文章编号: 1001-0920(2017)09-1647-06 DOI: 10.13195/j.kzyjc.2016.0856

一种基于改进灰色关联分析的变量选择算法

韩 敏†, 张瑞全, 许美玲

(大连理工大学 电子信息与电气工程学部, 辽宁 大连 116024)

摘 要: 针对灰色绝对关联度模型和灰色相似关联度模型存在的问题,提出一种基于相对变化面积的改进灰色关联度模型. 以序列几何形状的相似程度为基础,构建反应折线相似程度的相对变化面积,并以此作为关联系数的计算依据,同时以局部关联度的平均值度量整体的相似性,定义灰色关联度模型. 此外,根据关联度计算结果,提出一种基于集合思想的变量选择算法,有效去除变量间的无关和冗余变量. 仿真结果验证了所提出算法的有效性和合理性.

关键词: 变量选择; 灰色关联分析; 灰色相似关联度; 灰色绝对关联度

中图分类号: TP183 **文献标志码:** A

A variable selection algorithm based on improved grey relational analysis

HAN Min†, ZHANG Rui-quan, XU Mei-ling

(Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China)

Abstract: An improved grey relation model based on relative area change is proposed for the deficiency of the similitude degree of grey incidence and the absolute degree of grey incidence. Depending on the similitude degree of sequence curves, the relative area change that reflects the similitude degree of the sequence curves is constructed, and then the new relational coefficient is defined. At the same time, the new grey relation model is defined based on the mean value of the relational coefficients. Furthermore, according to the result of the relational analysis, a variable selection method is also proposed in order to reduce the irrelevant and redundant variables. The simulation results show the rationality and the effectiveness of the proposed algorithm.

Keywords: variable selection; grey relational analysis; the similitude degree of grey incidence; the absolute degree of grey incidence

0 引 言

时间序列是当今科学研究领域中的一个研究热点,其广泛存在于水文、气象、经济、电子等众多领域[1],时间序列的普遍存在决定了对其研究的必要性.

与单元时间序列相比,多元时间序列包含了更多的系统信息,以多元时间序列为基础建立的预测模型能够获得更加理想的预测精度. 但是,多元时间序列本身也存在着许多不可忽视的局限性. 多元时间序列之间往往具有复杂多变的关联特性,如果不对这些复杂的关联特性进行合理利用,非但不能提高模型的预测性能,反而会增大模型规模,导致预测精度降低,产生“维数灾难”[2]. 因此,对多元时间序列之间的相关性进行分析,进而选择合适的输入变量,降低输入变量维数是十分必要的.

相关性一般用来描述两个随机变量之间的密切程度. 目前,变量间相关性分析方法仍然以统计学分析方法为主,典型的相关性分析方法主要包括Granger因果关系分析[3]、Copula分析[4]、互信息[5]以及灰色关联分析[6]等方法. 其中,传统的Granger因果分析只能对变量进行定性的分析,难以给出定量的描述,且不能直接应用于非线性系统;Copula分析对于分布不规则的数据,很难寻找到合适的边缘分布和Copula函数;互信息虽然对数据的分布类型没有要求,可以度量变量间任意类型的关系,但对于高维数据,计算复杂度高;而灰色关联分析是一种定性与定量相结合的方法,计算简单,对样本量和样本分布规律无特殊要求. 因此,灰色关联分析已被广泛应用到

收稿日期: 2016-07-04; **修回日期:** 2016-10-05.

基金项目: 国家自然科学基金项目(61374154).

作者简介: 韩敏(1959–), 女, 教授, 博士生导师, 从事控制理论、神经网络预测等研究; 张瑞全(1991–), 男, 硕士生, 从事时间序列相关性分析的研究.

†通讯作者. E-mail: minhan@dlut.edu.cn

各个领域[7].

