基于虚假最近邻点 GT 准则的化工模型变量选择*

侯 杰[1], 李太福[2], 苏盈盈[2], 易 军[2]
(1. 重庆大学 自动化学院, 重庆 400044; 2. 重庆科技学院 电气与信息工程学院, 重庆 401331)

摘 要: 针对传统变量选择方法对复杂非线性化工模型进行变量选择时,由于缺乏输出变量的有效监督,导致所选择输入变量不能有效解释输出变量的问题,提出基于虚假最近邻点 Gamma 检验(Gamma test,GT)准则的变量选择方法。首先借鉴虚假最近邻点法,实现对所有变量的全面搜索;再采用能够在输出变量监督下进行非线性系统噪声估计的 GT 准则,计算各输入变量置零前后数据噪声的伽马统计量,得到输出变量对各输入变量的敏感度,以此为依据进行变量选择。使用线性、非线性模型验证了该方法的有效性。最后对氢氰酸复杂非线性化工过程建模进行变量选择,结果表明合理的变量选择有效地提高了模型精度。

关键词: 变量选择; Gamma 检验; 虚假最近邻点; 化工建模
中图分类号: TP181 **文献标志码**: A **文章编号**: 1001-3695(2012)11-4108-05
doi:10.3969/j.issn.1001-3695.2012.11.027

Variable selection approach based on false nearest neighbors and Gamma test for chemical industrial modeling

HOU Jie[1], LI Tai-fu[2], SU Ying-ying[2], YI Jun[2]
(1. *College of Automation, Chongqing University, Chongqing* 400044, *China*; 2. *School of Electrical & Information Engineering, Chongqing University of Science & Technology, Chongqing* 401331, *China*)

Abstract: To against traditional variable selection method could not obtain the input variables, which could always reasonably explain the output, because the chemical industrial modeling is nonlinearity and the method belongs to unsupervised learning. This paper put forward a method for complex nonlinear chemical industrial modeling based on the false nearest neighbors and Gamma test(GT). Firstly, it inspired by the false nearest neighbors, thinking about only one variable every time though setting it zero, could search all variables simply. Secondly, it calculated the Gamma statistics for every variable using the Gamma test when it's zero and not zero. Finally, it obtained the sensitivity of output to input to carryout the selection of variables. The example of linearity and nonlinearity were given to validate the method. Finally, confirming the input variables of BP artificial neural networks for the hydrocyanic acid by the method, the modeling is high precision. Therefore, it provides a new method for the variable selection of the complex nonlinear chemical industrial modeling.

Key words: variable selection; Gamma test(GT); false nearest neighbors; chemical industrial modeling

0 引言

随着数据获取、存储能力的增强,化工过程领域的样本维数迅速增长,冗余、噪声变量的存在容易产生维度灾难,使得采用所有变量建立的化工过程模型在计算时间、复杂度、解释能力等方面的性能变差。为了有效解决上述问题,借助变量选择技术在化工过程模型确定之前进行变量选择,选择恰当的变量作为模型输入,可降低模型复杂性,获得高精度模型[1]。研究化工过程建模的变量选择方法具有重要的科学意义和学术价值。

变量选择常用的搜索策略包括传统搜索策略、禁忌搜索策略[2]、群体智能全局搜索策略[3]。传统搜索策略主要有前向选择法(forward selection)[4]、后向剔除法(backward elimination)[5]、逐步回归法(stepwise regression)[6]等。鉴于化工过程普遍存在的复杂非线性特征,以上搜索策略存在各自的缺点。前向选择法和后向剔除法是单向增加或者减少变量数目,可快速地实现变量搜索,但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减,可以根据评估函数灵活地浮动,但其问题在于 l 和 r 的大小难以确定,且此方法也只适合于线性模型。禁忌搜索策略的特点类似于传统搜索策略。群体智能全局搜索策略使用先进群体智能方法(如遗传算法、蚁群算法等)进行全局搜索,但该方法计算量大、耗时严重。鉴于以上缺点,借鉴恢复混沌运动轨迹的特征筛选方法,即虚假最近邻点法[7],能够有效克服以上缺陷,快速全面地进行变量搜索。虚假最近邻点法已成功用于变量选择搜索[8,9]。

传统的虚假最近邻点法常采用相似度准则进行变量选择。相似度准则进行变量选择缺点是明显的,它只是从输入变量的

收稿日期: 2012-04-04; **修回日期**: 2012-05-12 **基金项目**: 国家自然科学基金资助项目(61174015,51075418);重庆市自然科学基金资助项目(CSTC2010BB2285)

作者简介:侯杰(1990-),男,云南曲靖人,硕士研究生,主要研究方向为智能控制与模式识别(ynhj88311@163.com);李太福(1971-),男,四川资阳人,教授,硕导,博士,主要研究方向为智能控制与模式识别;苏盈盈(1982-),女,黑龙江伊春人,讲师,博士研究生,主要研究方向为复杂系统故障诊断与变量选择;易军(1973-),男,重庆人,讲师,博士,主要研究方向为化工过程建模与优化.

