

基于支持向量机的区域运量滚动预测模型

刘 强¹, 陆化普¹, 王庆云², 张红亮³

(1. 清华大学 交通研究所, 100084 北京, liuqiang2007@tsinghua.org.cn; 2. 国家发展和改革委员会 交通运输部, 100824 北京; 3. 北京交通大学 交通运输学院, 100044 北京)

摘 要: 为寻求反映区域交通需求特性机理的运量预测方法, 针对一般区域运量数据小样本的问题及其诱发因素的随机性和不可控制性, 在分析区域交通需求特性及现有运量预测方法缺陷的基础上, 采用以统计学习理论为基础的专门研究小样本情况下机器学习规律的支持向量机, 建立了区域运量预测支持向量机模型. 该模型通过预测值与统计值不断交互, 实现区域运量的滚动预测, 避免了建立和求解非线性函数的过程. 以京津冀区域客运量预测为例, 验证建立模型的合理性. 结果表明, 基于支持向量机的区域运量滚动预测较传统的预测方法提高了预测精度.

关键词: 区域交通需求特性; 区域运量; 支持向量机; 滚动预测

中图分类号: U116.5

文献标志码: A

文章编号: 0367-6234(2011)02-0138-06

Rolling forecasting model of regional transportation volume based on support vector machine

LIU Qiang¹, LU Hua-pu¹, WANG Qing-yun², ZHANG Hong-liang³

(1. Institute of Transportation Engineering, Tsinghua University, 100084 Beijing, China, liuqiang2007@tsinghua.org.cn;
2. Department of Transportation, National Development and Reform Commission, 100824 Beijing, China;
3. Traffic and Transportation Department, Beijing Jiaotong University, 100044 Beijing, China)

Abstract: To get the volume forecasting method which can reflect the mechanism of regional transportation demand characteristics, and considering the limited data of regional transportation volume and the stochastic and uncontrollable of the inducing factors, a rolling forecasting model of regional transportation volume was developed based on the Support Vector Machine and the analysis of the characteristics of regional transportation demand and the defects of existing forecasting methods. Through continuous interaction between predictive value and statistical value to update the training samples, the model realized the rolling forecasts of the regional transportation volume, and avoided the process of establishment and calculation of the nonlinear function. Taking the transportation volume of Beijing - Tianjin - Hebei area for instance, the rationality of the model has been validated and the results indicate that the model enhances the forecasting accuracy of regional volume, and has better prospects.

Key words: characteristics of regional transportation demand; regional transportation volume; support vector machine; rolling forecasting

客货运量作为社会经济发展对交通运输产生

的量化反映, 是确定交通网络发展规划的重要依据, 其预测的可靠性直接影响到交通运输项目决策的成败, 因此, 科学的区域运量预测是进行区域交通规划、优化区域交通网络资源的基础和前提.

由于区域客货运量诱发因素的随机性和不可控制性, 其预测是一个非线性随机过程, 受到许多

收稿日期: 2009-04-01.

基金项目: 国家发展和改革委员会交通运输部资助项目.

作者简介: 刘 强(1980—), 男, 博士;

陆化普(1956—), 男, 教授, 博士生导师;

王庆云(1955—), 男, 教授, 博士生导师.

模糊不确定因素的影响,例如气候、季节、产业结构、经济指标和国家政策倾向等因素,一般运用定性预测和因果预测的方法难以取得理想的预测结果.在实际中,由于区域涉及范围广、行政单位多、管理层次复杂等原因,开展大规模交通量调查非常困难,加之区域缺乏长期、连续、保持一致性的统计调查数据,难以完全按照传统“四阶段法”进行区域客货运量预测^[1].因此,探索适合区域特点和反映区域交通需求特性机理的客货运量预测方法已成为一项重要的研究课题.

1 区域交通需求特性及预测方法分析

1.1 区域交通需求特性分析

由于区域与城市在辖属范围、经济、人口、产业等方面的不同,导致了区域与城市交通需求特性存在较大差异.区域交通需求主要来源于社会生产和消费领域,受人们社会经济活动、资源分布与生产力布局的影响,更多的是生产活动带来的交通需求.而城市交通需求产生的机理在于土地利用形态,主要是生活和消费性需求.此外,由于区域一些大宗货物如煤、石油、矿石等具有明显的流动方向及产业生产的季节性,使得区域交通需求在方向和时间上具有较大的不平衡性和一定的不均衡性^[2].因此,区域交通需求特性决定了其运量预测具有较大的复杂性和非线性特点.

