第 35 卷　第 4 期
2011 年 8 月

武汉理工大学学报(交通科学与工程版)
Journal of Wuhan University of Technology
(Transportation Science & Engineering)

Vol. 35　No. 4
Aug. 2011

基于混沌高效遗传算法优化 SVM 的交通量预测*

康海贵　李明伟　周鹏飞　赵泽辉

(大连理工大学工程建设学部　大连　116024)

摘要:针对交通量预测本身所存在的小样本、非线性和复杂性等特点,利用支持向量机建立了基于 RBF 核函数的 SVM 交通量预测模型,采用基于混沌映射和加速遗传算法的混沌高效遗传算法对 SVM 模型参数 C,ε 和 δ^2 进行优选,结合某市 1978～2008 年交通量实测资料进行了仿真验证,与 GA-SVM 模型和 BP 神经网络模型的仿真预测结果对比表明:该模型取得了较好的预测效果,可有效应用于城市交通量的预测.

关键词:公路交通流量预测; 支持向量机; 加速遗传算法; 混沌

中图法分类号:U116.5　　**DOI**:10.3963/j.issn.1006-2823.2011.04.001

支持向量机(SVM)在经验风险最小化的基础上同时采用了结构风险最小化准则[1],很好地解决了小样本、非线性、高维数、局部极小等实际问题[2],避免了神经网络的参数难以选择、易陷入局部极值和过拟合等缺陷.但是,支持向量机本身并没有给出选择合适参数的方法,目前主要依靠经验法和试算法,很大程度上影响了模型的推广使用.考虑到基于混沌理论改进的混沌高效遗传算法(CHEGA)在参数优选中表现出的快速、高效的特点,本文建立了基于 CHEGA 进行参数优选的 CHEGA-SVM 交通量预测模型,并进行了仿真验证.对比结果表明该模型在交通量预测过程中具有一定的实用性.

1　支持向量回归的基本原理

在解决函数回归问题时,SVM 方法的基本思想是:通过事先定义的非线性映射 $\varphi: R_n \to R_m (m \geqslant n)$,把输入空间的数据 x 映射到一个高维特征空间,然后在该空间中做线性回归 $\varphi: R_n \to R_m (m \geqslant n)$.给定数据点集为 $\{(\boldsymbol{x}_i, \boldsymbol{y}_i), i=1,2,\cdots,N\}$,式中:$\boldsymbol{x}_i \in R_n$ 为输入变量;$\boldsymbol{y}_i \in R_n$ 为与 x_i 相对应的输出向量;N 为数据点总数.SVM 通过下式进行函数回归估计

$$f(x) = \boldsymbol{\omega} \cdot \boldsymbol{\varphi}(x) + b \tag{1}$$

式中:$\boldsymbol{\omega}$,$\boldsymbol{\varphi}$ 为 m 向量;"·"表示特征空间中的点积;$b \in R$ 为阈值.SVM 采用最小化结构 R_{str} 来确定参数 w 和 b,即

$$\min R_{\mathrm{str}} = \frac{1}{2}\|\boldsymbol{\omega}\|^2 + CR_{emp} \tag{2}$$

式中:

$$R_{emp} = \frac{1}{l}\sum_{i=1}^{l} L_\varepsilon[x_i, y_i - f(x_i)] \tag{3}$$

$$L_\varepsilon = (x, y - f(x)) = \max\{0, |y - f(x) - \varepsilon|\}$$

式(2)等号右端第 1 项 $\frac{1}{2}\|\boldsymbol{\omega}\|^2$ 为决策函数复杂性的表达能力项,是正规化项;第 2 项经验风险 R_{emp} 为训练误差项;平衡因子 $c(c \geqslant 0)$ 为权重系数.经验风险 R_{emp} 由惩罚函数(loss function)来度量,通常采用 ε -不敏感损失函数 $L_\varepsilon[x_i, y_i - f(x_i)]$.