灰色关联分析[6]最早由邓聚龙教授提出,此后众多学者对其进行了系统的研究,并取得了许多显著的成果[8]. 从整体而言,灰色关联分析模型可以分为3类:基于距离的度量模型,基于斜率的度量模型和基于面积的度量模型. 其中,灰色绝对关联度模型[9]和灰色相似关联度模型[10]是两种典型的基于面积的度量模型,但这两种模型均是基于有向面积的,当比较序列围绕着参考序列存在上下波动时,在积分过程中会出现正负面积相互抵消的现象,使得关联分析的准确性大大降低. 为此,王靖程等[11]和刘震等[12]均利用先取绝对值后积分的思想,有效地克服了正负面积相互抵消的问题,但这两种方法均需要判断序列曲线是否相交,然后根据是否相交采用不同的计算公式计算关联系数,计算相对复杂. 此外,王靖程等提出的算法中,当改变比较序列组中的一个比较序列时,其他比较序列的关联度计算结果也会随之发生改变,且对于不同的比较序列,ρ的取值也不尽相同;刘震等提出的算法不能有效地反映序列几何形状的接近性,且只能衡量变量间的正相关关系,此外,当两序列平行时,其不满足规范性原则.

基于上述分析,本文提出一种基于相对变化面积的灰色关联度模型,以序列几何形状的相似程度为基础,通过序列曲线的相对变化面积计算序列间的关联系数;同时,以局部关联度的平均值度量整体的相似性,计算关联度. 此外,根据关联度分析结果,提出一种基于集合思想的变量选择算法,有效去除变量间的无关和冗余变量.

1 灰色关联分析

灰色关联分析是灰色系统理论的重要组成部分,其基本思想是通过比较各比较序列与参考序列间序列曲线几何形状的相似程度来判断序列间的关联程度,几何形状越相似,关联度越大.

灰色绝对关联度模型是一类基于有向面积的关联度度量模型,其可以有效地度量变量间的关联程度. 设存在参考序列$X_0=[x_0(1),\cdots,x_0(n)]$和比较序列$X_1=[x_1(1),\cdots,x_1(n)]$,使用始点零化算子$D$对序列进行无量纲处理,得到序列始点零化像,如下所示:

$$\begin{aligned}&X_0^0=[x_0^0(1),\cdots,x_0^0(n)]=[x_0(1)d,\cdots,x_0(n)d],\\&X_1^0=[x_1^0(1),\cdots,x_1^0(n)]=[x_1(1)d,\cdots,x_1(n)d],\end{aligned}\tag{1}$$

其中$x_i^0(k)=x_i(k)d=x_i(k)-x_i(1)$. 进而,$X_0^0$和$X_1^0$的灰色关联度可表示为

$$\gamma(X_0^0,X_1^0)=\frac{1+|s_0|+|s_1|}{1+|s_0|+|s_1|+|s_1-s_0|}.\tag{2}$$

其中

$$\begin{aligned}&|s_0|=\Big|\sum_{k=2}^{n-1}x_0^0(k)+\frac{1}{2}x_0^0(n)\Big|,\\&|s_1|=\Big|\sum_{k=2}^{n-1}x_1^0(k)+\frac{1}{2}x_1^0(n)\Big|,\\&|s_1-s_0|=\\&\Big|\sum_{k=2}^{n-1}(x_1^0(k)-x_0^0(k))+\frac{1}{2}(x_1^0(n)-x_0^0(n))\Big|.\end{aligned}$$

虽然灰色绝对关联度模型得到了较为广泛的应用,但仍存在许多不可避免的局限性[12]. 为此,李思峰等[10]从相似性角度出发,提出了灰色绝对关联度模型的改进模型——灰色相似关联度模型. 该模型从相似性视角度量序列间的相关性,通过比较序列曲线几何形状的相似程度,判断序列间的关联程度,两序列几何形状越相似,关联度越大.