1.2 区域运量预测方法分析

传统的运量预测方法如时间序列方法、回归模型分析法、灰色系统预测法以及基于多种方法综合的组合预测法,是一种基于大数定律的统计方法,只有样本数量趋于无穷大时才有理论上的保证,并且这些方法主要集中在对其因果关系回归模型和时间序列模型的分析上,所建立的模型不能全面、科学和本质地反映预测数据的内在机理,丢失信息量较大.加之区域运量数据通常是小样本和宏观数据,其预测受到许多模糊不确定因素的影响,是一个非线性随机过程预测,因此,采用传统统计方法难以取得较好的预测效果.

神经网络作为一种机器学习方法,具有识别复杂非线性系统的特性.但是传统神经网络的学习方法采用经验风险最小化原则,不能使期望风险最小化,同时存在收敛速度慢、结构选择和局部极小点等问题,并且通常需要大量数据进行训练,对数据的质量要求也较高.当学习样本数量较多时,又陷入“维数灾”,导致泛化性能不高^[3-6].而区域很难收集到大量质量高的运量数据,因此,限制了神经网络在区域运量预测领域的应用.

支持向量机(Support Vector Machine,简称SVM)作为一种新的机器学习方法^[7-10],主要研究如何从有限数据中挖掘出目前尚不能通过原理分析得到的规律,并利用这些规律去分析客观对象,对未知数据或无法观测的新现象进行预测和判断.由于SVM采用结构风险最小化准则,克服了神经网络的固有缺陷,较好地解决小样本、非线性、高维数和局部极小点等问题.因此,选取SVM研究区域运量预测,建立基于SVM的区域运量滚动预测模型,从本质上避开了从归纳到演绎的传统预测过程,提高了区域运量预测的精度.

2 基于SVM的区域运量滚动预测模型

2.1 支持向量机非线性回归分析

支持向量机非线性回归的基本思想是通过一个非线性映射 Φ ,将数据 x 映射到一个高维特征空间 F ,并在这个空间进行线性回归^[11-13].

$$f(x) = (\omega \cdot \Phi(x)) + b, \quad \Phi: R^{d_L} \rightarrow R^{d_F}.$$

式中: b 为阈值, $\omega = \{\omega_1, \omega_2, \dots, \omega_n, \dots\}$, $\Phi(x) = \{\Phi_1(x), \dots, \Phi_n(x), \dots\}$.

这样,在高维特征空间的线性回归便对应于低维空间的非线性回归,免去了在高维空间 ω 和 $\Phi(x)$ 点积的计算.于是函数回归问题就等价于下列泛函最小:

$$R(\omega) = R_{\text{emp}} + \lambda \|\omega\|^2 = \sum_{i=1}^n c(f(x_i) - y_i) + \lambda \|\omega\|^2.$$

式中: n 表示样本的数目, $c(\cdot)$ 为损失函数, λ 为规则化常数, R_{emp} 为经验风险.最小化 $R(\omega)$ 得到用数据点表示的 ω ,即

$$\omega = \sum_{i=1}^n (a_i - a_i^*) \Phi(x_i).$$

a_i 和 a_i^* 为最小化 $R(\omega)$ 的解.引入核函数,问题变为

$$\begin{aligned} \text{minimize} \quad & - \sum_{i=1}^n y_i (a_i^* - a_i) + \varepsilon \sum_{i=1}^n (a_i + a_i^*) + \\ & \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (a_i^* - a_i)(a_j^* - a_j) K(x_i, x_j). \end{aligned} \quad (1)$$

约束条件为

$$\begin{cases} \sum_{i=1}^n (a_i - a_i^*) = 0, \\ 0 \leq a_i, a_i^* \leq \frac{C}{n}, i = 1, 2, \dots, n. \end{cases}$$

式中: $K(x_i, x) = \Phi(x_i) \cdot \Phi(x)$ 为核函数,满足Mercer条件.

