

## 基于灰色理论的铁路客运量预测影响因素优化

王文莉, 杨俊红

(郑州铁路职业技术学院 信息工程系, 河南 郑州 450052)

**摘要:** 为了更好地反映各种相关因素对客运量的影响, 实现铁路客运量预测模型中影响因素的优化选择, 采用灰色理论对影响因素进行分析, 并针对传统灰关联分析在具体应用中存在的关联评价值趋于均匀化、分辨系数取值影响排序结果等不足, 提出一种采用动态分辨系数的铁路客运量灰关联分析方法, 从而得到各因素对客运量的关联度, 实现铁路客运量预测模型中影响因素的优化选择。仿真实验以河南省铁路客运量为例, 结果表明预测精度得到了提高, 此方法可行并且有效。

**关键词:** 铁路客运量预测; 动态分辨系数; 影响因素优化选择; 灰关联分析

**中图分类号:** TP391.9

**文献标识码:** A

**文章编号:** 1000-7180(2011)10-0164-04

## Optimal Selection of Factors Influencing Grey-theory-based Forecast of Railway Passenger Traffic Volume

WANG Wen-li, YANG Jun-hong

(Department of Information Engineering, Zhengzhou Railway Vocational & Technical College, Zhengzhou 450052, China)

**Abstract:** In order to better unfold the influences of related factors on railway passenger traffic volume to optimize the selection of factors influencing railway passenger traffic forecast modeling, this paper analyzes the influencing factors by the grey theory and introduces a novel grey correlation analysis approach to railway passenger traffic volume with dynamic resolution coefficients, considering the weaknesses of traditional correlation analysis embodied in specific applications such as relation appraisal tending to be equalized and sorting results subject to the impact of discrimination coefficient value. Hence the relations between diverse factors and passenger traffic volume are uncovered, which can help to optimize the selection of factors influencing railway traffic volume forecast modeling. With the railway passenger traffic volume in Henan Province taken for instance, the simulation results testify that the accuracy of forecast has been increased, proving that the approach adopted here is feasible and effective.

**Key words:** forecast of railway passenger traffic volume; dynamic resolution coefficient; optimal selection of influencing factors; grey correlation analysis

### 1 引言

铁路运输是为国民经济服务的产业, 通过对客运量各影响因素的分析与把握, 实现铁路客运量的预测, 从而为铁路部门实现客流宏观调控提供参考依据。同时由于影响客运量的因素很多, 这就需要对这些相关影响因素进行优化选择, 才能提高客运量预测的精度。

灰关联分析方法是灰色理论的重要组成部分, 是一种多因素统计分析方法, 是以各因素的样本数据为依据, 根据因素之间发展态势的相似或相异程度衡量因素间的接近程度。相对于传统的回归分析、主成分分析等方法, 灰关联分析法具有对样本要求低、计算量小等优点, 成功地用于经济、生态、农业等不确定系统。但传统的灰关联分析法在具体应用中还存在着关联评价值趋于均匀化, 分辨系数取值影

收稿日期: 2011-04-20; 修回日期: 2011-06-18

基金项目: 河南省教育厅自然科学研究指导计划项目(2009C520010)

响结果排序等不足,反映在灰关联系数表达式和关联度计算公式上<sup>[1]</sup>. 本文通过分析分辨系数对关联系数的影响,并从信息论的角度出发,提出一种新的采用动态分辨系数的区域铁路客运量灰关联分析方法,分析各因素对铁路客运量的影响程度.

## 2 指标选取

依据现有研究<sup>[2-4]</sup>,可将铁路客运量影响因素归纳为:经济体制、国民生产总值、人口数量、居民平均消费水平、运价、工商业发展、运输服务水平、运输方式结构八项因素. 同时由于预测的提前性,有些无法预料到的因素却造成实际值波动剧烈,例如 2003 年我国爆发的 SARS 病症. 由于条件的限制和不可预见性,在预测中就很少考虑到这方面因素对铁路客运运营造成的影响.

从指标选取的代表性、可获得性等原则出发,拟选取以下 8 个指标,从不同侧面寻找铁路客运量的主要影响因素(表 1).