数据结构出发,没有考虑输入/输出变量之间的相关关系,属于无监督变量选择准则;同时,相似度准则对非线性系统进行变量选择的有效性缺乏相关文献证明,因此,用其对复杂非线性化工过程建模进行变量选择,效果有限。作为非参数噪声估计准则的 GT[10] 准则,无须关注输入/输出变量之间任何参数关系,只依赖于输入/输出变量本身,可以在输出数据的监督下,将变量选择问题转换为噪声求解问题。其已经显示出强大的变量选择能力,可有效克服相似度准则进行变量选择的缺陷。文献[11]采用 GT 对降雨径流影响参数进行变量选择,从 18 个变量选择得到 5 个变量,建立了很好的 SVM 模型;文献[12]使用 GT 选择变量进行房价预测建模,得到很好预测效果;文献[13]在洪水预测中,使用 GT 进行变量选择,得到了很好效果;文献[14]在非线性降雨径流回归模型的确定中,使用 GT 对模型进行变量选择,从 25 个预测变量中有效地选出 19 个变量。

本文提出一种基于虚假最近邻点 GT 准则判别的变量选择方法。

1　虚假最近邻点相似度变量选择方法

虚假最近邻点法是 1992 年提出的用于确定相空间重构中嵌套维数的算法,也即在高维相空间重构过程中,随着嵌入维数 m 的增加,虚假最近邻点被逐步剔除,从而使混沌运动的轨迹得到恢复的一种特征筛选方法。

在 d 维相空间中,每一个相点矢量 $x(i)=\{x(i),x(i+\tau),\cdots,x(i+(d-1)\tau)\}$ 都有一个某距离内的最近邻点 $x^{NN}(i)$,其距离为

$$R_d(i)=\|x(i)-x^{NN}(i)\| \tag{1}$$

当相空间的维数从 d 维增加到 $d+1$ 维时,这两个相点的距离就会发生变化,两者的距离称为 $R_{d+1}(i)$,且有

$$R_{d+1}^2(i)=R^2d(i)+\|x(i+\tau d)-x^{NN}(i+\tau d)\| \tag{2}$$

如果 $R_{d+1}(i)$ 比 $R_d(i)$ 大很多,可以认为这是由于高维混沌吸引两个不相邻的点投影到低维轨道上时变成相邻的点造成的,因此这样的邻点是虚假的。

传统上,常采用计算相点的相关性来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 $A=(x_1,x_2,\cdots,x_i,\cdots,x_n)$,令 $x_i=0$,求出样本 A 在 i 维空间内的投影 $B=(x_1,x_2,\cdots,0,\cdots,x_n)$。计算 A、B 的相关性:

$$\cos(\theta)=\frac{A\cdot B^{\mathrm{T}}}{\|A\|\cdot\|B\|} \tag{3}$$

若 $\cos(\theta)$ 值接近于 1,则说明 A 与 B 的相似度大,变量 x_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 x_i 对样本解释能力较大,此时 B 即 A 的虚假最近邻点。

2　虚假最近邻点 GT 准则变量选择方法

2.1　GT 准则

Gamma test 是 1997 提出的一种对所有光滑函数均适用(输入到输出的转换是连续的,并且在输入空间中一阶导数是有界的)、无须关注输入数据和输出数据之间任何参数关系、只依赖于输入输出本身来估计误差的非参数估计方法。

对于如下形式的数据集:

$$\{(x_i,y_i),1\leqslant i\leqslant M\} \tag{4}$$

其中:$x\in\mathrm{R}^m$ 是输入,对应的输出标量为 $y\in\mathrm{R}$。

GT 假定的模型关系是:

$$y=f(x_1,\cdots,x_M)+r \tag{5}$$

其中:f 表示光滑函数,r 表示噪声变量。不失一般性,假定 r 的均值为 0,方差为 var(r),也即伽马统计量 Γ。

为了估计 var(r),GT 首先使用 kd-tree 算法在输入空间对各输入变量 $x_i(1\leqslant i\leqslant M)$ 进行计算,得到输入样本 $x_i(1\leqslant i\leqslant M)$ 的第 $K(1\leqslant K\leqslant P)$ 近邻域点 $x_{N[i,K]}(1\leqslant i\leqslant M)$,一般 $10\leqslant P\leqslant 50$。下一步计算所有 $x_i(1\leqslant i\leqslant M)$ 的第 P 近邻域点的最小均方距离 $\delta_M(K)$ 以及输出空间相应的最小均方距离 $\gamma_M(K)$。最后,对 $(\delta_M(K),\gamma_M(K))(1\leqslant K\leqslant P)$ 按式(8)进行一次线性回归,得到一次线性函数的截距,即为伽马统计量 Γ。

$$\delta_M(K)=\frac{1}{M}\sum_{i=1}^{M}|X_{N[i,k]}-X_i|^2\quad 1\leqslant K\leqslant P \tag{6}$$

$$\gamma_M(K)=\frac{1}{M}\sum_{i=1}^{M}|Y_{N[i,k]}-Y_i|^2\quad 1\leqslant K\leqslant P \tag{7}$$

$$\gamma=A\delta+\Gamma \tag{8}$$

当输入数据的距离统计量无限接近时,对应的输出数据距离统计量就是噪声的方差,则 $\delta_M(K)\to 0$,$\gamma_M(K)\to\mathrm{var}(r)$。

2.2　基于虚假最近邻点 GT 准则的变量选择方法

基于虚假最近邻点 GT 准则的变量选择方法借鉴虚假最近邻点法的思想快速有效地对所有变量进行搜索,再采用对所有光滑函数均适用的 GT 准则作为变量选择准则,对输入数据进行敏感度计算,根据敏感度排序进行变量选择。该方法可以有效克服传统虚假最近邻点相似度准则在变量选择时,无输出变量监督以及准则对非线性模型效果有限的缺点。