求解问题(1),得到最优解 $\bar{a} = (\bar{a}_1, \bar{a}_1^*, \dots, \bar{a}_n, \bar{a}_n^*)^T$.

构造决策函数:

$$f(x) = \sum_{i=1}^n (\bar{\alpha}_i^* - \bar{\alpha}_i) (\Phi(x_i) \cdot \Phi(x)) + \bar{b} = \sum_{i=1}^n (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_i, x) + \bar{b}.$$

由于 SVM 理论只考虑高维特征空间的点积运算 $K(x_i, x) = \Phi(x_i) \cdot \Phi(x)$, 而不直接使用函数 Φ , 从而解决了因映射函数 Φ 未知而 ω 无法显示表达的问题.

根据 KKT(karush - kukn - tucher)条件,可得

$$\bar{b} = y_j - \sum_{i=1}^n (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_i, x_j) + \varepsilon,$$

$$\bar{b} = y_k - \sum_{i=1}^n (\bar{\alpha}_i^* - \bar{\alpha}_i) K(x_i, x_k) - \varepsilon.$$

2.2 基于 SVM 的区域运量滚动预测模型

考虑到区域交通需求特性和 SVM 不具备根据实际情况动态调整预测值的能力,通过预测值与统计值不断交互、不断更新训练集,构建了基于支持向量机的区域运量滚动预测模型. 具体流程如图 1 所示.

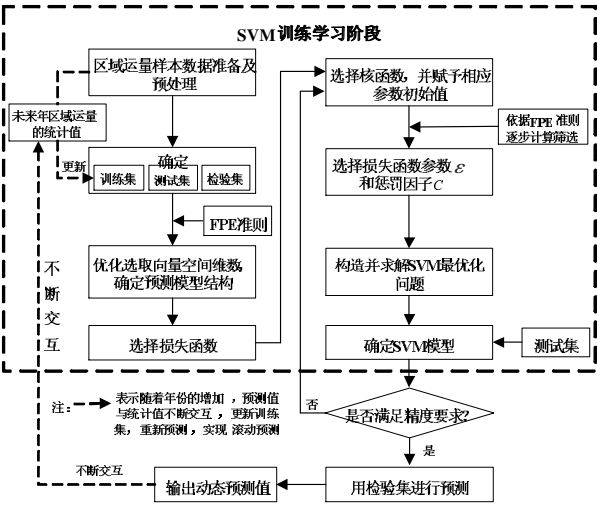


图 1 区域运量滚动预测 SVM 模型流程图

2.2.1 样本数据集的确定

本文将区域运量样本数据分为 3 部分:用于建立 SVM 模型的训练样本集、用于测试 SVM 模型预测能力的测试样本集以及检验预测模型预测效果(推广能力)的检验样本集,其中检验样本集的数据不参与训练学习及参数筛选等建模过程.

2.2.2 模型结构的确定

建立自相关输入 $x_{t+1} = \{x_t, x_{t-1}, \dots, x_{t-m}\}$ 与输出 $y_{t+1} = \{x_{t+1}\}$ 之间的映射关系 $f: R^m \rightarrow R$, 其中 m 值为输入向量的空间维数. 采用最小最终误差预测准则 FPE (Final Prediction Error, 简称

FPE)评价模型的预测误差,并根据误差大小来优化选取嵌入维数. 经过变换之后,得到用于 SVM 模型的学习样本:

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \\ x_2 & x_3 & \cdots & x_{m+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n_{tr}-m} & x_{n_{tr}-m+1} & \cdots & x_{n_{tr}-1} \end{bmatrix},$$
$$Y = \begin{bmatrix} x_{m+1} \\ x_{m+2} \\ \vdots \\ x_{n_{tr}} \end{bmatrix}.$$

$$E_{FP}(m) = \frac{n_{tr} + m}{n_{tr} - m} \sigma^2. \quad (2)$$

$$\sigma^2 = E(a_{n_{tr}}) = \frac{1}{n_{tr} - m_{t=m+1}} \sum_{t=m+1}^{n_{tr}} [y_t - \left(\sum_{i=1}^{n_{tr}-m} (a_i - a_i^*) K(x_i \cdot x_t) + b \right)]^2.$$

式中: n_{tr} 为用于训练的数据个数; m 为需要确定的嵌入维数.

