第 37 卷 第 2 期
Vol.37 No.2

计 算 机 工 程
Computer Engineering

2011 年 1 月
January 2011

• 人工智能及识别技术 • 文章编号：1000—3428(2011)02—0178—02 文献标识码：A 中图分类号：TP181

基于 SVR 与微分进化策略的话务量预测

韩 锐[1]，贾振红[1]，覃锡忠[1]，常 春[2]，王 浩[2]

(1. 新疆大学信息科学与工程学院，乌鲁木齐 830046；2. 中国移动新疆分公司，乌鲁木齐 830063)

摘 要：采用支持向量回归机(SVR)与微分进化策略相结合的方法，对新疆 2 个地区的月平均忙时话务量进行预测。由微分进化策略良好的全局搜索性质，以预测平均相对误差为目标函数，对 SVR 的超参数进行寻优，利用优化后的 SVR 月平均忙时话务量进行预测。与传统的网格寻优算法和 RBF 神经网络方法进行比较，结果表明，SVR 的泛化能力与微分进化策略的搜索能力相结合，可以得到更好的预测效果。

关键词：微分进化策略；支持向量回归机；话务量预测

Telephone Traffic Load Prediction Based on SVR with DE-strategy

HAN Rui[1], JIA Zhen-hong[1], QIN Xi-zhong[1], CHANG Chun[2], WANG Hao[2]

(1. College of Information Science and Engineering, Xinjiang University, Urumqi 830046, China;
2. Xinjiang Mobile Communication Company, Urumqi 830063, China)

【Abstract】 Telephone traffic load of monthly busy hour in two states of Xinjiang are predicted by the method of Support Vector Regression(SVR) combining with Differential Evolution strategy(DE-strategy). The hyper-parameter of SVR is optimized via the DE-strategy and the MAPE criteria is defined as the objective function. Telephone traffic load of monthly busy hour is forecasted by the optimized SVR, the predicted result is compared with the method of grid search and RBF neural network. A better prediction result is obtained by the generalization property of SVR combining with searching property of DE-strategy.

【Key words】 Differential Evolution strategy(DE-strategy); Support Vector Regression(SVR); telephone traffic load prediction

DOI: 10.3969/j.issn.1000-3428.2011.02.061

1 概述

支持向量机理论自从提出以来[1]就因其良好的泛化能力和严密的数学结构受到学者们的广泛关注[2]。近几年，支持向量回归机(Support Vector Regression, SVR)作为一种预测工具，已经应用于医疗诊断[3]、链路负荷[4]以及浅海混响时间序列预测[5]等方面。SVR 的学习性能与其核函数的超参数有着很重要的联系[6]，因此，合适的超参数对 SVR 的预测能力有很大的帮助。微分进化策略[7]作为一种实值的群体智能优化算法，具有实现简单、可控参数少、鲁棒性强等特点，适用于求解全局优化问题，并对目标函数的可微性与约束条件没有要求，已在实际应用中取得良好效果[8]。利用微分进化策略优秀的全局搜索能力，可以对 SVR 的超参数进行搜索，从而达到预期的优化目标。本文通过忙时话务量的定义，计算新疆地区的每月话务总量，从而得到每月忙时平均话务量的时间序列，通过微分进化策略对支持向量回归机的超参数进行寻优，利用训练好的支持向量回归机对每月忙时平均话务量进行了时间序列预测。

2 相关原理及算法

2.1 月忙时平均话务量的定义

对每天产生的话务量进行每隔一小时的统计，一天就会统计 24 个话务量的值，对应每天的 0 点到 24 点，记录当天 24 个话务量值里的最大值作为对此日忙时话务量的统计。对当月每天的忙时话务量进行排序，去除最小的 8 个值及最大的 2 个值，再对剩余的数据取平均，由此构成当月的月忙时平均话务量。

2.2 支持向量回归机理论

对于一组给定的数据集：

$$T=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_i,y_i)\}\subset R^d\times R$$

其中，向量 $\boldsymbol{x}_i$ 为训练数据集的大小；R^d 为输入特征空间；y_i 为与之相对应的输出数据的大小，回归问题就是要估计出 $\boldsymbol{x}_i$ 与 y_i 的关系：

$$y=f(\boldsymbol{x})=<\omega,\Phi(\boldsymbol{x})>+b,\quad \boldsymbol{x}\in R^d,\quad y,b\in R \tag{1}$$