表 1 铁路客运量影响因素评价指标简表

| 序号 | 指标        | 表征含义       |
|----|-----------|------------|
| 1  | 国民生产总值    | 区域综合经济发展水平 |
| 2  | 人口数量      | 区域人口数量     |
| 3  | 居民平均消费水平  | 居民消费能力     |
| 4  | 工业企业增加值   | 工业发展程度     |
| 5  | 社会消费品零售总额 | 商业发展程度     |
| 6  | 旅客列车旅行速度  | 运输服务水平     |
| 7  | 公路客运量     | 运输方式结构状况   |
| 8  | 民航客运量     | 运输方式结构状况   |

其中铁路内部经济体制的变革将促使新型票价体制的调整,导致运输服务水平和交通运输结构的调整,经济体制和运价作为间接因素体现.

## 3 动态分辨系数的灰关联分析算法设计

设参考数列为  $\{X_0(t)\} = \{X_{01}, X_{02} \cdots, X_{0n}\}$ , 与参考数列作关联度比较的比较数列为

$$\{X_1(t), X_2(t), \cdots, X_m(t)\}$$
$$= \begin{Bmatrix} X_{11} & X_{12} \cdots X_{1n} \\ X_{21} & X_{22} \cdots X_{2n} \\ \cdots, & \\ X_{m1} & X_{m2} \cdots X_{mn} \end{Bmatrix}$$

其中:  $n$  为数列的数据个数,  $m$  为参考数列的个数. 于是, 第  $i$  个比较数列  $X_i(t)$  与参考数列  $X_0(t)$  在时刻  $t$  的关联系数可以表示为

$$\gamma(x_0(t), x_i(t)) = \frac{\min + \xi \max}{\Delta_i(t) + \xi \max} \tag{1}$$

$$\begin{aligned} \text{式中, } \min &= \min_i \min_t |x_0(t) - x_i(t)| \\ \max &= \max_i \max_t |x_0(t) - x_i(t)| \\ \Delta_i(t) &= |x_0(t) - x_i(t)| \end{aligned}$$

式中,  $\xi$  为分辨系数.

关联度以关联系数为基础, 分辨系数  $\xi$  是直接影响关联系数的一个因子, 它的取值直接决定着关联系数的分布区间. 如果关联度分布区间非常小, 那么即使能得到关联度的大小顺序, 也会由于关联度之间的差异很小而不能很清楚的区别序列之间的相似性. 所以在灰关联分析算法设计中, 关键是要确定  $\xi$  的取值.

### 3.1 $\xi$ 的取值原则

在实际应用中, 关于  $\xi$  的取值一直没有一个量化的方法, 而是延用大多数文献中的方法, 取  $\xi = 0.5$ . 但是现在许多研究都表明, 这种取值方式并不一定符合实际情况, 甚至可能会改变关联度的大小排序, 造成分析结果出现错误<sup>[6]</sup>.

下面从序列是否平稳的角度说明  $\xi$  的取值对关联度分布区间的影响. 通过对数据的初始化处理, 使  $\min_i \min_t |x_0(t) - x_i(t)| = 0$ , 那么式(1) 可变为

$$\gamma(x_0(t), x_i(t)) = \frac{\xi \max}{\Delta_i(t) + \xi \max} \tag{2}$$

$$\text{令 } \theta_i(t) = \frac{\Delta_i(t)}{\max} \tag{3}$$

则式(2) 变为

$$\gamma(x_0(t), x_i(t)) = \frac{\xi}{\theta_i(t) + \xi} \tag{4}$$

当序列不平稳时, 经常出现  $\max \gg \Delta_i(t)$  的情况, 这时  $\theta_i(t)$  非常小, 若  $\xi$  取较大值, 例如取  $\xi = 1$ , 则  $\gamma(x_0(t), x_i(t))$  完全受  $\xi$  支配,  $\gamma(x_0(t), x_i(t))$  的值接近 1, 因此通过式(4) 求得的关联系数非常接近, 关联系数的分布区间非常小, 很难分辨参考序列与比较序列的相似性, 这种情况在实际中经常发生的, 例如, 在将灰色关联度分析应用于机械故障诊断时, 待检信号在采集时往往会混入随机干扰而出现信号突变. 因此对于这种情况, 在实际应用时, 应尽量减小  $\xi$  的值, 削弱  $\max$  对  $\gamma(x_0(t), x_i(t))$  支配.

