行了改进，提出了偏互信息 (Patial Mutual Information，PMI)的概念[3]。偏互信息计算了当一个新的变量加入时，输入集合对输出集合相关性的增量。

收稿日期：2012-12-26
基金项目：国家自然科学基金项目(51309104, 51239004)；湖北省自然科学基金（2013CFB184）；武汉市科技计划项目（2014060101010064）
作者简介：陈璐（1985-），女，讲师，E-mail: chl8505@126.com。

偏互信息类似于偏相关性，与互信息方法相比，它的优势在于剔除了已选入的变量对结果的影响。Bowden 等（2005）在比较多种方法之后认为基于偏互信息的方法较好[4]。Fernando 等（2009）在 Bowden 等（2005）研究的基础上，为提高 PMI 值的计算效率，改进了 PMI 的计算方法，并提出了新的算法停止准则[5]。May 等（2008）对 PMI 算法进行了计算精度评价和复杂性分析，证明这一方法不仅可以简化神经网络的结构，还能够有效提高预报精度[6]。但上述 PMI 计算方法仍然存在如下不足：①降雨、径流等水文变量均为连续的，但上述计算方法主要基于离散数据的假设；②以上方法需要估计变量的边缘和联合概率密度分布，对于 d 维的多元分布，联合概率密度函数难以求得。

考虑到偏互信息在水文预报因子选择中的作用，针对已有 PMI 计算方法存在的问题，本文引入 Copula 熵的概念，提出采用 Copula 熵计算 PMI 值，并借助模拟试验检验了该方法的合理性和计算精度；最后，将该方法应用到三峡水库的水文预报中。

1 Copula 熵理论

令 $x \in R_d$ 为 d 维随机变量，其边缘分布函数为 $F_i(X)$，$U_i = F_i(X)$，$i=1,2,\ldots,d$。其中，U_i 为服从均匀分布的随机变量，u_i 为随机变量 U_i 的具体数值。Copula 函数的熵可用下式表示：

$$H_C(u_1,u_2,\cdots,u_d) = -\int_0^1\cdots\int_0^1 c(u_1,u_2,\cdots,u_d)\log(c(u_1,u_2,\cdots,u_d))\mathrm{d}u_1\cdots\mathrm{d}u_d \tag{1}$$

式中：$c(u_1,u_2,\cdots,u_d)$ 为 Copula 函数的概率密度函数，可以表示为 $\dfrac{\partial C(u_1,u_2,\cdots,u_d)}{\partial u_1\partial u_2\cdots\partial u_d}$。

2 神经网络模型

人工神经网络在水文领域的应用，最为直接的方式是模拟径流-径流或降雨-径流的相关关系，本研究模拟径流之间的相关关系，表达式为：

$$y(t) = F_{ANN}(y(t-1),y(t-2)\cdots y(t-p),\boldsymbol{X}(t-1)\cdots\boldsymbol{X}(t-q)) \tag{2}$$

式中：$y(t)$ 为流域出口处 t 时刻的洪水预报值；$\boldsymbol{X}$ 为其相关上游站点的流量值；p、q 为滞时。

人工神经网络中有多种模型，其中广义回归神经网络（Generalized Regression Neural Network, GRNN）是 Specht 于 1991 年提出的一种新型神经网络。本文利用该神经网络结构固定、训练时间短和自学习能力强的特点，建立 GRNN 模型，进而实现水文预报。

3 预报因子的选择

3.1 偏互信息 PMI

Sharma（2000）对互信息的方法进行了改进，引入了偏互信息（PMI）的概念[3]。偏互信息度量了在消除其它变量影响的条件下，某两变量之间的互信息。针对本研究而言，偏互信息估计了在剔除已选变量影响的情况下，新加入的预报因子与输出变量之间的相关性。PMI 的表达式为[3]：

$$PMI = \iint f_{X',Y'}(x',y')\ln[\frac{f_{X',Y'}(x',y')}{f_{X'}(x')f_{Y'}(y')}]\mathrm{d}x'\mathrm{d}y' \quad x' = x - E[x\,|\,\boldsymbol{z}];\ y' = y - E[y\,|\,\boldsymbol{z}] \tag{3}$$

式中：E 表示期望值；x 表示待选 ANN 网络的预报因子；y 表示 ANN 网络的输出；$\boldsymbol{Z}$ 表示已选入的预报因子集合。$E[x|\boldsymbol{z}]$、$E[y|\boldsymbol{z}]$可采用 GRNN 神经网络进行计算。