灰色相似关联度可表示如下:

$$\gamma(X_0^0,X_1^0)=\frac{1}{1+|s_1-s_0|}.\tag{3}$$

灰色相似关联度模型不仅满足偶对称性、规范性和接近性原则,而且关联度的大小只与序列曲线的几何形状有关,与其相对空间位置无关.

2 基于改进灰色关联分析的变量选择算法

虽然灰色绝对关联度模型和灰色相似关联度模型得到了较为广泛的应用,但这两种模型均是基于有向面积的方法,当积分过程中出现正负面积相互抵消时,关联度的计算结果往往与定性分析不符. 为此,本文从曲线相似性角度出发,通过两曲线的相对变化面积,构建反应两曲线相似程度的相对面积变化比,从而定义一种新的灰色关联模型,该模型有效地解决了原始算法中可能存在的正负面积相互抵消的问题. 同时,两序列的相似程度只与曲线的几何形状有关,与其空间相对位置等均无关.

2.1 改进灰色关联度模型

定义1 设存在序列$X_i=[x_i(1),\cdots,x_i(n)]$,其中$i$表示序列标号,$n$表示样本量,则$X_i$的始点零化像可以表示为

$$X_i^0=X_iD=[x_i^0(1),\cdots,x_i^0(n)].\tag{4}$$

其中:$x_i^0(k)=x_i(k)d=x_i(k)-x_i(1)$,$D$为始点零化算子.

定义2 设序列X_i^0为1-时距序列,则序列X_i^0在区间$[k,k+1]$上所对应的折线可表示为$X_i^0(t)$,其中

$t \in [k, k+1], k = 1, 2, \cdots, n-1$.

定义3　设存在序列X_i^0,则折线$X_i^0(t)$在区间$[k, k+1]$上的面积变化量可以表示为

$$\Delta s_i(k) = \int_k^{k+1} X_i^0(t) - X_i^0(k)\mathrm{d}t. \tag{5}$$

进一步,在区间$[k, k+1]$上,可将式(5)的积分看作求取直角三角形的面积. 其中,该直角三角形的一条直角边长度为1,则式(5)可进一步表示为

$$\Delta s_i(k) = \frac{1}{2}(x_i^0(k+1) - x_i^0(k)),\ k = 1, 2, \cdots, n-1. \tag{6}$$

定义4　设存在序列X_i^0和X_j^0,$X_i^0(t)$和$X_j^0(t)$在区间$[k, k+1]$上的面积变化量记为$\Delta s_i(k)$和$\Delta s_j(k), k = 1, 2, \cdots, n-1$,则称

$$\gamma_{i,j}(k) = \begin{cases} \mathrm{sgn}(\Delta s_i \cdot \Delta s_j)\dfrac{\min||\Delta s_i|, |\Delta s_j||}{\max||\Delta s_i|, |\Delta s_j||},\ \Delta s_i \cdot \Delta s_j \neq 0; \\ 0,\ \Delta s_i \cdot \Delta s_j = 0 \end{cases} \tag{7}$$

为序列X_i^0和X_j^0的关联系数. 其中:

1) 若$\Delta s_i(k) \cdot \Delta s_j(k) > 0$,则关联系数为正,即$X_i^0(t)$与$X_j^0(t)$在区间$[k, k+1]$上的面积变化趋势相同,序列间存在正相关;

2) 若$\Delta s_i(k) \cdot \Delta s_j(k) < 0$,则关联系数为负,即$X_i^0(t)$与$X_j^0(t)$在区间$[k, k+1]$上的面积变化趋势相反,序列间存在负相关;

3) 若$\Delta s_i(k) \cdot \Delta s_j(k) = 0$,则序列不相关.