从式(2)可以看出, $E_{FP}(m)$ 取最小值时的 m 就是模型的最佳输入向量的维数,即

$$E_{FP}(m_0) = \min_{0 \leq m \leq m_h} E_{FP}(m).$$

式中: m_h 为输入向量空间维数的上界.

因此,确定了最佳预测模型的输入向量的空间维数,也就确定了 SVM 预测模型的拓扑结构. 区域运量预测的 SVM 模型基本结构如图 2 所示.

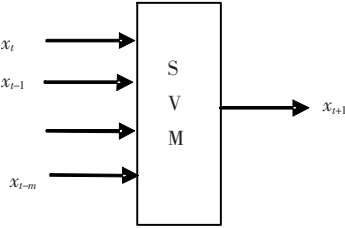


图 2 SVM 模型结构

在图 2 中 $x_t, x_{t-1}, \dots, x_{t-m}$ 分别表示 $t, t-1, \dots, t-m$ 年的区域运量, x_{t+1} 表示 $t+1$ 年的运量, 即要预测的区域运量. 确定了预测模型的结构后, 利用学习样本就可以得到回归函数:

$$y_t = \sum_{i=1}^{n_{tr}-m} (a_i - a_i^*) K(x_i \cdot x_t) + b, t = m+1, \dots, n_{tr}.$$

则有一步预测模型:

$$y_{n_{tr}+1} = \sum_{i=1}^{n_{tr}-m} (a_i - a_i^*) K(x_i \cdot x_{n_{tr}-m+1}) + b.$$

式中: $x_{n_{tr}-m+1} = \{x_{n_{tr}-m+1}, x_{n_{tr}-m+2}, \dots, x_{n_{tr}}\}$, 得到 $n+1$ 点的值后, 又可得到一个样本数据 $x_{n_{tr}-m+2} = \{x_{n_{tr}-m+2}, x_{n_{tr}-m+3}, \dots, x_{n_{tr}}, \hat{x}_{n_{tr}+1}\}$, 其中 $\hat{x}_{n_{tr}+1}$ 表示第

$n + 1$ 点的预测值. 依次递推, 可得到 k 步预测模型为

$$y_{n_{tr}+k} = \sum_{i=1}^{n_{tr}-m} (a_i - a_i^*) K(x_i \cdot x_{n_{tr}-m+k}) + b.$$

式中: $x_{n_{tr}-m+k} = \{x_{n_{tr}-m+k}, \cdots, \hat{x}_{n_{tr}+1}, \cdots, \hat{x}_{n+k-1}\}$.

2.2.3 损失函数的选择

SVM 区域运量预测模型采用 ε - 不敏感损失函数作为支持向量机的损失函数.

2.2.4 核函数的选择

选用高斯核函数 $K(x, x_i) = \exp\left(-\frac{|x - x_i|^2}{2\sigma^2}\right)$ 进行预测, 并采取逐步筛选的方法确定核函数的参数.

2.2.5 损失函数参数 ε 和惩罚因子 C 的选择

对于这两个参数的选取, 目前尚未有统一的理论和经验. 本文运用逐步计算筛选的方法依据 FPE 准则进行选择, 从而确定最优的 SVM 预测模型.

3 实例分析

以京津冀地区 1990 ~ 2005 年客运量数据作为全部学习样本, 选取 1990 ~ 2000 年的客运量数据作为训练集, 选取 2001 ~ 2005 年的客运量数据作为测试集, 构建 SVM 模型. 并通过建立不同的 SVM 模型预测测试集进行参数优选, 选取平均相对误差作为评价指标, 部分结果见表 1.

表 1 不同参数取值时单步预测的平均相对误差

参数组合	C	ε	σ^2	平均相对误差
1	10	0.000 1	3.5	0.013 23
2	100	0.000 1	3.5	0.000 14
3	1 000	0.000 1	3.5	0.000 14
4	10 000	0.000 1	3.5	0.000 14
5	100	0.010 0	1.5	0.013 37
6	100	0.001 0	1.5	0.001 41
7	100	0.000 5	1.5	0.000 69
8	100	0.000 1	1.5	0.000 14
9	100	0.000 1	0.5	0.000 14
10	100	0.000 1	2.0	0.000 14
11	100	0.000 1	3.0	0.000 14
12	100	0.000 1	4.0	0.000 60
13	100	0.000 1	4.5	0.001 87
14	100	0.000 1	5.0	0.003 29