借鉴虚假最近邻点法思想,假设一个 n 维变量组空间 Q,其中一个样本点 $A=(x_1,x_2,\cdots,x_i,\cdots,x_n,y)$,令 $x_i=0$,求出样本 A 在 i 维空间内的投影 $B=(x_1,x_2,\cdots,0,\cdots,x_n,y)$,再对投影前后的数据 A、B 分别进行 GT 计算,求得 A、B 对应的伽马统计量 Γ_A、Γ_B,按式(9)求得输出变量 y 对输入变量 x_i 的敏感度 Γ_i。

$$\Gamma_i=\frac{|\Gamma_A-\Gamma_B|}{\Gamma_B} \tag{9}$$

$x_i=0$ 时,敏感度越小,表示剔除变量 x_i,只使用剩余输入变量对输出变量进行建模,噪声误差变化越小;输出变量对输入变量 x_i 的敏感度小,则变量 x_i 对输出变量的作用也越小,变量 x_i 越不重要。

$x_i=0$ 时,敏感度越大,表示剔除变量 x_i,只使用剩余输入变量对输出变量进行建模,噪声误差变化越大;输出变量对输入变量 x_i 的敏感度大,则变量 x_i 对输出变量的作用也越大,变量 x_i 越重要。

对 n 维变量组空间 Q,使用虚假邻点法对数据进行 n 次遍历搜索,每次均利用 GT 准则对数据进行 Γ_i 计算,得到所有变量对应的 Γ_i 排序,再依据 Γ_i 从大到小的排序结果进行变量选择。

2.3　复杂度分析

虚假最近邻点 GT 变量选择方法借鉴虚假最近邻点思想

对变量进行搜索,再采用GT准则进行变量敏感度的计算,确定输入变量被选择顺序,时间复杂度为$O(n(2m^2+2mp+p))$。其中,n为变量维数,m为样本个数,p为GT准则邻域数。而采用其他搜索算法进行变量搜索(对群体智能全局搜索策略中的遗传算法和蚁群算法进行分析),再采用GT准则进行输入变量被选择顺序的确定,时间复杂度分别如表1所示。

表1 时间复杂度

方法		时间复杂度
传统搜索策略	前向选择法	$O(n!(2m^2+2mp+p))$
	后向剔除法	$O(n!(2m^2+2mp+p))$
	逐步回归法	$O(n^3(2m^2+2mp+p))$
禁忌搜索策略		$O((2m^2+2mp+p)tn)$
群体智能全局搜索策略	遗传算法	$O((2m^2+2mp+p)tab)$
	蚁群算法	$O((2m^2+2mp+p)tcd(d-1))$

其中:n为变量维数,m为样本个数,p为GT准则选择邻域数;禁忌搜索策略中,t为进化次数;遗传算法中,t为进化次数,a为染色体长度,b为群体大小;蚁群算法中,t为迭代次数,c为蚂蚁数,d为城市数。

由结果可知,该方法的时间复杂度主要由GT准则决定。该方法的时间复杂度比采取相同准则进行变量选择的传统方法、禁忌搜索方法小。当$tab<n$时,该方法的时间复杂度大于遗传算法;当$tcd(d-1)<n$时,该方法的时间复杂度大于蚁群算法,但是$tab<n$或$tcd(d-1)<n$的情况通常不存在,一般$n<<tab$,$n<<tcd(d-1)$,该方法的时间复杂度远远小于群体智能全局搜索方法。通过虚假最近邻点搜索策略进行变量选择的搜索,有效降低了变量选择的时间复杂度,能够快速对所有变量进行遍历搜索。

3 仿真研究

3.1 仿真数据

这里分别使用构造线性、非线性模型对本文方法的有效性进行验证。

模型1 模型如式(10)所示,为输出y关于x_4、x_5的线性模型,系统还包括冗余、噪声变量x_1、x_2、x_3、ε。

$$y=51+3x_4+4x_5+\varepsilon \tag{10}$$

其中:$x_1,\cdots,x_5\sim N_{700}(0,1)$,$\varepsilon\sim N_{700}(0,\sigma^2 I)$,$\sigma=2.5$。

模型2 模型如式(11)所示,为输出y关于x_1、x_2、x_3的非线性模型,系统还包括冗余变量x_5、x_6。

$$y=\cos(2\pi x_1)\cos(4\pi x_2)\exp(x_2)\exp(2x_3) \tag{11}$$

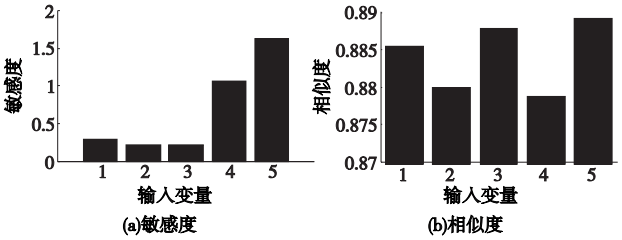
其中:$x_1,\cdots,x_5\sim[0,1]_{1500}$。

3.2 仿真结果及分析

根据GT准则的一般经验$10\leqslant P\leqslant 50$,本文采用$P=20$的GT准则进行敏感度计算,并进行变量选择。

借鉴虚假最近邻点思想,分别将变量$x_1,\cdots,x_M$依次置零,再使用GT准则计算相应变量置零前后数据的伽马统计量,据式(9)计算输出对输入的敏感度,以敏感度从大到小的排序结果进行变量选择。