同时引入松弛变量 ξ_i 和 ξ_i^* $(i=1,2,\cdots,l)$,式(2)可改写为

$$\min \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \tag{4}$$

收稿日期:2011-02-10

康海贵(1945-):男,教授,博士生导师,主要研究领域为交通规划、高速公路后评价等

* 教育部博士点专项基金项目(批准号:200901411105)、河南省交通厅科技计划项目(批准号:2010D107-4)资助

$$\text{s.t.}\quad [\boldsymbol{\omega}\cdot\boldsymbol{\varphi}(x_i)]+b-y_i\leqslant\varepsilon+\xi_i \tag{5}$$
$$y_i-[\boldsymbol{\omega}\cdot\boldsymbol{\varphi}(x_i)]-b\leqslant\varepsilon+\xi_i$$
$$\xi_i,\xi_i^*\geqslant 0$$

上式便是 SVM 的原始问题,可以看出原始问题是一个有线性约束的二次规划问题. 根据强对偶定理,引入 Lagrange 乘子 α_i 和 α_i($i=1,2,\cdots,l$),建立 Lagrange 函数,并对 w,b,ξ_i 和 ξ_i^* 求偏导并令其为 0,于是得到上述 SVM 原始问题的对偶问题为

$$\min\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i,y_i)+\varepsilon\sum_{i=1}^{l}(\alpha_i^*-\alpha_i)-\sum_{i=1}^{l}y_i(\alpha_i^*-\alpha_i)$$
$$\text{s.t.}\quad\sum_{i,j=1}^{l}(\alpha_i^*-\alpha_i)=0$$
$$0\leqslant\alpha_i^*,\alpha_i\leqslant c,\ i=1,2,\cdots,l \tag{6}$$

式中:$K(x_i,y_i)=[\varphi(x_i)\cdot\varphi(x_j)]$为核函数,其作用是不必知道从低维输入空间到高维特征空间非线性映射 $\varphi(x)$的具体形式,通过引入核函数就可得到决策回归方程. 常用的核函数主要有

多项式函数 $K[x,x_i]=[x\cdot x_i+1]^q$

RBF 函数 $K(x,x_i)=\exp\{-\|\boldsymbol{x}-\boldsymbol{x}_i\|^2/2\delta^2\}$

Sigmoid 函数 $K(x,x_i)=\tanh(\upsilon(x,x_i)+c)$

本文选用的核函数是 RBF 函数[3],其中 δ 为 RBF 函数的宽度函数. 设对偶问题的解为 $\overline{\alpha}=(\overline{\alpha_1},\overline{\alpha_1^*},\cdots,\overline{\alpha_l},\overline{\alpha_l^*})$,根据 SVM 的稀疏性一对 α_i 和$\overline{\alpha_i}$最多只有一个不为零,即$\overline{\alpha_i},\overline{\alpha_i^*}=0$,$\forall_i\in(1,2,\cdots,l)$且只有少数$\overline{\alpha_i}$和$\overline{\alpha_i^*}$可不为零,这些不为零的参数所对应的输入向量 $\boldsymbol{x}_i$ 称为支持向量,则决策方程只由支持向量决定,与非支持向量无关. 决策回归方程式(1)可改写为

$$f(x)=\sum_{i=1}^{l}(\overline{\alpha_i^*}-\overline{\alpha_i})\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}_i\|^2}{2\delta^2}\right)+\overline{b} \tag{7}$$

式中:

$$\overline{b}=\begin{cases}y_i-\sum_{i=1}^{l}(\overline{\alpha_i^*}-\overline{\alpha_i})K(x_i,x_j)+\varepsilon & \overline{\alpha_j}\in(0,c)\\ y_i-\sum_{i=1}^{l}(\overline{\alpha_i^*}-\overline{\alpha_i})K(x_i,x_j)+\varepsilon & \overline{\alpha_j^*}\in(0,c)\end{cases} \tag{8}$$

式中:x_i 为训练年份影响因子向量;x 为预测年份的影响因子向量;$f(x)$为指标集合向量. 式(7)即为 SVM 的交通量预测模型.

由上述可知,SVM 的参数主要为惩罚系数 C、损失函数参数 ε 和 RBF 核函数参数 δ^2,它们的选择在很大程度上影响着 SVM 的预测精度和泛化性能,只有正确选择参数,才可以使 SVM 回归估计得到较好的拟合效果[4],因此为了提高 SVM 预测模型的预测精度和泛化性能,本文利用 CHEGA 进行参数 C,ε 和 δ^2 的优选.