另一种情况, 当序列非常平稳时,  $\max$  与  $\Delta_i(t)$  差别并不是很大, 这时  $\theta_i(t)$  较大, 若  $\xi$  取很小的值, 则  $\gamma(x_0(t), x_i(t))$  较小而且同样很接近, 因此, 关联系数的值也会很接近, 分布区间很小, 很难分辨参考

序列与比较序列的相似性. 所以, 在这种情况下,  $\xi$  应该取较大的值, 增加  $\max$  对  $\gamma(x_0(t), x_i(t))$  的作用, 充分体现关联系数的整体性.

通过以上分析可知, 在实际中对  $\xi$  的取值的应该遵循以下原则:

(1)  $\xi$  应根据实际观测序列的具体情况动态取值;

(2) 当观测序列不平稳时,  $\xi$  应取较小的值, 以克服异常值的支配作用;

(3) 当观测序列比较平稳时,  $\xi$  应取较大的值, 充分体现关联度的整体性.

### 3.2 $\xi$ 的取值方法

从信息论的角度出发, 分辨系数的取值只要能使关联系数具有最大信息量和最大信息分辨率即可. 已知系统的信息量为  $H = \sum_i H_i = -\sum_i P_i \ln P_i$ , 要使  $\gamma(x_0(t), x_i(t))$  具有最大信息量和最大信息分辨率, 我们仅需考察下式即可:

$$H_i = \frac{\xi_{\max}}{\Delta_i(t) + \xi_{\max}} \ln \left( \frac{\xi_{\max}}{\Delta_i(t) + \xi_{\max}} \right) \quad (5)$$

欲使式(5)有最大信息量需满足:

$$\frac{dH_i}{d\xi_{\xi=\xi_1}} = 0 \text{ 并且 } \frac{d^2 H_i}{d\xi_{\xi=\xi_1}^2} < 0 \quad (6)$$

欲使式(5)有最大信息分辨率需满足:

$$\frac{d^2 H_i}{d\xi_{\xi=\xi_2}^2} = 0 \quad (7)$$

由式(6)、(7)可得

$$\begin{aligned} \xi_1 &= \frac{\Delta_i(t)}{\max} \cdot \frac{1}{e-1} \\ \xi_2 &= (e-1)\xi_1 \end{aligned} \quad (8)$$

这样, 就得到了每一时刻的最大信息量和最大信息分辨率所要求的值. 对于整个系统的全过程, 根据观测序列的平稳与否, 采用最大信息量  $\xi$  值或最大信息分辨率  $\xi$  值, 进一步计算关联系数.

### 3.3 计算关联度

由于关联系数的数目很多, 信息过于分散, 不利于比较, 为此有必要将各个时刻的关联系数集中为一个值, 求平均值便是作这种信息处理的一种方法, 也称为关联度. 因为本文分析的是相同时间截面上的数据, 因此赋予  $1 \sim n$  个点上的关联系数相同的权重, 其一般表达式为

$$r_i = r(x_0, x_i) = \frac{1}{n} \sum_{t=0}^{n-1} r_i(t) \quad (9)$$

### 3.4 排关联序

从上述的关联分析可见, 各因素的关联度数值实际意义并不大, 关联度只是因素间关联性比较的度量, 只能衡量因素间密切程度的相对大小, 比较各比较序列对于同一参考序列的关联程度孰大孰小. 因此, 将  $m$  个比较序列对参考序列的关联度按大小顺序排列起来, 便组成关联序.

### 3.5 相关影响因素的优化选择

通过用灰关联分析得到各个相关影响因素的关联序列, 可以直观了解每个影响因素对客运量的影响程度, 然后根据预测精度等要求, 把对客运量预测结果影响很小的相关影响因素去除, 这样可以更加合理的选择预测器的输入变量, 在不牺牲预测准确性的前提下, 极大程度的减少预测的时间、降低预测的难度、简化预测模型, 降低预测成本.