进一步,序列X_i^0和X_j^0的关联度可以表示为

$$\gamma(X_i^0, X_j^0) = \frac{1}{n-1}\sum_{k=1}^{n-1}\gamma_{i,j}(k). \tag{8}$$

由上述定义可知,采用相对变化面积之比的形式定义关联系数,使得关联系数的大小、正负只与序列曲线的几何形状有关,与其空间相对位置无关. 因此,本文算法可以有效地克服原算法积分过程中正负面积相互抵消的现象. 同时,与文献[11]相比,利用相对变化面积之比的形式不仅能够更加形象地反映序列曲线几何形状的相似程度,而且还可以有效地度量变量间的负相关关系.

假设1　改进灰色关联模型满足规范性原则.

证明　由改进灰色关联模型关联系数的定义式可知$0 \leqslant |\gamma_{i,j}(k)| \leqslant 1$,当且仅当两序列曲线的面积变化量绝对值相同时,$|\gamma_{i,j}(k)| = 1$. 灰色关联度是以局部关联度的平均值度量整体的相似性,因此,改进的灰色关联模型满足规范性原则. □

假设2　改进灰色关联模型满足偶对称性原则.

证明　对于给定的序列X_i^0和X_j^0,由关联系数$\gamma_{ij}(k) = \gamma_{ji}(k)$可得,$\gamma(X_i^0, X_j^0) = \gamma(X_j^0, X_i^0)$,因此,改进的灰色关联模型满足偶对称性原则. □

假设3　改进灰色关联模型满足接近性原则.

证明　由关联系数定义式可知,两序列曲线的几何形状越接近,面积变化量的比值越大,关联度越大. 因此,改进的灰色关联模型满足接近性原则. □

假设4　改进灰色关联模型关联度的大小只与序列曲线的几何形状有关,与其空间相对位置和距离等无关.

证明　由改进灰色关联模型关联度的定义式可知,关联度的大小、正负只与序列曲线的相对变化面积有关,而相对变化面积只与序列曲线的几何形状有关. 因此,改进灰色关联模型关联度的大小只与序列曲线的几何形状有关,与其空间相对位置和距离等无关. □

2.2　基于改进灰色关联分析的变量选择算法

对于多元时间序列,其高维特性在带来更多信息的同时,也会产生无关和冗余变量. 而无关和冗余信息的产生不仅增大了后续模型的规模,而且还降低了模型的预测精度和效率. 因此,合理地分析变量间的相互关系,进而选择合适的变量维数是十分必要的.

目前,基于灰色关联分析的变量选择算法主要思想是: 首先使用灰色关联分析对变量进行相关性排序; 然后根据预先设定的关联度阈值或变量个数对变量进行选择[13-14]. 但是,由灰色关联分析的总体性和非唯一性可知: 灰色关联分析的重点是排出序列间的关联序,而非得到具体的关联度数值; 关联度数值随着数据无量纲处理方法的不同、参考序列和比较序列的不同而不同. 因此,单纯地依靠人为设定的关联度阈值或变量个数进行变量选择具有很大的局限性和不确定性.

此外,文献[10]指出,灰色关联度是一种“孤立”的关联度,它只能反映单一的比较序列与参考序列间的关联程度. 因此,若只根据自变量(比较序列)与因变量(参考序列)间的相关性,没有考虑自变量间的相关性,单纯利用一个阈值进行变量选择,则可能会删除有用信息或导致冗余的产生. 为此,文献[15] 提出一种组合式的变量选择方法,有效地衡量了自变量间的相关性,但其需要遍历所有可能的变量子集,计算复杂度高,应用较为困难.

文献[16]指出,根据变量间的相关性强弱,可以将变量划分为强相关变量、弱相关变量和无关变量,而冗余信息是由弱相关变量引入的,其可以被视为弱相关变量中的一类.

基于上述分析,利用集合思想,本文提出一种前向变量选择算法.具体的算法步骤如下.

1) 相关性排序. 利用改进的灰色关联模型对变量进行相关性分析,并按关联度降序方式对变量进行排序,得到变量集合$S=\{s(1),s(2),\cdots,s(n)\}$.