2 混沌映射

设 Logistic 映射产生的混沌变量为$\{x(i)\}$,$i=1,2,\cdots,n$[5],则有

$$x(i+1)=u\cdot x(i)\cdot(1-x(i)),\ x(i)\in[0,1] \tag{9}$$

式中:$x(i)$为变量 x 在第 i 次的迭代值,其中 u 是控制参量.

理论上已经证明当 $u\geqslant 4$ 时系统完全处于混沌状态,所以 $x(0)$可任意设为[0,1]区间内的初值,但不能为 0.25,0.5 和 0.75,然后根据式(9)得到混沌变量[6].

图 1 给出了 $n=500$,$u=4$,$x(0)=0.45$ 时混沌变量的分布情况. 图 2 给出了 $n=500$ 一般随机均匀分布系统产生的均匀随机变量分布情况.

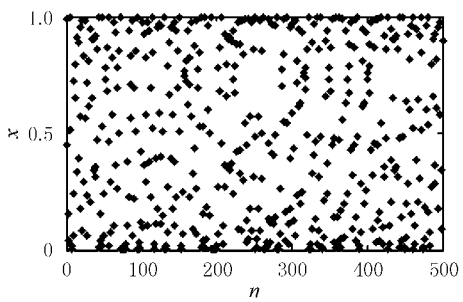

图 1 Logistic 映射混沌变量图

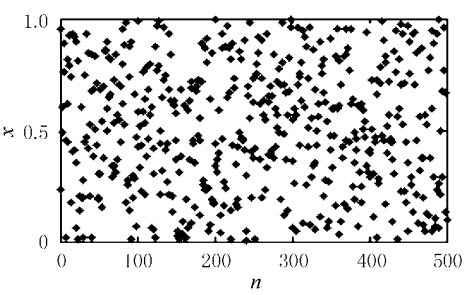

图 2 均匀分布随机变量图

从图 1 和图 2 可以看出,Logistic 混沌映射轨道点能够布满整个区域的内部和边界,并且区域的内部和边界的轨道点的数量分布比较均匀,而均匀分布点虽然能够布满整个区域,但多分布在区域的内部,在区域的边界上的数量分布很少. 在大量的非线性优化问题中,最优点的分布可能在区域的内部,也可能在边界上,因此采用 Logistic 混沌映射来生成初始群体,并利用混沌变异改

进的混沌高效遗传算法，能够更好地搜索整个区域，保持种群的多样性，进行精细搜索.

3　CHEGA-SVM 交通量预测模型算法

CHEGA 优化 SVM 参数问题可以转化为如下最小化问题

$$\min\ Q(\gamma,\delta^2)=\frac{1}{n}\sum_{t=1}^{n}(Y^1(t)-Y(t))^2$$
$$\text{s.t.}\quad c\in[c_{\min},c_{\max}],\varepsilon\in[\varepsilon_{\min},\varepsilon_{\max}],\quad \delta^2\in[\delta^2_{\min},\delta^2_{\max}] \tag{10}$$

式中：n 为输入样本个数；$Y^1(t)$ 为 SVM 拟合序列；$Y(t)$ 为实际序列值. 基于 CHEGA 优化 CHEGA-SVM 交通量预测模型的计算步骤如下.

步骤 1　样本数据和参数初始区间归一化. 对输入的样本数据和参数的初始化区间进行归一化处理，归一化方法如下.

$$x(j)=a(j)+y(j)\cdot(b(j)-a(j)) \tag{11}$$

式中：$b(j)$ 和 $a(j)$ 分别为第 j 类数据取值范围的最大值和最小值.

步骤 2　混沌随机父代群体初始化. 由 Logistic 混沌映射在参数初始区间内生成初始父代群体，m 个初始父代群体记为 $y(j,i)$，$(j=1,2,3;i=1,2,\cdots,m)$，其中 $y(j,i)$ 为第 i 个父代个体上的第 j 个基因，本文取 $m=50$，$u=4.0$，$y(j,0)=0.7$.

步骤 3　父代个体的适应度评价. 将第 i 个父代个体的基因作为 SVM 参数，以训练集样本为输入输出，对 SVM 模型进行训练，将 SVM 训练结束后返回的输出序列方差 Q 的倒数作为第 i 个父代个体的适应度值 $F(i)=1/Q$.