## 4 仿真预测

把铁路客运量(万人)作为母序列  $X_0$ , 其影响因素作为子序列, 按顺序分别记为国民生产总值(亿元) —  $X_1$ , 人口数量(万人) —  $X_2$ , 居民平均消费水平(元) —  $X_3$ , 工业企业增加值(亿元) —  $X_4$ , 社会消费品零售总额(亿元) —  $X_5$ , 旅客列车旅行速度(公里/小时) —  $X_6$ , 公路客运量(万人) —  $X_7$ , 民航客运量(万人) —  $X_8$ . 数据见表 2.

表 2 影响河南省铁路客运量发展的主要因素数据表

| 序号 | 年份   | $X_0$ | $X_1$    | $X_2$ | $X_3$    | $X_4$    | $X_5$    | $X_6$ | $X_7$  | $X_8$ |
|----|------|-------|----------|-------|----------|----------|----------|-------|--------|-------|
| 0  | 1997 | 3 843 | 4 041.09 | 9 243 | 3 378.02 | 1 662.00 | 1 427.53 | 54.9  | 65 786 | 82    |
| 1  | 1998 | 4 133 | 4 308.24 | 9 315 | 3 415.65 | 1 720.20 | 1 565.88 | 55.1  | 6 9917 | 77    |
| 2  | 1999 | 4 366 | 4 517.94 | 9 387 | 3 497.53 | 1 761.84 | 1 691.20 | 57.5  | 73 493 | 74    |
| 3  | 2000 | 4 727 | 5 052.99 | 9 488 | 3 830.71 | 2 045.15 | 1 869.80 | 58.4  | 79 017 | 77    |
| 4  | 2001 | 4 980 | 5 533.01 | 9 555 | 4 110.17 | 2 251.19 | 2 071.93 | 63.4  | 80 259 | 78    |
| 5  | 2002 | 5 085 | 6 035.48 | 9 613 | 4 504.68 | 2 508.73 | 2 292.75 | 63.4  | 85 078 | 85    |
| 6  | 2003 | 4 864 | 6 867.70 | 9 667 | 4 941.60 | 3 009.14 | 2 539.33 | 63.2  | 76 301 | 95    |

续表 2

| 序号 | 年份   | $X_0$ | $X_1$     | $X_2$ | $X_3$    | $X_4$    | $X_5$    | $X_6$ | $X_7$   | $X_8$ |
|----|------|-------|-----------|-------|----------|----------|----------|-------|---------|-------|
| 7  | 2004 | 5 695 | 8 553.79  | 9 717 | 5 294.19 | 3 837.18 | 2 938.26 | 66.7  | 85 016  | 218   |
| 8  | 2005 | 5 842 | 10 587.42 | 9 768 | 6 038.02 | 4 896.01 | 3 358.43 | 66.7  | 91 920  | 240   |
| 9  | 2006 | 6 313 | 12 362.79 | 9 820 | 6 685.18 | 6 031.21 | 3 880.47 | 66.7  | 101 345 | 297   |
| 10 | 2007 | 6 585 | 15 012.46 | 9 869 | 7 826.72 | 7 508.34 | 4 597.54 | 75.1  | 115 460 | 352   |
| 11 | 2008 | 7 476 | 18 407.78 | 9 918 | 8 837.46 | 9 546.08 | 5 662.55 | 76.2  | 122 414 | 356   |

- 4.1  $\xi$ 取值 0.5 的关联序列
- 取  $\xi = 0.5$ , 计算关联度结果见表 3.
- 4.2 采用动态  $\xi$  值的关联序列
- 由式(8) 计算表 1 中数据的  $\xi_1$  和  $\xi_2$ , 计算结果为
- $\xi_1 = 0.084\ 3, \xi_2 = 0.144\ 8$ , 即对于本系统而言, 分辨系数取值在 0.084 3 至 0.144 8 之间为佳, 本文取  $\xi$  值为 0.114 6. 计算关联度结果如表 4 所示.