3.2 基于 Copula 熵的 MI(PMI)的计算

1）Copula 熵与 MI 的关系研究

互信息和 Copula 函数的熵有着内在的关系，经推导，其满足下式[7]：

$$H(X_1,X_2,\cdots,X_d) = \sum_{i=1}^{d} H(X_i) + H_C(u_1,u_2,\cdots,u_d) \tag{4}$$

式（4）表明多维变量的联合熵可分解为两部分，即边缘熵的和与 Copula 函数的熵。本文仅考虑二维

变量的互信息，即当 d=2 时，根据式（4）有：

$$T(X_1,X_2)=H(X_1)+H(X_2)-H(X_1,X_2)=-H_C(X_1,X_2) \tag{5}$$

2） Copula 熵的计算

可以采用以下两种方法计算 Copula 函数的熵：

① 多重积分方法：根据 Copula 熵的定义，Copula 熵可以直接通过对式（1）积分得到。

② Monte Carlo 方法：当变量较多时，计算被积函数的多重积分往往比较困难，可采用 Monte Carlo 模拟的方法计算 Copula 熵。

令 $\boldsymbol{U}\in[0,1]$，$i=1,\cdots,d$，Copula 熵可表示为：

$$H_C(u_1,u_2,\cdots,u_d)=-\int_{[0,1]^d} c(\boldsymbol{U})\ln c(\boldsymbol{U})\mathrm{d}\boldsymbol{U}=-E[\ln c(\boldsymbol{U})] \tag{6}$$

因此，Copula 函数的熵就等于 ln[$c(\boldsymbol{U})$]的期望值。

3.3 算法停止准则

PMI 算法需要制定一个可靠和有效的标准，以确定偏互信息有多大时，可将变量纳入模型的输入集。Fernando 等（2005）和 May 等（2008）推荐采用 Hampel 检验作为算法的停止准则[5,6]。Hampel 准则的表达式为：

$$Z_j=\frac{d_j}{1.4826d_j^{(50)}}\qquad d_j=\left|PMI-PMI^{(50)}\right| \tag{7}$$

式中：$d_j^{(50)}$ 表示所有 d_j 的中位数。如果 Hampel 距离大于 3，那么此输入变量被选入输入变量集，即确定为模型的预报因子。

3.4 预测因子的选择步骤

首先，确定神经网络模型的所有可能的输入变量，定义此变量集为 $\boldsymbol{C}_{\mathrm{in}}$；通过 Copula 熵的方法计算 PMI 值，确定最终的输入变量集为 $\boldsymbol{C}$。算法的具体步骤如下：①在输入变量集 $\boldsymbol{C}$ 的条件下，依据 Copula 熵计算 $\boldsymbol{C}_{\mathrm{in}}$ 中每个可能输入与输出之间的的 PMI。其中，x、y 的条件期望值通过 GRNN 计算；②采用式（7），计算所有 PMI 值所对应的 Z_j 值；③如果最大的 PMI 值的 Z_j 值大于 3，那么将此变量作为模型的输入，添加到集合 $\boldsymbol{C}$ 中；④重复步骤①至③，直至所有相关变量都被选入 $\boldsymbol{C}$。

4 模拟试验

在进行实际应用之前，有必要采用统计试验来验证所提方法的合理性与实用性。检验的原理为：首先，随机生成一组已知相关性的序列；其次，采用本文所提方法对变量的相关性进行验证，找出有关联的变量。

采用以下两种函数进行模型的检验，模型的表达式如下：

1）AR9

$$x_t=0.3x_{t-1}-0.6x_{t-4}-0.5x_{t-9}+e_t \tag{8}$$

式中：e_t 是一个均值为 0 和标准差为 1 的高斯随机噪声。

2）TAR2-门限自回归模型

$$x_t=\begin{cases}-0.5x_{t-6}+0.5x_{t-10}+0.1e_t & \text{if } x_{t-6}\leqslant 0\\ 0.8x_{t-10}+0.1e_t & \text{if } x_{t-6}>0\end{cases} \tag{9}$$