从表 1 可以看出, 在相同的 σ^2 和 ε 下, 当 C 值较大时预测效果比较好. 当 $C = 10\ 000$ 时, 预测误差最小; 在相同的 C 值和 σ^2 下, 随着 ε 的增大,

相应的误差也随之增大. 当 $\varepsilon = 0.000\ 1$ 时, 相应的误差最小; 在相同的 C 值和 ε 下, 当 σ^2 较小时, 误差较小, 随着 σ^2 的增大, 平均相对误差增大.

因此, 本文选取最佳的模型参数 $C = 100$ 、 $\varepsilon = 0.000\ 1$ 、 $\sigma^2 = 0.5$ 行京津冀地区客运量预测, 2001 ~ 2005 年客运量的预测结果及相对误差分别见图 3、4.

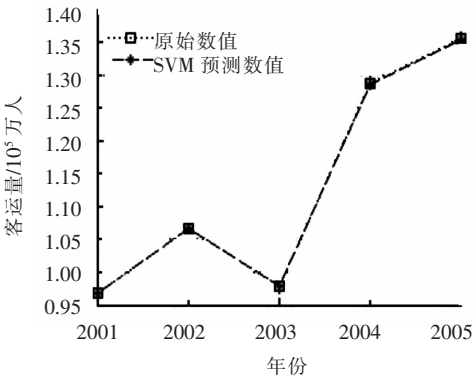


图 3 SVM 模型客运量预测结果

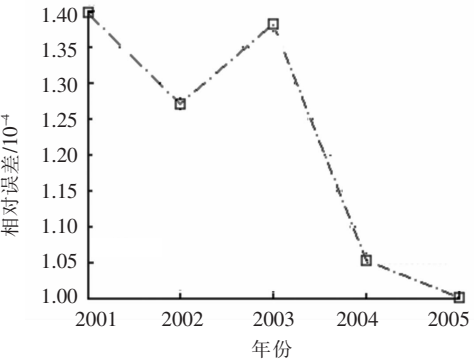


图 4 SVM 模型客运量预测相对误差

经过计算, 本文建立的基于 SVM 的区域客运量预测模型的平均相对误差为 0.15%, 预测误差在可接受的范围内. 从图 4 可以看出, 预测值与实际值基本吻合, 整体变化趋势一致, 反映了用支持向量机预测区域客运量的准确性. 在此基础上, 以 1990 ~ 2005 年的数据作为训练集, 采用支持向量机滚动预测 2006 ~ 2020 年京津冀地区客运量, 采用不同模型参数的预测结果与采用其他方法的预测结果比较如图 5 所示.

由图 5 可以看出, 随着年份的增加, 用支持向量机进行远期预测的误差呈现递减的趋势, 究其原因主要是用支持向量机近期预测比较准, 而对于远期预测无论是用支持向量机还是用其他方法预测结果都不准, 所以误差越来越小. 因此, 可以利用支持向量机近期预测准确性高的优势, 通过预测值与统计值不断交互、不断更新训练集, 以实现区域运量的滚动预测, 提高预测的精度.