表 3  $\xi$  为 0.5 时灰色关联度

| $X_1$     | $X_2$     | $X_3$     | $X_4$     | $X_5$     | $X_6$     | $X_7$    | $X_8$     |
|-----------|-----------|-----------|-----------|-----------|-----------|----------|-----------|
| 0.794 439 | 0.855 907 | 0.909 195 | 0.757 667 | 0.804 027 | 0.900 385 | 0.966 26 | 0.741 693 |

关联序为  $X_7 > X_3 > X_6 > X_2 > X_5 > X_1 > X_4 > X_8$

表 4  $\xi$  为 0.1146 时灰色关联度

| $X_1$     | $X_2$     | $X_3$     | $X_4$     | $X_5$     | $X_6$     | $X_7$     | $X_8$     |
|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 0.601 399 | 0.614 648 | 0.737 189 | 0.565 548 | 0.580 207 | 0.699 062 | 0.874 651 | 0.496 649 |

关联序为  $X_7 > X_3 > X_6 > X_2 > X_5 > X_1 > X_4 > X_8$

可以看到  $\xi$  取值 0.5 和取值 0.114 6 的关联序结果不同, 这主要是因为  $X_1$ 、 $X_4$ 、 $X_5$  观测序列不平稳有异常值存在, 采用最大信息量  $\xi$  值 0.099 1, 重新计算关联度结果如表 5 所示.

表 5  $\xi$  为 0.099 1 时灰色关联度

| $X_1$     | $X_5$    |
|-----------|----------|
| 0.582 508 | 0.556 75 |

所以, 关联序为  $X_7 > X_3 > X_6 > X_2 > X_5 > X_1 > X_4 > X_8$ .

综上所述, 可以看出传统方法  $\xi$  值关联序列和采用动态  $\xi$  值的关联序列并不相同, 在本文的具体案例中, 差别在于  $X_1$ 、 $X_5$  影响因素排序不同, 这将

会对  $G(0, 6)$  灰色预测模型的变量选择产生影响.

4.3 模型精度评估

采用  $G(0, 6)$  预测模型, 分别用残差检验和后验差检验法对预测结果进行检验, 评价变量选取对预测模型的精度和可靠性的影响. 运用 MatLab 编程对  $GM(0, 6)$  模型进行预测, 并采用  $X_7$ 、 $X_3$ 、 $X_6$ 、 $X_2$ 、 $X_1$  和  $X_7$ 、 $X_3$ 、 $X_6$ 、 $X_2$ 、 $X_5$  为影响因素的预测结果对比, 进行模型精度检验, 结果如表 6 所示.

表 6 模型精度检验表

| $G(0, 6)$ 预测模型误差检验                    |               |                                 |               |       |
|---------------------------------------|---------------|---------------------------------|---------------|-------|
| 影响因素                                  | 平均相对误差 / %    | $\Delta n / \%$ ( $n$ 为最后一个观测值) | 均方差比值         | 小概率误差 |
| $X_7$ 、 $X_3$ 、 $X_6$ 、 $X_2$ 、 $X_5$ | 2.312 298 009 | 2.210 704 871                   | 0.152 172 105 | 1     |
| $X_7$ 、 $X_3$ 、 $X_6$ 、 $X_2$ 、 $X_1$ | 2.023 540 343 | 0.588 029 876                   | 0.131 170 127 | 1     |

由表 5 可知, 选择  $X_7$ 、 $X_3$ 、 $X_6$ 、 $X_2$ 、 $X_1$  这几个相关因素的组合作为  $G(0, 6)$  预测模型的输入变量时所得到的误差较小, 这与用动态  $\xi$  值修正关联序列得结果较为一致. 所以通过动态  $\xi$  值修正关联序列对输入变量进行优化选择后, 改善了预测精度.

5 结束语

本文在灰色关联度的基础上, 通过分析分辨系数对关联度的影响, 指出分辨系数的取值应采用动  
(下转第 172 页)

## 5 结束语

Web 数据外挖掘可以为商业带来市场的行情,了解市场是商家有的放矢的基础.文中通过网店店主的应用需求,利用正则表达式法和分词法分别开发了基于 Web 的手机价格数据挖掘,实验结果表明,利用正则表达式法对网络环境的依赖性比较强,但速度快,而分词法对网络带宽的依赖性比较低,且对网页框架及内容的变化适应性强,但对计算机的运算性能要求高.随着技术的日益进步,除商业应用外,Web 挖掘应用将可以为人们提供更多的个性化服务,如为人们提供 Web 新闻<sup>[8]</sup>、分析 E-mail 邮件<sup>[9]</sup>和找工作<sup>[10]</sup>等.相信随着 Web 应用的丰富,基于 Web 的挖掘一定将存在于我们生活的方方面面.