步骤 4　选择操作. 产生第 1 个子代群体 $\{y_1(j,i),j=1,2,3;i=1,2,\cdots,m\}$，取比例选择方式，则父代个体 $y(j,i)$ 的选择概率为 $p_s(i)=F(i)/\sum_{i=1}^{m}F(i)$，令 $p(i)=\sum_{k=1}^{i}p_s(k)$，则序列 $\{p(i),i=1,2,\cdots,m\}$ 把 $[0,1]$ 区间分成 m 个子区间，生成 $m-5$ 个随机数 $\{u(k),k=1,2,\cdots,n-5\}$，若 $u(k)$ 在 $[p(i-1),p(i)]$ 中，则第 i 个个体 $y(j,i)$ 被选中，即 $y_1(j,k)=y(j,i)$，为增强 CHEGA 进行全局优化搜索能力，这里把最优秀的 5 个父代个体直接加入子代群体中，即进行移民操作

$$y_1(j,m-5+i)=y(j,i),\ i=1,2,\cdots,5 \tag{12}$$

步骤 5　杂交操作. 产生第 2 个子代群体 $\{y_2(j,i),j=1,2,3;i=1,2,\cdots,m\}$，根据选择概率随机选择一对父代个体 $y(j,i_1)=y(j,i_2)$ 作为双亲，并进行如下随机线性组合，产生一个子代个体 $y_2(j,i)$

$$\begin{aligned}y_2(j,i)&=u_1y(j,i_1)+(1-u_1)y(j,i_2),u_3<0.5\\ y_2(j,i)&=u_2y(j,i_1)+(1-u_2)y(j,i_2),u_3\geqslant 0.5\end{aligned} \tag{13}$$

式中：u_1,u_2,u_3 都为随机数，通过杂交操作，共产生 m 个子代个体.

步骤 6　混沌变异操作. 产生第 3 个子代群体 $\{y_3(j,i),j=1,2,3;i=1,2,\cdots,m\}$，CHEGA 的混沌变异操作采用自适应变异概率 $p_m(i)=1-p(i)$ 来代替个体 $y(j,i)$，从而得到子代个体 $y_3(j,i)$，即

$$\left.\begin{aligned}y_3(j,i)&=u(j),u_m<p_m(i)\\ y_3(j,i)&=y(j,i),u_m\geqslant p_m(i)\end{aligned}\right\} \tag{14}$$

式中：$u(j)$ 为混沌变量；u_m 为 $[0,1]$ 区间上的均匀随机数.

步骤 7　演化迭代. 由步骤 4 到 6 得到的 $3m$ 个子代个体，按照其适应度值 $F^*(i)$ 从大到小进行排序，判断当前种群中最优个体是否满足终止准则，若满足转入步骤 9，否则取排在最前面的 m 个子代个体作为新的父代群体. 然后转入步骤 3 进入下一轮演化过程. 停止准则采用最大进化代数 $G_{\max}$ 与相邻进化代数最优个体适应值相对误差 E 相结合.

步骤 8　引进加速遗传算子. 每演化迭代两次，取进化得到的 $6m$ 个子代个体中的 m 个优秀个体的变化区间作为新的初始变化区间，转入步骤 2，重新开始迭代过程.

步骤 9　训练模型输出预测结果. 以当前种群中具有最大适应值的个体的染色体基因 C，ε 和 δ^2 作为 SVM 参数训练模型，输入待测样本，输出预测值.

计算流程见图 3.

4　模型仿真验证

4.1　CHEGA-SVM 交通量预测模型的训练与预测

结合某市客运量统计数据进行仿真验证，指标集为 1978～2008 年全市客运量，经相关性分析得出影响因子集为 1978～2008 年全市消费品总额、人均收入、能源总量、生产总值、汽车拥有量和总人口. 在总体样本中抽取 1982、1986、1990、

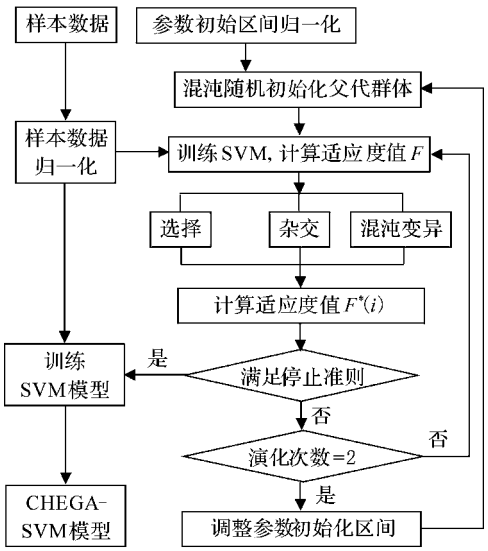

图 3 算法流程图

1995、2001、2007 年的影响因子和指标作为测试集,其余为训练集,应用 CHEGA-SVM 交通量预测模型分别进行训练和预测.