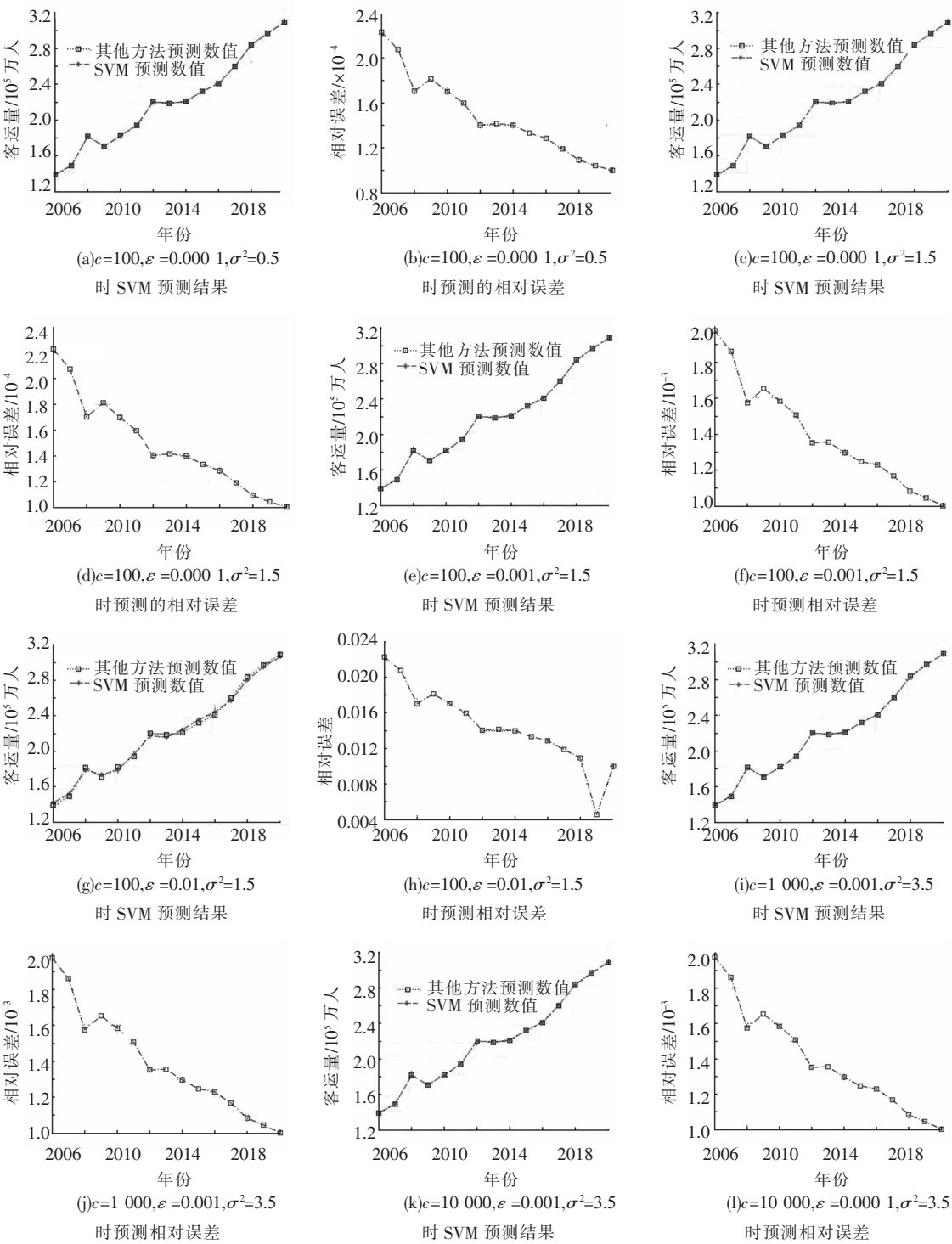


图 5 预测结果比较

4 结 语

本文在分析区域交通需求特性和现有运量预测方法存在问题的基础上,建立了基于支持向量机的区域运量滚动预测模型,给出了基于支持向量机的区域运量滚动预测模型的流程图,并以京津冀地区客运量预测为例,验证了模型的有效性.选取不同的核函数参数、损失函数参数和惩罚因子值建立的 SVM 模型,其预测结果有较大的差

异,因此,如何选取和优化参数的取值是需要进一步研究的问题.

参考文献:

[1] 刘强. 区域交通网络资源优化配置辅助决策理论与方法研究[D]. 北京: 清华大学土木工程系, 2007.
[2] 杨浩, 张秀媛. 综合运输与区域经济[M]. 北京: 中国铁道出版社, 1995:28-34.
[3] VAPNIK V N. The nature of statistical learning theory