每个模型生成 1020 数据点，丢弃前 20 个数据点，以减少数据初始化的影响[8]。以 t 时刻以前 15 个变量 $x_{t-1},\cdots,x_{t-15}$ 作为模型的可能输入，采用 Copula 熵方法选择输入变量，最终结果如表 1 所示。以 AR9 模型为例，经 PMI 计算可得，x_{t-1}、x_{t-4} 和 x_{t-9} 被视为与 x_t 相关的变量。这与式（8）相吻合，x_{t-1}、x_{t-4}、x_{t-9} 与 x_t 之间存在线性相关关系。可知，所提变量能够有效、准确地确定具有相关关系的变量，能够用来选取 ANN 模型的预报因子。

表 1 函数模拟试验结果
Table 1 Final input variable sets selected for the two models

函 数	选择的变量
AR9	$x_{t\text{-}4}$, $x_{t\text{-}9}$, $x_{t\text{-}1}$
TAR2	x_{10}, x_6

5 应用研究

5.1 研究区域

选用金沙江屏山站、岷江高场站、沱江李家湾站、嘉陵江北碚站、乌江李家湾站以及长江干流宜昌站，汛期 1998-2007 年的同步日流量资料参与计算。将上游金沙江屏山站、岷江高场站、沱江李家湾站、嘉陵江北碚站、乌江武隆站和宜昌站上游前期 13 天的实测日流量，即 $x_{t-13},x_{t-12},\cdots,x_{t-1}$，作为可能的输入变量，宜昌站 t 时刻的流量作为输出变量。

5.2 预报因子的选择

基于上述所提偏互信息的理论，Bowden 等（2005）提出了两阶段的方法来选择预报因子[8]。本文按照此思路，将预报因子的选择问题亦分为两个阶段：第一阶段，在各站内，采用 Copula 熵方法选择对宜昌站日流量影响较为显著的预报因子，结果如表 2 所示；在此阶段，原有的 78 个预报因子（13×6）减少到 12 个；宜昌站的前期流量对 y_t 的影响较大，因此有 5 个因子入选；对于高场站、李家湾站、北碚站以及武隆站，选择变量得到的滞时，基本与洪水到宜昌站的传播时间相一致。

表 2 PMI 方法第一阶段选择结果
Table 2 Input variables selected in step one using PMI method

站名	前期入选时刻
屏山	t-1, t-4, t-2
高场	t-3
北碚	t-2
李家湾	t-3
武隆	t-2
宜昌	t-1, t-2,t-3, t-4,t-5

第二阶段，将第一步选中的预报因子组成一个集合，即表 2 中数据，再次运用 Copula 熵方法，确定最终结果。在这个阶段，12 个输入变量减少到了 6 个。最后的输入变量包括宜昌站 t-1 时刻的流量、高场站 t-3 时刻的流量、李家湾站 t-3 时刻的流量、北碚站 t-2 时刻的流量、武隆站 t-2 时刻的流量以及屏山站 t-1 时刻的流量。

5.3 与其它方法的比较分析

Bowden 等（2005）指出相关系数法是应用最为广泛的预报因子选择方法[4]。因此，本文将 PMI 的结果与线性相关系数得出的结果进行比较分析。首先，计算预报因子与输出变量之间的 Pearson 线性相关系数，各站选取相关系数最大的因子，作为模型的输入，结果见表 3。由表 3 可知，本文提出方法和线性相关系数法得到的结果存在一定差异。如 PMI 的方法选择高场站滞后时间 3 天为输入，相关系数法选择高场站滞后时间 4 天为输入，而高场和宜昌之间实测数据的洪水传播时间为 3 天。将用两种方法得到的输入集合，用于宜昌站的洪水预报。采用 1998-2007 年十年的汛期数据参与计算，80%的数据（1998-2005）用于 ANN 模型的训练，20%的数据（2006-2007）用于模型的检验。基于两种预报因子集合，运用广义自回归神经网络（GRNN）预报宜昌站的流量，计算预报结果的确定性系数、合格率和实测值和预报值的均方根误差 RMSE，结果如表 4 所示。综上表明，基于 PMI 方法的预报结果明显优于线性相关系数法。图 1 给出了 2006 年的洪水预报结果，表明实测值与理论值拟合较好。

采用现行方法预报了 2006-2007 年汛期的日流量，相应的确定性系数、合格率和 RMSE 值为 0.934、0.95 和 2425。比较分析可知，本文所提方法的计算结果均优于现行方法。由于本文研究受水文资料所限，

未能考虑库区产流以及区间入流，限制了计算指标的进一步提高。

表 3　Copula 熵与线性相关系数选择结果的比较分析
Table 3　Comparisons of the input variable sets selected by the method of Pearson linear correlation coefficients and the proposed PMI method