模型1 使用所有5个输入变量作为输入数据,使用GT准则进行伽马统计量的计算,求得变量置零前的伽马统计量为$\Gamma=2.6321$;再分别将$x_1,\cdots,x_5$置零,使用GT准则对相应变量置零后的数据进行计算,得到相应伽马统计量依次为$\Gamma_1=3.4324$,$\Gamma_2=3.2143$,$\Gamma_3=3.2323$,$\Gamma_4=5.4364$,$\Gamma_5=6.9121$。根据式(9)计算得到相应变量的敏感度值如图1(a)所示;而采用相似度准则进行计算时,所得相似度值分别为$\cos(\theta_1)=0.8854$,$\cos(\theta_2)=0.8798$,$\cos(\theta_3)=0.8878$,$\cos(\theta_4)=0.8787$,$\cos(\theta_5)=0.8891$,相应变量的相似度值如图1(b)所示。



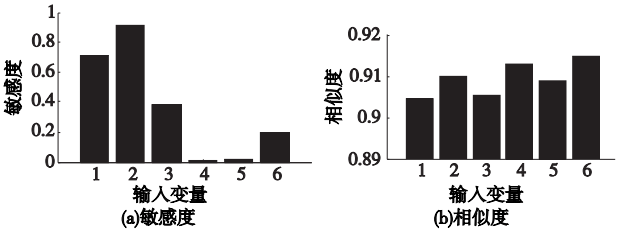
(a)敏感度 (b)相似度

图1 模型1仿真结果

由以上结果可以看出,使用本文变量选择方法,敏感度由大到小的排序依次为x_5、x_4、x_1、x_3、x_2,变量被选择顺序依次为x_5、x_4、x_1、x_3、x_2,所得结果与已知模型由x_4、x_5确定相一致。由于有输出变量的监督,采用虚假最近邻点GT准则对线性模型进行变量选择的效果良好。

使用传统虚假最近邻点相似度准则进行变量选择时,输入变量的相似度由小到大的排序依次为x_4、x_2、x_1、x_3、x_5,变量被选择顺序依次为x_4、x_2、x_1、x_3、x_5,所得结果与已知模型由x_4、x_5确定不一致。传统相似度准则进行变量选择,由于缺乏输出数据的监督,其效果有限。

模型2 使用所有6个输入变量作为输入数据,使用GT准则进行伽马统计量的计算,求得变量置零前的伽马统计量为$\Gamma=0.9668$;再分别将$x_1,\cdots,x_6$置零,使用GT准则对相应变量置零后的数据进行计算,得到伽马统计量依次为$\Gamma_1=1.6503$,$\Gamma_2=1.8452$,$\Gamma_3=1.335$,$\Gamma_4=0.8335$,$\Gamma_5=0.9836$,$\Gamma_6=1.1591$。根据式(9)计算得到相应变量的敏感度值如图2(a)所示,而采用虚假最近邻点相似度准则进行计算时,所得相似度值分别为$\cos(\theta_1)=0.9039$,$\cos(\theta_2)=0.9093$,$\cos(\theta_3)=0.9047$,$\cos(\theta_4)=0.9122$,$\cos(\theta_5)=0.9083$,$\cos(\theta_6)=0.9139$,相应变量的相似度值如图2(b)所示。



(a)敏感度 (b)相似度

图2 模型2仿真结果

从以上结果可以看出,使用本文变量选择方法时,敏感度由大到小的排序依次为x_2、x_1、x_3、x_6、x_5、x_4,变量被选择顺序依次为x_2、x_1、x_3、x_6、x_5、x_4,所得结果与已知模型由x_1、x_2、x_3确定一致。采用虚假最近邻点GT准则方法对线性模型进行变量选择,由于有输出变量的监督,其效果良好。

使用传统虚假最近邻点相似度准则进行计算时,输入变量的相似度由小到大的排序依次为x_1、x_3、x_5、x_2、x_4、x_6,变量被选择顺序依次为x_1、x_3、x_5、x_2、x_4、x_6,所得结果与已知模型由x_1、x_2、x_3确定不一致。相似度准则进行变量选择,由于缺乏输出数据的监督,其效果有限。

通过以上仿真结果可以看出,基于虚假最近邻点GT准则的变量选择方法,能够有效地对线性、非线性模型进行变量

选择。

4　实例研究

氢氰酸(HCN)生产过程复杂,监控参数繁多,存在复杂非线性关系。在 HCN 的生产过程中,设备都和空气接触,存在过多的监控参数,包括各种气体的温度、压力、流量和设备参数,这其中存在一定的无关、冗余变量,若采用所有变量进行建模,对 HCN 的后续优化控制将产生不良影响。为了建立在计算时间、复杂度、解释能力等方面的能力均较好的模型,采用本文变量选择技术在 HCN 模型确定之前进行变量选择,选择恰当的变量作为模型输入,剔除冗余特征参数,降低模型复杂性,得到精确 HCN 模型。