2) 最优子集选择. 由上述分析可知,冗余信息是由弱相关变量产生的,而冗余和无关变量的引入会导致模型预测精度的降低. 因此,本文采用前向选择法,对变量进行选择.

i) 最优子集初始化,即$\Gamma=\{s(1)\}$,同时令$i=1$;

ii) 使用最优子集Γ进行预测,得到预测误差$\varepsilon(i)$;

iii) 更新最优子集$\Gamma=\Gamma+\{s(i+1)\}$,同时使用更新后的最优子集进行预测,得到预测误差$\varepsilon(i+1)$;

iv) 比较$\varepsilon(i)$与$\varepsilon(i+1)$之间的大小,若$\varepsilon(i)>\varepsilon(i+1)$,则认为第$i+1$个变量为有效变量,保留第$i+1$个变量,同时令$i=i+1$,返回步骤2);否则,令$\Gamma=\Gamma-\{s(i+1)\}$,结束循环.

3 仿真结果及分析

为验证本文所提出改进灰色关联模型的有效性,分别利用Friedman数据集和Gas furnac数据集进行仿真. 同时基于改进关联度模型的分析结果,将变量选择算法应用于大连市气象数据.

3.1 Friedman数据集

Friedman数据集是一种广泛使用的标杆数据集,其数学模型如下:

$$Y=10\sin(\pi x_1x_2)+20(x_3-0.5)^2+10x_4+5x_5+\varepsilon. \tag{9}$$

其中:$x_1,\cdots,x_5$表示相关变量,$x_6,\cdots,x_{10}$表示无关变量,$x_i(i=1,2,\cdots,10)$均为[0,1]之间的均匀分布,ε为标准正态分布的噪声.

将Y作为参考序列,$x_1,\cdots,x_{10}$作为比较序列,分别利用邓式关联度模型[6]、灰色绝对关联度模型[9]、灰色相似关联度模型[10]、文献[11]模型和本文所提出关联度模型对数据进行相关性分析,关联分析结果如表1所示.

由表1可知,在本文所提出模型的关联序中,前5维变量均为相关变量,而在其他几种模型的关联序中,前5维变量均包含无关变量,这说明本文所提出模型可以更为有效地度量变量间的相关性.

表 1 Friedman数据集关联度排序结果

算法	关联度排序
邓氏关联度[6]	$x_4,x_5,x_8,x_6,x_1,x_{10},x_7,x_9,x_2,x_3$
灰色绝对关联度[9]	$x_8,x_4,x_5,x_6,x_1,x_{10},x_7,x_9,x_2,x_3$
灰色相似关联度[10]	$x_8,x_4,x_5,x_6,x_1,x_{10},x_7,x_9,x_2,x_3$
文献[11]模型	$x_4,x_8,x_5,x_6,x_1,x_{10},x_2,x_7,x_9,x_3$
本文所提出模型	$x_2,x_4,x_1,x_3,x_5,x_8,x_7,x_{10},x_9,x_6$

3.2 Gas furnace数据集

Gas furnace数据来源于UCI数据集,是一组常用的多元时间序列分析数据,其输入为气体速率$u(t)$,输出为CO2的百分比浓度$y(t)$. 对$u(t)$和$y(t)$进行相空间重构,设置延迟时间为1,嵌入维数为6和4,则可以得到10维输入变量

$$X=\{u(t-6),\cdots,u(t-1),y(t-4),\cdots,y(t-1)t\}. \tag{10}$$

目标输出变量为$y(t)$,共计290组数据.分别利用不同的灰色关联模型对输入输出数据进行相关性分析,并按相关性强弱对变量进行降序排列. 为了验证不同灰色关联模型的分析结果,利用本文的变量选择算法对各排序结果进行最优子集选择和预测,其中选取极端学习机作为神经网络模型,前200组数据作为训练数据(前150组用于变量相关性排序,后50组用于最优子集选择),剩余90组数据作为测试数据. 引入均方根误差(RMSE)定量评价预测性能,其定义如下:

$$\mathrm{RMSE}=\Big(\frac{1}{n-1}\sum_{i=1}^{n}|\hat{y}_i-y_i|^2\Big)^{1/2}. \tag{11}$$

其中:n为样本量,y_i为观测值,$\hat{y}_i$为预测值. 具体仿真结果如表2所示.