河流	站点	相关系数（滞时）	Copula 熵（滞时）
金沙江	屏山	*t*-1	*t*-1
岷江	高场	*t*-4	*t*-3
沱江	李家湾	*t*-3	*t*-3
嘉陵江	北碚	*t*-2	*t*-2
乌江	武隆	*t*-2	*t*-2
长江	宜昌	*t*-1	*t*-1,*t*-2

表 4 不同预报因子对预报结果影响的比较分析
Table 4　Comparison of forecasting results obtained with different input variables

方法	确定性系数		RMSE		合格率	
	率定	检验	率定	检验	率定	检验
相关系数	0.9231	0.9036	1476	2932	0.9857	0.8566
Copula 熵	0.9402	0.9341	1281	2423	0.9898	0.9590

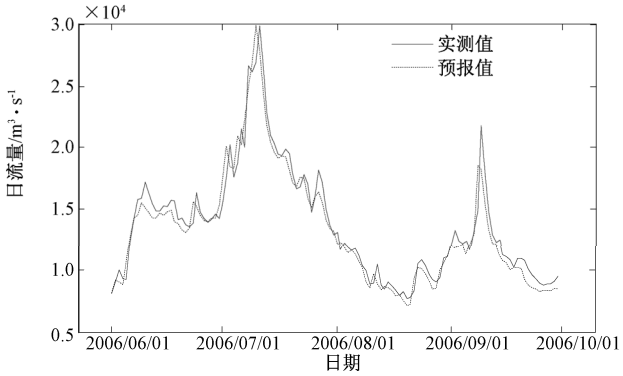

图 1　2006 年实测与预报流量的比较分析
Fig. 1　Comparisons between the observed and predicted runoff of the year 2006

6　结论

本研究率先提出了 Copula 熵的概念，探讨了 Copula 熵和互信息的关系，基于 Copula 熵计算了偏互信息 PMI 值，利用该值选择了神经网络的水文预报因子，得出如下结论：

① 互信息等于 Copula 函数的负熵；Copula 熵是度量线性、非线性相关关系的有效工具，它具有非参数性以及对变量分布函无限制等优点。通过 Copula 熵计算互信息、偏互信息，无需计算变量的边缘分布和联合分布，使得计算更为方便，避免了误差的累积。

② 精度和函数检验的结果表明，所提方法计算互信息值精度较高；能够准确可靠的找出具有相关性的变量，且对线性相关和非线性相关均适用。

③ 采用 Copula 熵对神经网络的水文预报因子进行选择，并用所建模型预报了宜昌站的流量，结果表明，该模型提高了水文预报的精度。

参考文献：

[1] 朱永英, 周惠成, 彭慧. 粗集-模糊推理技术在水文中长期预报中的应用研究[J]. 水力发电学报, 2009, 28(1):45-49.
ZHU Yongying, ZHOU Huicheng, PENG Hui. Rough fuzzy inference model and its application for mid-long term hydrological forecast[J]. Journal of Hydroetectric Engineering, 2009, 28(1):45-49.　(in Chinese).

(下转至第 90 页)

者提供参考和决策依据。

④ 由于城市水安全问题的复杂性及不同城市水安全问题的差异性，因此，在对某一具体城市进行水安全评价时，上述指标体系中的具体评价指标可根据各城市的实际情况进行适当的增减和调整，以得到更为客观、能反映各城市水安全特征的评价结果。

参考文献：

[1] 畅明琦，刘俊萍，黄强．水资源安全 Vague 集多目标评价及预警 [J]．水力发电学报，2008，23(6): 81-87，100.
CHANG Mingqi，LIU Junping，HUANG Qiang． Multiobjective assessment and early warning of water resources security based on Vague set [J]. Journal of Hydroelectric Engineering, 2008, 23(6): 81-87, 100. (in Chinese)

[2] Sullivan, Caroline. Calculaiing a Water Poverty Index. World Development, 2002(7): 1195-1210.

[3] Souro D Joardar. Carrying Capacities and Standards as Bases towards Urban Infrastructure Planning in India: A Case of Urban Water Supply and Sanitation [J]. Urban Infrastructure Planning in India, 1998, 22(3): 327-337.