### 参考文献:

- [1] Chen Qi, Hou Ming. XML-based data mining design and implementation [C]//International Conference on Computer Design and Applications. Qinhuaangdao, China; IEEE, 2010: 610-613.
- [2] Antony S, Wu Ping, Agrawal D. et al. Aggregate skyline: analysis for online users [C]//Ninth Annual International Symposium on Applications and the Internet. Bellevue, Washington, USA; IEEE, 2009: 50-56.
- [3] Alla H, Al-Ghreamil N. A novel efficient classification algorithm for search engines [C]//Computational Intelligence for Modelling Control & Automation. Vienna, Austria; IEEE, 2008: 773-778.
- [4] Atanasova T, Kasheva M, Sulova S, et al. Analysis of the possible application of Data Mining, Text Mining and Web Mining in Business Intelligent Systems [C]//Proceedings of the 33rd International Convention. Opatija, Croatia; IEEE, 2010: 1294-1297.
- [5] 何波,涂飞,程勇军. Web 日志挖掘数据预处理研究[J]. 微电子学与计算机, 2011, 28(4): 111-114.
- [6] Salin S, Senkul P. Using semantic information for web usage mining based recommendation [C]//International Symposium on Computer and Information Sciences. Northern Cyprus; IEEE, 2009: 236-241.
- [7] 张瑞雪,宋明秋,公衍磊. 逆序解析 DOM 树及网页正文信息提取[J]. 计算机科学, 2011, 38(4): 213-215.
- [8] Xu Cheng Zhong, Ibrahim T I. A keyword-based semantic prefetching approach in Internet news services [J]. Knowledge and Data Engineering, 2004, 16(5): 601-611.
- [9] Grobelnik M, Mladenic D, Fortuna B. Semantic technology for capturing communication inside an organization [J]. Internet Computing, 2009, 13(4): 59-67.
- [10] Litecky C, Aken A, Ahmad A, et al. Mining for computing jobs [J]. Software, 2010, 27(1): 78-85.

### 作者简介:

王红艳 女, (1979-), 硕士, 讲师. 研究方向为计算机应用;  
朱全银 男, (1966-), 教授. 研究方向为智能信息处理、接口与通信;  
严云洋 男, (1967-), 博士, 教授. 研究方向为模式识别、数字图像处理.

(上接第 167 页)

态方式,并从信息论的角度出发,给出分辨系数取值的具体算法,解决了以往分辨系数难以量化的问题.并将该算法应用于河南省铁路客运量实例,给出了基于动态灰关联分辨系数的河南省铁路客运量影响因素诊断模型.仿真结果表明,该算法是合理的,而且能有效提供关联度分辨力,使关联分析更符合实际,是一种合理的选择输入变量的有效方法.

### 参考文献:

- [1] 申卯兴,薛西锋,张小水.灰色关联分析中分辨系数的选取[J].空军工程大学学报,2003,4(1):68-70.
- [2] 刘琦.影响铁路客流的因素及相关度分析[J].上海铁道大学学报,1999,20(2):19-20.
- [3] 汪健雄,刘春煌,单杏花.基于双层次正交神经网络模型的铁路客运量预测[J].中国铁道科学,2010,31(3):126-132.
- [4] 夏国恩,金炜东,张葛祥.改进 SVR 及其在铁路客运量预测中的应用[J].西南交通大学学报,2007,42(4):494-498.
- [5] 吕锋.灰色系统关联度之分辨系数的研究[J].系统工程理论与实践,1997,17(6):49-54.
- [6] 范凯,吴皓莹.灰色系统关联度中一种新的分辨系数确定方法[J].武汉理工大学学报,2002,24(7):87-89.

### 作者简介:

王文莉 女, (1978-), 硕士, 讲师. 研究方向为数据挖掘、计算机网络;  
杨俊红 女, (1970-), 硕士, 副教授. 研究方向为数据挖掘、数字视频处理.