由表2可知,使用同一种变量选择算法对不同灰色关联模型变量排序结果进行变量选择和预测时,基于本文所提出关联模型变量排序结果的最优子集可以得到最好的预测精度,这说明本文所提出灰色关联度模型可以更加有效地分析变量间的相关性.同时

表 2 Gas furnace数据集关联度排序结果

算法	最优子集	RMSE
邓氏关联度[6]	$y(t-1),y(t-2),y(t-3),y(t-4),\ u(t-1),u(t-2),u(t-3)$	0.306 6
灰色绝对关联度[9]	$u(t-1),y(t-1),u(t-2),\ y(t-2),u(t-3)$	0.320 7
灰色相似关联度[10]	$u(t-1),u(t-2),u(t-3),\ u(t-4),y(t-1)$	0.317 3
文献[11]模型	$y(t-1),y(t-2),y(t-3),y(t-4),\ u(t-1),u(t-2),u(t-3)$	0.306 6
本文所提出模型	$y(t-1),u(t-4),u(t-5),\ y(t-2),u(t-3)$	0.288 1

对比其他几种灰色关联模型的最优子集个数,本文所提出模型的最优子集个数为 5, 小于邓式关联度模型和文献[14]中关联度模型的最优子集个数,这也说明本文所提出的关联度模型可以更加有效地对变量间的相关性进行分析,进而使得在后续变量选择过程中可以使用最少的变量来近似表示整体变量.

3.3　大连市气象数据集

为了进一步验证本文算法的有效性,将其应用到大连市气象数据集的仿真预测中. 大连市气象数据集包含1951-01-01～2010-07-01共计715个月的当地月平均风速、日照百分比、气压、降雨量、相对湿度和气温等六维变量. 选取气温Y作为因变量,即参考序列,其他5维变量依次记为x_1、x_2、x_3、x_4和x_5,作为自变量,即比较序列. 同时,选取前530组数据作为训练数据集,其中前390组数据作为训练子集,用于变量的相关性排序,后140组数据作为确认子集,用于最优子集选择;剩余185组数据作为测试数据集,用于最优子集的评价.

根据先验信息,对变量进行定性分析可知,降雨量、湿度与气温之间存在较强的正相关关系,气压与气温之间存在较强的负相关关系,风速与气温之间存在较弱的负相关关系;而日照百分比基本不受温度影响.

利用不同的灰色关联模型对数据进行定量分析,结果如表3所示,其中γ_1、γ_2、γ_3、γ_4和γ_5分别表示x_1、x_2、x_3、x_4和x_5与Y之间的关联度大小,灰色相似关联度的数值单位为1e-04,其他均为1e-0.

表 3　大连市气象数据集关联度分析结果

算法	γ_1	γ_2	γ_3	γ_4	γ_5
邓氏关联度[6]	0.700 6	0.712 3	0.724 0	0.732 5	0.672 6
灰色绝对关联度[9]	0.535 0	0.524 7	0.500 0	0.500 0	0.751 3
灰色相似关联度[10]	0.956 9	0.965 5	0.430 4	0.868 4	0.927 4
文献[11]模型	0.895 7	0.894 5	0.792 5	0.883 1	0.811 0
本文所提出模型	−0.035 5	−0.033 8	−0.542 5	0.074 1	0.219 9

由表3可知,本文算法不仅可以度量变量间的正相关关系,还可以度量变量间的负相关关系,而其他几种算法对所有变量均作出了正相关的判断,这明显与定性分析不符. 同时,由定性分析可知,风速和日照百分比对气温的影响较小,相关性较弱,而前3种算法均作出了相关性较大的判断. 此外,由表3可知,不同关联度模型计算出的关联度相差较大,因此单纯地利用阀值确定变量个数的方法是不可靠的. 对比其他几种关联度分析结果,本文算法的关联度计算结果相对较小,这是由关联度计算公式和输入数据导致的. 本文算法的关联度分析结果完全取决于相对变化面积,而在其他几种算法中,相对变化部分在比值中所占比例较小,对结果的影响较弱.