[4] Michiel A Rijsberman, Frans H M. van de Ven. Different Ap2proaches t o Assessment of Design and Management of Sustainable Urban Water System [J]. Environment In pact Assessment Review , 2000, 129 (3): 333 -345.

[5] Patrick J Sullivan, Franklin J Agardy, James J Clark. Living with the Risk of Polluted Water [M]. Amsterdam;Burlington, MA: Elsevier Butterworth-Heinemann, 2005:143-196.

[6] 李永，朱明，李嘉．基于 Vague 集相似度量模型的城市水安全应急保障能力评价 [J]，水利学报，2009(40) 5:609-622.
LI Yong , ZHU Ming , LI Jia. Evaluation method of urban water security assurance capability for emergency rescue based on similarity measure of Vague sets [J]. Journal of Hydraulic Engineering, 2009(40) 5:609-622. (in Chinese)

[7] 史正涛，刘新有，黄英，等.基于边际效益递减原理的城市水安全评价方法 [J].水利学报，2010(41) 5:545-552.
SHI Zhengtao, LIU Xinyou, HUANG Ying, et al. Evaluation method for urban water safety based on law of diminishing marginal utility [J]. Journal of Hydraulic Engineering, 2010, 41(5): 545-552. (in Chinese)

[8] 贡力. 基于 WPI 的水安全评价体系研究 [J].中国农村水利水电，2010(9):4-7.
Gong Li. Evaluation system of water security based on water poverty index [J]. China Rural Water and Hydropower, 2010(9):4-7. (in Chinese)

[9] Peter Lawrence, Jeremy Meigh, Caroline Sullivan. The Water Poverty Index: International comparisons [R]. Center for Ecology &hydrology, Wallingford, United Kingdom.

[10] 陈金凤，傅铁.水贫乏指数在社会经济干旱评估中的应用 [J].水电能源科学，2011(29) 9:130-133.
CHEN Jinfeng, FU Tie. Application of water poverty index to socioeconomic drought assessment [J].Water Resources and Power, 2011(29) 9:130-133. (in Chinese)

[11] 张凤太，王腊春，苏维词，等.基于熵权集对耦合模型的表层岩溶带“二元”水资源安全评价[J].《水力发电学报》，2012,31(6): 70-76.
ZHANG Fengtai,WANG Lachun，SU Weici，et al. Evaluation on the safety of epikarst dualistic water resources by coupling model of entropy weight set pair[J]. Journal of Hydroelectric Engineering, 2012, 31(6): 70-76. (in Chinese)

[12] 金菊良,吴开亚,李如忠,等.信息熵与改进模糊层次分析法耦合的区域水安全评价模型[J].《水力发电学报》，2007,26(6): 61-66，110.
JIN Juliang ,WU Kaiya, LI Ruzhong, et al. Region water security evaluation method based on information entropy and improved fuzzy analytic hierarchy process[J]. Journal of Hydroelectric Engineering, 2007, 26(6): 61-66,110. (in Chinese)

(上接第 29 页)

[2] 赵铜铁钢，杨大文. 神经网络径流预报模型中基于互信息的预报因子选择方法[J]. 水力发电学报, 2011, 20(1), 24-30.
ZHAO Tongtiegang, YANG Dawen. Mutual information-based input variable selection method for runoff-forecasting neural network model[J]. Journal of Hydroelectric Engineering, 2011, 20(1): 24-30. (in Chinese).

[3] Sharma A. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 a strategy for system predictor identification [J]. Journal of Hydrology, 2000: 239, 232-239.

[4] Bowden Gavin J, Dandy Graeme C, Maierb Holger R. Input determination for neural network models in water resources applications. Part 1-background and methodology [J]. Journal of Hydrology, 2005, 301: 1-4: 93-107.

[5] 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]. Journal of Hydrology, 2009, 367:165-176.

[6] May Robert J, Maier Holger R, Dandy Graeme C, Fernando T.M.K. G. Non-linear variable selection for artificial neural networks using partial mutual information [J]. Environmental modeling & Software, 2008, 23: 1312-1326.

[7] Calsaverini R. S, Vicente R. An information-theoretic approach to statistical dependence: Copula information[J]. Europ. Phys. Lett., 2009,88(6): 3-12.

[8] Bowden Gavin J, Maierb Holger R, Dandy Graeme C. Input determination for neural network models in water resources applications. Part 2. Case study: forecasting salinity in a river[J]. Journal of Hydrology, 2005, 301 (1-4): 93-107.