本文使用 HCN 生产过程中所得的 591 组实际生产数据(表 2)进行化工过程变量选择的研究,输入变量为 $X=\{x_1,x_2,\cdots,x_{10}\}$,输出变量为转换率 Y。其中:x_1 表示氨的补偿温度,x_2 表示氨的流量,x_3 表示天然气的流量,x_4 表示空气的流量,x_5 表示氨的补偿压力,x_6 表示天然气的补偿压力,x_7 表示空气的补偿压力,x_8 表示气泡压力,x_9 表示大混合器出口温度,x_{10}表示入塔器 HCN 百分比,Y 表示氨的转换率。其中转换率的计算公式为

$$Y=\frac{17}{27}\times\frac{0.771x_2+0.717x_3+1.293x_4}{0.771x_2}x_{10} \tag{12}$$

表 2　HCN 生产实绩数据汇总

参数	数目				
	1	2	…	590	591
x_1	31.6	31.3	…	31.8	32.1
x_2	601	598	…	599	600
x_3	707	699	…	699	709
x_4	3744	3733	…	3761	3780
x_5	2.00	2.02	…	1.99	2.00
x_6	1.99	2	…	1.98	2.00
x_7	1.99	1.99	…	1.99	2.00
x_8	2.90	3.01	…	2.99	2.99
x_9	78	82	…	79	78
x_{10}	8.88	8.78	…	7.69	8.79
Y	71.784 46	71.887 71	…	71.440 96	71.327 1

首先分别将 $x_1,\cdots,x_{10}$依次置零,再使用 GT 准则依次计算相应变量置零前后的伽马统计量,最后根据式(9)计算得到相应变量的敏感度,进而进行变量选择,以确定化工过程模型。使用所有 10 个输入变量作为输入数据,采用 GT 准则进行伽马统计量的计算,求得变量置零前伽马统计量的值为 $\Gamma=4.4223$,再分别将 $x_1,\cdots,x_{10}$置零,使用 GT 准则对变量置零后的数据进行计算,所得的伽马统计量依次为 $\Gamma_1=4.5258$,$\Gamma_2=5.4811$,$\Gamma_3=5.5228$,$\Gamma_4=3.0939$,$\Gamma_5=4.4223$,$\Gamma_6=4.4223$,$\Gamma_7=4.4223$,$\Gamma_8=4.4258$,$\Gamma_9=4.5210$,$\Gamma_{10}=4.0207$。根据式(9)计算得到相应变量的敏感度值如图 3(a)所示,而采用虚假最近邻点相似度准则进行计算时,所得相似度值分别为 $\cos(\theta_1)=1$,$\cos(\theta_2)=0.9879$,$\cos(\theta_3)=0.9835$,$\cos(\theta_4)=0.2393$,$\cos(\theta_5)=1$,$\cos(\theta_6)=1$,$\cos(\theta_7)=1$,$\cos(\theta_8)=1$,$\cos(\theta_9)=0.9998$,$\cos(\theta_{10})=1$,相应变量的相似度值如图 3(b)所示。

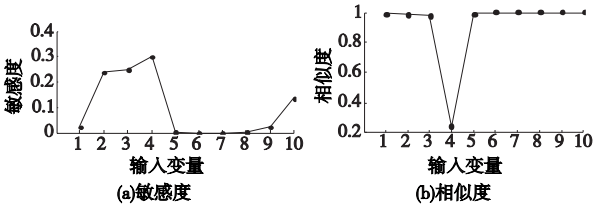


图3　HCN计算结果

通过以上结果可以看出,本文方法对复杂非线性化工过程进行变量选择时,输入变量敏感度由大到小的排序依次为 x_4、x_3、x_2、x_{10}、x_1、x_9、x_8、x_5、x_6、x_7,变量被选择顺序依次为 x_4、x_3、x_2、x_{10}、x_1、x_9、x_8、x_5、x_6、x_7。转换率与变量之间的关系如式(12),变量选择结果符合经验。而使用传统虚假最近邻点相似度准则变量选择时,输入变量的相似度由小到大的排序依次为 x_4、x_3、x_2、x_9、x_8、x_5、x_6、x_7、x_{10}、x_1,变量被选择顺序依次为 x_4、x_3、x_2、x_9、x_8、x_5、x_6、x_7、x_{10}、x_1,所得结果与已知模型不一致,对于会对模型造成很大影响的 x_{10},则没能很好地挖掘出它们之间的关系。说明相似度准则进行变量选择,由于缺乏输出数据的监督,效果有限。

下面按敏感度依次递减的方向加入输入变量,进行 BP 神经网络建模。BP 神经网络结构选择为(变量数 -8 -1),隐层函数为 tansig,输出层函数为 logsig,训练方式为 traingdx,设置训练次数为 700,期望误差为 10^{-6}。依次加入被选变量进行相应模型确定,模型结果如表 3 所示。

表 3　神经网络模型精度

序号	选择变量	模型精度			
		训练样本精度(500)		检验样本精度(91)	
		SSE	MSE	SSE	MSE
1	x_4	1.672 1	0.003 3	2.842 5	0.031 2
2	x_4,x_3	1.619 5	0.003 2	2.952 9	0.032 4
3	x_4,x_3,x_2	1.660 1	0.003 3	2.703 8	0.029 7
4	x_4,x_3,x_2,x_{10}	0.050 0	0.000 1	0.066 5	0.000 7
5	x_4,x_3,x_2,x_{10},x_1	1.909 5	0.003 8	2.722 0	0.029 9
6	$x_4,x_3,x_2,x_{10},x_1,x_9$	1.089 1	0.022 2	1.933 5	0.021 2
7	$x_4,x_3,x_2,x_{10},x_1,x_9,x_5,x_8,x_6,x_7$	0.215 8	0.000 4	0.158 3	0.001 7