利用 Matlab 7.1 编制模型程序,运行环境为:Core(TM)2CPU,1.81 MHz, 2 GB 内存的微机.操作系统:WindowsXP.仿真中,C 的取值范围为[0.01,200];ε 的取值范围为[0.01,0.8];δ^2 取值范围为[0.01,50],最大进化代数 $G_{\max}=100$,邻代最优个体适应值相对误差 $E=0.000\ 01$.为了验证模型的使用性能,同时选用基于 GA 优化的 GA-SVM 模型和基于梯度下降法的传统 BP 神经网络模型进行对比仿真预测,并采用以下三个评价指标进行预测性能对比分析.

1) 平均绝对相对误差

$$mrerr=\frac{1}{n}\sum_{t=1}^{n}\left|\frac{Y^1(t)-Y(t)}{Y(t)}\right| \tag{15}$$

2) 最大绝对相对误差

$$mxarer=\max\left|\frac{Y^1(t)-Y(t)}{Y(t)}\right| \tag{16}$$

3) 均方根误差

$$rmse=\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\frac{Y^1(t)-Y(t)}{Y(t)}\right)^2} \tag{17}$$

式中:$Y^1(t)$为 SVM 拟合序列;$Y(t)$为实际序列值;n 为检验数据个数.

结合实测数据,进行模型训练和客运量仿真预测,图 4 为 CHEGA-SVM 交通量预测模型训练曲线,由图 4 可以看出训练输出和实际值拟和的较好,拟合平均绝对相对误差为 2.42%,拟合均方根误差为 3.21%,说明 CHEGA-SVM 交通量预测模型对客运量的历史序列拟合较稳定.

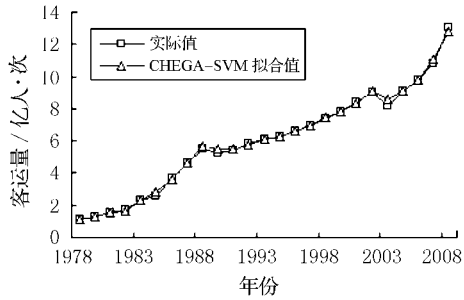

图 4 CHEGA-SVM 模型训练图

利用得到的 CHEGA-SVM 交通量预测模型对 1982、1986、1991、1996、2002、2007 年的客运量进行预测.图 5 为 CHEGA-SVM 交通量预测值与实际值对比图,仿真预测平均绝对相对误差为 2.92%,仿真预测均方根误差为 3.89%.

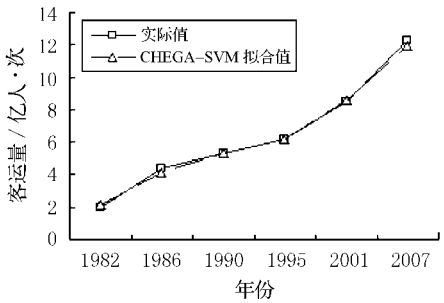

图 5 CHEGA-SVM 模型预测图

4.2 CHEGA-SVM 交通量预测模型性能对比分析

将预测得到的客运总量与传统 BP 神经网络模型和 GA-SVM 模型仿真预测结果进行比较分析,其中 BP 神经网络模型选择的是单隐层 BP 神经网络,BP 神经网络参数:最大迭代次数为 100 次,隐层节点数为 7,学习率为 0.25,允许的平方根误差为 0.000 1;GA 参数:种群大小为 50,进化最大代数为 100,交叉概率 P_c 为 0.55,变异概率 P_m 为 0.005.仿真预测值如表 1 所列.