其中，$<>$ 对应在 R^d 空间的内积；$\Phi(\cdot)$ 为核函数，把训练数据映射到高维空间 F 上 $R^d\to F$，因此，在原空间上解决非线性问题就等同于在新的高维空间上解决线性回归问题。

机器学习理论对这一问题可以表述为在一组函数 $\{f(\boldsymbol{x},\omega)\}$ 中寻求一个最优的函数 $f(\boldsymbol{x},\omega^*)$ 使得预期的期望风险 $R(\omega)$ 达到最小：

$$R(\omega)\le R_{\text{emp}}(\omega)+\sqrt{\frac{h(\ln(2n/h)+1)-\ln(\eta/4)}{n}} \tag{2}$$

其中，n 为样本容量；h 为 VC 维。SVR 理论把式(2)转化为寻求如下问题的最优解：

$$\max W=-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)k(\boldsymbol{x}_i,\boldsymbol{x}_j)-\varepsilon\sum_{i=1}^{n}(\alpha_i+\alpha_i^*)+\sum_{i=1}^{n}y_i(\alpha_i-\alpha_i^*) \tag{3}$$

$$\text{s.t.}\begin{cases}\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)=0\\ 0\le\alpha_i,\alpha_i^*\le C,\quad i=1,2,\cdots,n\end{cases} \tag{4}$$

基金项目：中国移动新疆分公司研究发展基金资助项目

作者简介：韩 锐(1984－)，男，硕士研究生，主研方向：人工智能，模式识别；贾振红，教授、博士生导师；覃锡忠，副教授；常 春，高级工程师、硕士；王 浩，工程师、硕士

收稿日期：2010-07-18 **E-mail：**henry.006@163.com

其中，ε 由不敏感损失函数 $L(\boldsymbol{x}, y, f)$ 来定义，决定了回归曲线的平坦程度；C 为惩罚因子，表示对错分样本的惩罚。

由此，SVR 所求得的回归函数可以使式(1)改写为：

$$f(\boldsymbol{x})=\sum_{i=1}^{SV}(\alpha_i-\alpha_i^*)k(\boldsymbol{x}_i,\boldsymbol{x}_j)+b \tag{5}$$

其中，最常用的核函数为径向基核函数：

$$k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\gamma\left\|\boldsymbol{x}-\boldsymbol{x}_i\right\|^2) \tag{6}$$

由 KKT 条件可以知道，系数 $(\alpha_i-\alpha_i^*)$ 中只有一部分是非零值，并且训练样本的误差大于或等于 ε，这些训练样本就是支持向量。式(4)中的 C 以及式(6)中的 γ 被合称为 SVR 的超参数，对 SVR 的学习性能有着重要的影响。

2.3　微分进化策略理论

微分进化策略的基本思想为：对种群中的每个个体 i，从当前的种群中随机的选择 3 个点，以其中一个点为基础、另 2 个点为参照做一个扰动，所得点与个体 i 交叉以后进行“自然选择”，保留其中的较优者，实现种群的进化。不失一般性，设待求解的优化问题为 $\min\limits_{x\in R^n} f(x)$，则微分进化策略描述如下：

(1)初始化进化参数：种群规模 N，交叉概率 CR，交叉因子 F，进化代数 t，自变量下界 x_j^L 和上界 x_j^U，随机生成初始种群 $\{X_1(0), X_2(0),\cdots, X_N(0)\}$，其中，$X_i(0)=(x_1^{(i)}(0), x_2^{(i)}(0),\cdots, x_n^{(i)}(0))$。

(2)个体评价：计算每个个体 $X_i(t)$ 的目标值 $f(X_i(t))$。

(3)种群繁殖：对种群中的每个个体 $X_i(t)$，随机生成 3 个互不相同的整数 r_1，r_2，$r_3\in\{1,2,\cdots,N\}$ 以及随机整数 $j_{\text{rand}}\in\{1,2,\cdots,n\}$，

$$x_j^{(i)'}(t)=\begin{cases}x_j^{(r_1)}(t)+F(x_j^{(r_2)}(t)-x_j^{(r_3)}(t)) & \text{if } rand[0,1]<P_C \text{ or } j=j_{\text{rand}}\\ x_j^{(i)}(t) & \text{otherwise}\end{cases} \tag{7}$$