由表 3 结果可以看出,以上七个模型中,模型 4 的精度最高,建立了高精度神经网络模型,本文变量选择方法取得很好的效果。通过对变量进行相关性分析,得到 x_4、x_3、x_2、x_{10}、x_1、x_9、x_8、x_5、x_6、x_7 对输出的相关度值依次为 0.3167、0.2999、0.3047、0.9901、0.1578、0.044、-0.0695、0.0220、-0.0987、-0.2899。可以看出,前四个变量对输出的相关性比后面变量对输出的相关性高,本文方法可以有效选择出和输出有很强相关性的输入变量。而采用传统相似度准则时,由于缺乏输出数据的有效监督,且因为化工系统存在复杂非线性,导致所选变量对输出的解释能力有限。该方法较虚假最近邻点相似度准则变量选择效果好,可以有效地对复杂非线性化工过程模型进行变量选择。

5　结束语

针对包含无关、冗余变量的高维输入不利于建立精确化工

模型的问题,采用变量选择方法在模型确定之前进行变量选择,选择适当的输入进行模型确定,可以有效地改善模型精度。鉴于传统变量选择不能有效选择得到对复杂非线性化工过程模型的输出变量有很好解释能力的输入变量,本文分别从变量选择的搜索策略和变量选择准则两个方面进行考虑,提出基于虚假最近邻点 Gamma 检验准则的变量选择方法。该方法借鉴 FNN 思想,能够快速全面地对所有变量进行遍历搜索。通过分析可知,采用 FNN 进行搜索,在搜索计算上较其他搜索策略有更快的搜索能力。在 FNN 对变量进行快速、全面搜索的基础上,采用有输出变量监督的 GT 准则来确定最终的模型输入变量,可以选择得到对模型输出有很好解释作用的输入变量,而没有输出变量监督的相似度选择准则所选择输入变量对输出变量的解释作用有限。通过对构造模型和实际模型的仿真研究,结果表明该方法对复杂非线性化工过程模型进行变量选择取得了很好的效果。本文变量选择方法对输入/输出变量是连续数据的模型均适用,下一步可以对其他系统或领域的模型进行变量选择的研究。但是,通过对算法进行时间复杂度分析可知,该算法的时间复杂度取决于 GT 准则,采用 GT 准则对大样本数据进行变量选择会对算法的时间复杂度造成不良影响,因此对大样本进行变量选择,更加有效的有监督变量选择的准则应被采用。

参考文献:

[1] 王心哲,韩敏. 基于变量选择的转炉炼钢终点预报模型[J]. 控制与决策,2010,25(10):1589-1592.

[2] PACHECO J, CASADO S, NUNEZ L. A variable selection method based on Tabu search for logistic regression models[J]. European Journal of Operational Research,2009,199(2):506-511.

[3] 王艳玲,李龙澍,胡哲. 群体智能优化算法[J]. 计算机技术与发展,2008,18(8):114-117.

[4] BLANCHET F G, LEGENDRE P, BORCARD D. Forward selection of explanatory variables[J]. Ecology,2008,89(9):2623-2632.

[5] CHAN K Y,KWONG C H,DILLON T S, *et al*. Reducing overfitting in manufacturing process modeling using backward elimination based genetic programming[J]. Applied Soft Computing,2011,11(2):2623-2632.

[6] 梁朝林,沈本贤,刘纪昌,等. 用延迟焦化逐步回归法模型预测焦化产物的分布[J]. 东华大学学报:自然科学版,2009,35(2):185-191.

[7] ABARBANEL H D I, KENNEL M B. Local false nearest neighbors and dynamical dimensions from observed chaotic data[J]. Physical Review E,1993,47(5):3057-3068.

[8] 李太福,易军,苏盈盈,等. 基于 KCCA 虚假邻点判别的非线性变量选择[J]. 仪器仪表学报,2012,33(1):213-22.

[9] 李太福,易军,苏盈盈,等. 基于特征子空间虚假邻点判别的软传感器模型变量选择[J]. 机械工程学报,2011,47(12):7-12.

[10] STEFANSSON A, KONCAR N, JONES A J. A note on the Gamma test[J]. Neural Computing and Applications,1997,5(3):131-133.

[11] NOORI R, KARBASSI A R, MOGHADDAMNIA A, *et al*. Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction[J]. Journal of Hydrology,2011,401(3-4):177-189.

[12] JAAFAR W Z W, HAN D. Variable selection using the Gamma test forward and backward selections[J]. Journal of Hydrologic Engineering,2012,17(1):182-191.

[13] 李德详. 基于 Gamma test 的最小二乘支持向量机参数在线优化方法[D]. 大连:大连理工大学,2010.

[14] 任启伟,陈洋波. 基于 Gamma test 的非线性降雨径流回归模型研究[J]. 水文,2010,30(1):39-43.