表 1 CAGA-υ-SVR 与其他 3 个模型的预测比较

年份	真实值/万人	模型预测值/万人		
	原始序列	传统 BP 模型	GA-SVM	CHEGA-SVM
1982	20 129	17 759	18 449	18 556
1986	43 590	37 621	39 252	39 431
1990	53 567	56 812	51 464	51 632
1995	61 964	66 035	59 212	59 424
2001	85 412	81 010	88 521	88 031
2007	122 557	114 588	116 412	116 851

将表 1 中数据按式(15)、(16)和(17)分别进行误差指标处理,处理结果如表 2 所列.

表 2 仿真预测误差指标比较 %

模 型	*mrerr*	*mxarer*	*rmse*
传统 BP 模型	8.60	13.16	9.21
GA-SVM 模型	5.87	9.99	6.34
CHEGA-SVM 模型	2.92	7.42	3.89

从表 2 中可以看出，在对基于不同结构预测模型的仿真预测结果对比方面，SVM 模型的仿真预测结果要优于传统的 BP 神经网络模型；对采用不同的算法进行参数优选的 SVM 模型对比方面，基于混沌高效遗传算法优化的 SVM 模型的预测精度优于采用传统遗传算法优化的 SVM 模型.

5　结 束 语

本文提出了基于 CHEGA 进行参数优选的 CHEGA-SVM 交通量预测模型，并结合实际数据进行了仿真验证，对比结果表明：(1) 改进的混沌高效遗传算法可以遍历到整个区域的内部和边界，较好的保持了种群的多样性，提高了参数的优选精度，同时利用优秀个体群逐步缩小搜索空间，加快了寻优速度. 利用混沌高效遗传算法优选 SVM 模型的参数，克服了 SVM 模型参数确定难度大、精度低的问题，增强了 SVM 模型的泛化能力；(2) 应用 CHEGA-SVM 交通量预测模型进行交通量仿真预测，模型的仿真预测精度优于 GA-SVM 模型和传统 BP 神经网络模型，平均绝对相对误差控制在 3.0% 以内，提高了城市交通量的预测精度，说明采用该模型对城市交通量的预测是可行的；(3) 本文提出的 CHEGA-SVM 交通量预测模型，应用于城市交通量的预测，可以为公路建设项目的设计、施工和管理及相关政策、法规的出台提供科学的参考依据.

参 考 文 献

[1] Kim Kyoung-jae. Financial time series forecasting using support vector machines [J]. Neuron computing (S0925-2312), 2003, 55:307-319.

[2] Vapnik V N. An overview of statistical learning theory [J]. IEEE Transactions on Neural Networks (S1045-9227), 1999, 10(5):988-999.

[3] Keerthi K, Lin C J. Asymptotic Behaviors of Support Vector Machines with Gaussian Kernel [J]. Neural Computation, 2003,153 (3):1 667-1 689.

[4] Cherkessk V, Yunqian M A. Practical selection of SVM parameters and noise estimation for SVM regression [J]. Neural Networks, 2004, 17(1): 113-126.

[5] Rettemeir K, Falkenhagen B, Kongeter J. Risk assessment new trends in Germany[C]//ICOLD. The Proceedings of 21th Int Congress on Large Dams. Beijing: the International Commission on Large Dams, 2000: 625-641.

[6] Liu B, Wang L, Jin Y H, et al. Improved particle swarm optimization combined with chaos[J]. Chaos Solutions and Fractals, 2005, 25(5):1 261.

Prediction of Traffic Flow Using Support Vector Machine Optimized by Chaos Higher Efficient Genetic Algorithm

Kang Haigui　Li Mingwei　Zhou Pengfei　Zhao Zehui

(*Department of Engineering and Construction*, *Dalian University of Technology*, *Dalian* 116024, *China*)

Abstract: A SVM prediction model of traffic flow is built with kernel function of RBF, and chaos higher efficient genetic algorithm based on chaos map and genetic algorithm is used to preference the parameters C, ε and δ^2 of SVM, aiming at traffic flow prediction's feature of small sample, non-linear and complexity. At last simulation verification with observed passenger volume data over 1978-2008 was made. Compaing with GA-SVM model and tradition BP model, analysis represents that this model have better prediction result, and can be effectively applied to prediction of traffic flow.

Key words: traffic-flow prediction; support vector machine; acceleration genetic algorithm; chaos