为了更好地对变量进行选择,基于本文所提出的改进灰色关联度模型的分析结果,利用前向变量选择算法,剔除变量中的无关和冗余变量,其中选取极端学习机作为神经网络模型. 最优子集的选择结果如图1所示,具体预测误差如表4所示.

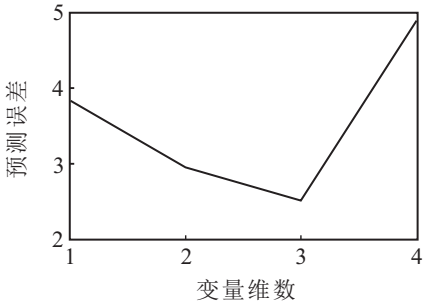

图 1　最优子集选择结果

表 4　使用不同维数变量的具体预测误差

变量维数	1	2	3	4
RMSE	3.729 9	2.860 2	2.368 2	4.765 3

由图1和表4可知,当加入第4维变量后,预测误差明显增大,根据所提出变量选择算法,停止后续变量的选择,选择最优子集为前3维变量. 进一步,为了验证最优子集的性能,使用测试数据集进行测试. 同时,为了验证本文算法的有效性,将基于本文提出的灰色关联分析结果的阀值法[13-14]、文献[15]提出的算法以及基于mRMR关联分析的变量选择算法[16]应用于仿真实验中,其中关联度阀值δ分别选取为0.100 0、0.300 0和0.500 0.

表 5　大连市气象数据气温预测结果

算法	阀值	最优子集	RMSE
阀值法	0.100 0	x_3、x_5	2.935 4
	0.300 0	x_3	3.830 0
	0.500 0	x_3	3.830 0
文献[15]	0	x_3、x_5、x_4	2.546 2
mRMR[16]	0	x_3、x_4、x_1、x_5	2.751 7
本文变量选择算法	0	x_3、x_5、x_4	2.546 2

由表5可知: 阀值法具有很大的随机性和不确定性,即对于不同的阀值,变量选择结果可能是相同的,且所选变量不能保证预测结果最优;文献[15]所提出算法虽然得到了与本文所提出算法相同的结果,但其需要遍历所有可能,计算量大,复杂度高; 而基于mRMR关联分析的变量选择算法,其选择出的最优子集变量个数为4,大于本文算法的最优子集个数,且

其预测误差也明显增大. 此外,文献[15]遍历所有可能后的选择结果与本文算法相同,表明了本文算法的有效性.

4 结 论

本文从相似性角度出发,通过相对变化面积构建灰色关联度模型,同时提出一种前向的变量选择算法,有效地去除了无关和冗余变量.与其他算法相比,本文算法具有以下特点:

1) 以相对变化面积比的形式代替面积差的形式,使所提出模型不仅可以充分地反映变量间的正负相关性,而且能有效地克服原模型积分过程中正负面积相互抵消的问题;

2) 所提出模型关联度大小只与序列曲线的几何形状有关,与相对空间位置和距离等均无关,更符合灰色关联度的定义;

3) 基于集合思想,利用前向选择方法进行变量选择,以预测误差作为指标,明确了变量选择的依据,有效克服了传统变量选择算法的不确定性.