(上接第4086页)点度为8。设10%的节点属于 X 集合,20%的节点属于 Y 集合,在 Z 集合中随机选取20%的节点,设其初始观点为+1,其余节点设为0。每次实验最大时间步设置为 10^6 次。表2给出了不同接收概率情况下系统的最终状态分布,同时显示了不同参数情况下系统最终状态中观点为+1所占的人数比例,所有数据经过了10 000次平均得到。

表2 不同接收概率情况下系统的最终状态分布

λ	0.1	0.3	0.5	0.7	1.0
q	0.501 3	0.501 9	0.495 1	0.505 4	0.498 7

4 结束语

考虑到用户的初始意见和交互概率,本文提出了一种新的舆情传播模型。考虑到实际生活中有些人的意见非常坚定,而有些人则是不会轻易受到邻居的意见左右的,本文假定模型中有些用户的意见不随环境变化,而有些则可以与邻居进行交互从而改变自己的意见。通过对模型进行解析分析和数值模拟,发现格子网络中持有各种观点的人数比例与初始状态人群所占的比例有关,而与交流概率无关;对特定的网络存在特殊的参数也可以使得系统达到平衡状态。下一步工作将着重解决系统的收敛时间以及与收敛时间相关因素等问题。

参考文献:

[1] CASTELLANO C, FORTUNATO S, LORETO V. Statistical physics of social dynamics[J]. Reviews of Modern Physics, 2009, 81(12):591-646.

[2] 潘新. 基于复杂网络的舆情传播研究[D]. 大连:大连理工大学,2010.

[3] SZNAID-WERON K, WERON R. A simple model of price formation[J]. International Journal of Modern Physics C,2002,13(1):115-123.

[4] DEFFUANT G, NEAU D, AMELAND F, *et al*. Mixing beliefs among interacting agents[J]. Advances in Complex Systems, 2000,3(1-4):87-98.

[5] HEGSELMANN R, KRAUSE U, COMPUT E. Opinion dynamics driven by various ways of averaging[J]. Computational Economics, 2005,25(4):381-405.

[6] WU Fang, HUBERMAN B A. Social structure and opinion formation[R]. Palo Alto, CA:HP Labs,2006.

[7] 聂哲,李粤平,温晓军,等. 个体相互影响的网络舆情演变模型[J]. 计算机工程与应用,2009,45(14):220-222.

[8] 潘新,邓贵仕,佟斌. 基于社会网络的舆情传播模型构建与分析[J]. 运筹与管理,2011,20(2):176-179.

[9] KARATZAS I, SHREVE S E. Brownian motion and stochastic calculus[M]. 2nd ed. [S.l.]:Springer,1997.

人与人之间的关联关系。系统参数定义如表 1 所示。

表 1 变量定义及其意义

N	系统的节点数
k	节点度
b_k	度为 k 持观点 -1 且不变的人数比例
w_k	度为 k 持观点 +1 且不变的人数比例
n_k	度为 k 的节点数
$p_k = n_k/n$	度分布
m	系统中观点为 +1 的人数
m_k	度为 k 且观点为 +1 的人数
$q = m/n$	观点为 +1 的人占总人数比例
$q_k = m_k/n$	度 k、观点为 +1 的人占总人数比例

2 舆情传播模型

假设系统中只存在两种观点,分别用 +1 和 -1 表示。X 表示持有观点 +1 且其观点不会改变;Y 表示持有观点 -1 且其观点也保持不变;Z 表示观点可以根据邻居的观点以一定概率变化的人群。

模型的演化规则如下:在每个时间步内,随机选择一个节点 A,并随机选择其一个邻居 B。如果 A 节点属于集合 Z,则 A 节点以概率 λ 接受其邻居 B 的观点,其中 $\lambda \in (0,1]$。不失一般性,假设 A 节点持有观点 -1 且度为 k,根据 A 节点与 B 节点观点的不同,系统的更新可以归纳为以下两种情况:a) 如果 B 节点持有观点 -1,则 A 节点的观点都不改变;b) 如果 B 节点持有观点 +1,则 A 的观点以概率 λ 由 -1 变为 +1。A 节点由 -1 变为 +1 的概率可用式(1)表示。

$$P_{w\to b}(u) = \lambda p_k (1 - q_k - w_k) \frac{\sum_j j p_j q_j}{\sum_j j p_j} \tag{1}$$

其中:p_k、$(1 - q_k - w_k)$ 分别表示 A 属于集合 Z、度为 k 且观点为 -1 的概率。相应地,A 节点的观点由 +1 变成 -1 的概率为

$$P_{b\to w}(u) = \lambda p_k (q_k - b_k) \frac{\sum_j j p_j (1 - q_j)}{\sum_j j p_j} \tag{2}$$

如果定义式(3)为 q_k 的加权平均

$$\langle q \rangle = \frac{\sum_j j p_j q_j}{\sum_j j p_j} \tag{3}$$

则式(1)和(2)可以写为

$$\begin{aligned} p_{w\to b}(u) &= \lambda p_k (1 - q_k - w_k)\langle q \rangle \\ p_{b\to w}(u) &= \lambda p_k (q_k - b_k)(1 - \langle q \rangle) \end{aligned} \tag{4}$$

为了考察系统最终的演化状态,考察随时间变化系统中观点为 +1 的个数 m_k 和其占所有人数的比例 q_k。在时间间隔 $(t, t + dt)$ 内,m_k 的变化可以通过如下公式得到。

$$\Delta m_k = \begin{cases} +1 & \text{以概率 } \lambda n_k p_k (1 - q_k - w_k)\langle q \rangle \\ -1 & \text{以概率 } \lambda n_k p_k (q_k - b_k)(1 - \langle q \rangle) \\ 0 & \text{其他} \end{cases}$$