- [M]. New York: Springer, 2000:73 – 88.
- [4] 边肇祺, 张学工. 模式识别[M]. 3版. 北京: 清华大学出版社, 2010:73 – 89.
- [5] VANAIAKSHI L, LAURENCE R R. A comparison of the performance of artificial neural networks and support vector machines for the prediction of traffic speed [C]// IEEE Intelligent Vehicles Symposium 2004. TX, USA: Texas A & M Univ, 2004: 194 – 199.
- [6] VAPNIK V, LEVIN E, LE C Y. Measuring the VC – dimension of a learning machine[J]. Neural Computation, 1994, 6(5): 851 – 876.
- [7] VAPNIK V N. An overview of statistical learning theory [J]. IEEE Transactions on Neural Networks, 1999, 10(5): 988 – 999.
- [8] SCHOLKOPF B, SUNG K K, BURGESS C, *et al.* Comparing support vector machines with gaussian kernels to radial basis function classifiers[J]. IEEE Trans on Signal Processing, 1997, 45(11): 2758 – 2765.
- [9] 陈永义, 俞小鼎, 高学浩, 等. 处理非线性分类和回归问题的一种新方法: 支持向量机方法简介[J]. 应用气象学报, 2004, 15(3): 345 – 354.
- [10] CRISTIANINI N, SHAWEETAYLOR J. An introduction to support vector machines and other kernel – based learning methods[M]. Beijing: Publishing House of Electronic Industry, 2004: 33 – 45.
- [11] 彭新俊. 支持向量机若干问题及应用研究[D]. 上海: 上海大学数学系, 2008.
- [12] 邓乃扬, 田英杰. 数据挖掘中的新方法: 支持向量机[M]. 北京: 科学出版社, 2004: 77 – 86.
- [13] THEODORIDIS S, MAVROFORAKIS M. Reduced convex hulls: a geometric approach to support vector machines[J]. IEEE Signal Processing Magazine, 2007, 24(3): 119 – 122.

(编辑 刘 彤)

(上接第 113 页)

- [4] 刘国君, 程光明, 杨志刚. 一种压电式精密输液微泵的试验研究[J]. 光学精密工程, 2006, 14(4): 612 – 616.
- [5] KOCH M, EVANS A, BRUNNSCHEWILER A. Microfluidic technology and applications[M]. Herts: Research Studies Ltd., 2000: 25 – 50.
- [6] STUDER V, PEPIN A, CHEN Y, *et al.* An integrated AC electrokinetic pump in a microfluidic loop for fast and tunable flow control[J]. The Analyst, 2004, 29: 944 – 949.
- [7] CAHILL B P, HEYDERMAN L J, GOBRECHT J, *et al.* Flow reversal at low voltage and low frequency in a microfabricated ac electrokinetic pump[J]. Physical Review E, 2007, 76: 036305 – 036314.
- [8] RAMOS A, Morgan H, GREEN N G, *et al.* AC electric-field-induced fluid flow in microelectrodes[J]. J Colloid and Interface Science, 1999, 217: 420 – 422.
- [9] RAMOS A, GONZÁNCHEZ A. A linear analysis of the effect of Faradaic currents on traveling – wave electroosmosis[J]. J Colloid Interface Sci, 2007, 309: 323 – 331.
- [10] JHUNJHUNWALA M. Multiphase flow and control of fluid path in microsystems[D]. Massachusetts: Massachusetts Institute of Technology, 2003.
- [11] CHEN Lingxin, LEE S. Continuous dynamic flow micropumps for microfluid manipulation[J]. Micromech Microeng, 2008, 18: 013001 – 013022.
- [12] LASER D, SANTIAGO J. A review of micropumps [J]. Micromech Microeng, 2004, 14: 35 – 64.
- [13] JIANG Hongyuan, REN Yukun, AO Hongrui. Electrohydrodynamic analysis based on conductivity gradient in microchannel[J]. Chin Phys, 2008, 17(12): 4541 – 4547.
- [14] STRATTON J A. Electromagnetic theory[M]. New York: McGraw-Hill, 1941: 508 – 510.
- [15] CASTELLANOS A, RAMOS A, GONZALEZ A, *et al.* Electrohydrodynamics and dielectrophoresis in microsystems: scaling laws [J]. J Phys D: Appl Phys, 2003, 36: 2584 – 2597.
- [16] NEWMAN J S, THOMAS-ALYE K E. Electrochemical systems [M]. Hoboken, NJ: Wiley-Interscience, 2004: 517 – 562.
- [17] CHEN CH, LIN H, LELE S K, *et al.* Convective and absolute electrokinetic instability with conductivity gradients[J]. J Fluid Mech, 2005, 524: 263 – 303.

(编辑 赵丽莹)