参考文献(References)

[1] Miranian A, Abdollahzade M. Developing a local least-squares support vector machines-based neuro-fuzzy model for nonlinear and chaotic time series prediction[J]. IEEE Trans on Neural Networks and Learning Systems, 2013, 24(2): 207-218.

[2] Autin F, Claeskens G, Freyermuth J M. Hyperbolic wavelet thresholding methods and the curse of dimensionality through the maxiset approach[J]. Applied and Computational Harmonic Analysis, 2014, 36(2): 239-255.

[3] Chen Y, Rangarajan G, Feng J, et al. Analyzing multiple nonlinear time series with extended Granger causality[J]. Physics Letters A, 2004, 324(1): 26-35.

[4] Patton A J. Copula methods for forecasting multivariate time series[M]. Amsterdam: Elsevier, 2013: 899-960.

[5] Fiedor P. Networks in financial markets based on the mutual information rate[J]. Physical Review E, 2014, 89(5): 052801.

[6] Julong D. Introduction to grey system theory[J]. J of Grey System, 1989, 1(1): 1-24.

[7] Liu S, Forrest J, Yang Y. A brief introduction to grey systems theory[J]. Grey Systems: Theory and Application, 2012, 2(2): 89-104.

[8] Liu S, Yang Y, Cao Y, et al. A summary on the research of GRA models[J]. Grey Systems: Theory and Application, 2013, 3(1): 7-15.

[9] Liu S F, Fang Z G, Lin Y. Study on a new definition of degree of grey incidence[J]. J of Grey System, 2006, 9(2): 115-122.

[10] 刘思峰, 谢乃明, Forrest Jeffery. 基于相似性和接近性视角的新型灰色关联分析模型[J]. 系统工程理论与实践, 2010, 30(5): 881-887.
(Liu S F, Xie N M, Forrest Jeffery. On new models of grey incidence analysis based on visual angle of similarity and nearness[J]. Systems Engineering — Theory & Practice, 2010, 30(5): 882-887.)

[11] 王靖程, 诸文智, 张彦斌. 基于面积的改进灰关联度算法[J]. 系统工程与电子技术, 2010, 32(4): 777-779.
(Wang J C, Zhu W Z, Zhang Y B. Improved algorithm of grey incidence degree based on area[J]. Systems Engineering and Electronics, 2010, 32(4): 777-779.)

[12] 刘震, 党耀国, 周伟杰, 等. 新型灰色接近关联模型及其拓展[J]. 控制与决策, 2014, 29(6): 1071-1075.
(Liu Z, Dang Y G, Zhou W J, et al. New grey nearness incidence model and its extension[J]. Control and Decision, 2014, 29(6): 1071-1075.)

[13] 苏博, 刘鲁, 杨方廷. 基于灰色关联分析的神经网络模型[J]. 系统工程理论与实践, 2008, 28(9): 98-104.
(Su B, Liu L, Yang F T. Research of artificial neural network forecasting model based on grey relational analysis[J]. Systems Engineering — Theory & Practice, 2008, 28(9): 98-104.)

[14] 钱晓山, 阳春华. 基于灰关联分析的 KFCM-LSSVM 蒸发过程软测量[J]. 控制与决策, 2012, 27(12): 1800-1804.
(Qian X S, Yang C H. Soft sensor of based on grey correlation analysis and KFCM-LSSVM in evaporation process[J]. Control and Decision, 2012, 27(12): 1800-1804.)

[15] Song Q, Shepperd M. Predicting software project effort: A grey relational analysis based method[J]. Expert Systems with Applications, 2011, 38(6): 7302-7316.

[16] 韩敏, 刘晓欣. 基于互信息的分步式输入变量选择多元序列预测研究[J]. 自动化学报, 2012, 38(6): 999-1006.
(Han M, Liu X X. Stepwise input variable selection based on mutual information for multivariate forecasting[J]. Acta Automatica Sinica, 2012, 38(6): 999-1006.)

（责任编辑：孙艺红）