其中:$n_k = n p_k$,则 Δm_k 的期望值可以用式(5)度量。

$$\begin{aligned} E[\Delta m_k] &= \lambda n_k (1 - q_k - w_k)\langle q \rangle - \lambda n_k (q_k - b_k)(1 - \langle q \rangle) = \\ &\lambda n_k (\langle q \rangle - q_k + b_k (1 - \langle q \rangle) - w_k \langle q \rangle) \end{aligned} \tag{5}$$

Δm_k 的二阶矩为

$$\begin{aligned} E[(\Delta m_k)^2] &= \lambda n_k (1 - q_k - w_k)\langle q \rangle + \\ &\lambda n_k (q_k - b_k)(1 - \langle q \rangle) \end{aligned}$$

因此,Δm_k 的方差为

$$\begin{aligned} \mathrm{var}[\Delta m_k] &= E[(\Delta m_k)^2] - (E[\Delta m_k])^2 = \\ &\lambda n_k (1 - q_k - w_k)\langle q \rangle + \\ &\lambda n_k (q_k - b_k)(1 - \langle q \rangle) + o(\lambda n_k) \end{aligned}$$

由定义 $q_k = \frac{m_k}{n_k}$,可以得到系统中持有观点 +1 人数变化期望值和方差。

$$E[\Delta q_k] = \frac{1}{n_k} E[\Delta m_k] = \lambda (\langle q \rangle - q_k)$$

$$\mathrm{var}[\Delta q_k] = \frac{1}{n_k^2} \mathrm{var}[\Delta m_k] = \frac{\lambda}{n_k} \sigma_k^2$$

其中:$\sigma_k^2 = (1 - q_k - w_k)\langle q \rangle + (q_k - b_k)(1 - \langle q \rangle)$。注意到 Δq_k 的增加步长为 $1/n_k$,当系统规模非常大时,该变化是非常小的,因此由上面两个公式可以近似地得到 q_k 在单位时间内的变化。

$$\begin{aligned} dq_k &= \lambda (\langle q \rangle - q_k + b_k (1 - \langle q \rangle) - w_k \langle q \rangle)\mathrm{d}t + \\ &\sqrt{\frac{1}{n_k}} \sigma_k dB_t^{(k)} \end{aligned}$$

其中:$B_t^{(k)}$ 表示 k 个独立的布朗运动粒子。公式两侧分别进行加权平均可以得到

$$d\langle q \rangle = \lambda [b_k (1 - \langle q \rangle) - w_k \langle q \rangle]\mathrm{d}t + \langle \sqrt{\frac{\lambda}{n_k}} \sigma_k dB_t^{(k)} \rangle$$

由于布朗运动的平均近似可以忽略不计[8],因此有

$$\frac{d\langle q \rangle}{\mathrm{d}t} = \lambda [\langle b \rangle (1 - \langle q \rangle) - \langle w \rangle \langle q \rangle]$$

在系统的平衡状态,方程的右侧等于 0,因此有

$$\langle b \rangle (1 - \langle q_\infty \rangle) - \langle w \rangle \langle q_\infty \rangle = 0$$

则有

$$\langle q_\infty \rangle = \frac{\langle b \rangle}{\langle b \rangle + \langle w \rangle}$$

即系统的平衡状态中,两种观点比例关系由初始状态属于 X 和 Y 集合的人数比例决定。

下面考察什么状态下系统中 $E[\Delta q_k]$ 等于零。在具有周期边界的格子网络中,所有节点具有相同的度,因此 $\langle q \rangle = q_k$,则 $E[\Delta q_k] = 0$。对于非格子系统,需要考察参数 λ 的影响。假设参数 λ 依赖于节点的度信息,即 $\lambda = \lambda(k)$,则式(5)可写为

$$\begin{aligned} E[\Delta q_k] &= \frac{1}{n_k} E[\Delta m_k] = \\ &\lambda(k)(\langle q \rangle - q_k + b_k (1 - \langle q \rangle) - w_k \langle q \rangle) \end{aligned} \tag{6}$$

如果 $E[\Delta q_k] = 0$,那么有

$$\langle q \rangle = \frac{\sum_k \lambda(k)(q_k - b_k)}{\sum_k \lambda(k)(1 - b_k - w_k)} \tag{7}$$

由式(3),取 $\hat{p}_i = 1 - b_i - w_i$,$\hat{q}_j = \frac{(q_j - b_j)}{(1 - b_j - w_j)}$,可以发现

$$\lambda(k) = ck\hat{p}_k \tag{8}$$

是式(6)的一个特解。$\hat{p}_i$ 对应着 Z 集合中度为 i 的人群规模,$\hat{q}_j$ 对应着度为 j 且不属于 X 集合的人群规模。因此,当 λ 满足式(8),且 $\hat{p}_i$、$\hat{q}_j$ 满足式(7)条件的系统中的人数比例都将保持在一个稳定水平。对于度分布比较复杂的系统,$\lambda(k)$ 非常难以确定,因此本文只在二维格子网络上进行数值模拟,数值模拟的结果与解析结果基本吻合。

3 数值实验

生成 10×10 的具有周期边界的格子网络(格子节点的边界节点与另一侧边界的节点也进行连接),节 (下转第 4112 页)