(4)选择：

$$X_i(t+1)=\begin{cases}x_j^{(i)'}(t) & \text{if } f(x_j^{(i)'}(t))<f(x_j^{(i)}(t))\\ x_j^{(i)}(t) & \text{otherwise}\end{cases} \tag{8}$$

(5)终止检验：如果种群 $X_i(t+1)$ 满足终止准则，则输出 $X_i(t+1)$ 中有最小目标值的个体作为最优解，否则转步骤(2)。

在本文的问题中，应当取 $n=2$，即需要优化的参数为 2 个，即 SVR 的超参数 C 及 γ。

3　实验及仿真结果

本文的实验数据来自新疆 2 个地区的月平均忙时话务量统计，从 2005 年 8 月到 2008 年 10 月共 39 个值，采用前 36 个值作为 SVR 的训练数据，后 3 个值作为测试数据。所有的数据首先都归整到 0~0.5 的范围内，测试的误差评价准则为平均相对误差准则，平均相对误差为：

$$MAPE=\frac{1}{p}\sum_{i=1}^{p}\frac{|y_i-\hat{y}_i|}{|y_i|} \tag{9}$$

其中，p 为预测步数；y_i 为数据的真实值；$\hat{y}_i$ 为 SVR 的预测值。ε 的取值设为 0.01，输入向量的维数选取 14，则支持向量回归机的输入值与目标值可以表述为：

$$\boldsymbol{X}=\begin{bmatrix}l_1\\ l_2\\ \vdots\\ l_{22}\end{bmatrix}=\begin{bmatrix}x_1 & x_2 & \cdots & x_{14}\\ x_2 & x_3 & \cdots & x_{15}\\ \vdots & \vdots & & \vdots\\ x_{22} & x_{23} & \cdots & x_{35}\end{bmatrix},\boldsymbol{Y}=\begin{bmatrix}x_{15}\\ x_{16}\\ \vdots\\ x_{36}\end{bmatrix} \tag{10}$$

对于微分进化策略的初始值设定，可以取得 $NP=\alpha\cdot n$，$\alpha\in[3,10]$，$F\geq 0.6$，$CR\geq 0.6$。本文取 $NP=12$，$F=0.8$，$CR=0.8$，迭代次数为 150 次。所要优化的目标函数为平均相对误差函数，选择常用的 2 种扰动策略作为微分进化策略的繁殖方式，并与传统的网格寻优法以及 RBF 神经网络进行比较。参数的搜索方式与取值范围如表 1 所示。

表 1　参数的搜索方式与取值范围

方法	扰动策略及搜索方式	取值范围
策略 1	$x_{\text{best}}+F\cdot(x_2-x_3)$	$C=100\sim 2\,600$，$\gamma=0.005\sim 0.95$
策略 2	$x_{\text{rand}}+F\cdot(x_2-x_3)$	$C=100\sim 2\,600$，$\gamma=0.005\sim 0.95$
网格寻优	$C=1.5^i$，$\gamma=1.5^j$	$i=1\sim 19$，$j=-9\sim 0$
RBF 神经网	径向基宽度参数 σ	$\sigma=5$

对新疆 2 个地区的月平均忙时话务量的后 3 个月的预测效果如图 1、图 2 所示。

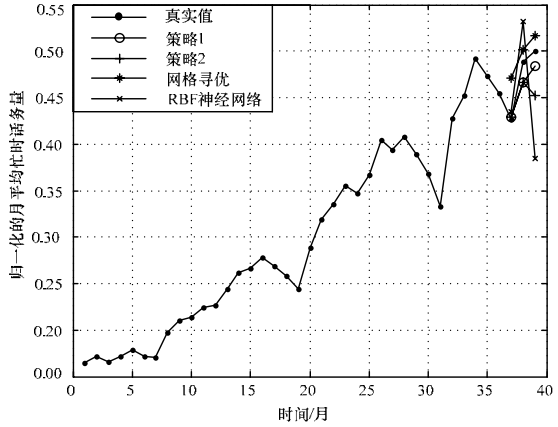

图 1　A 地区预测结果比较

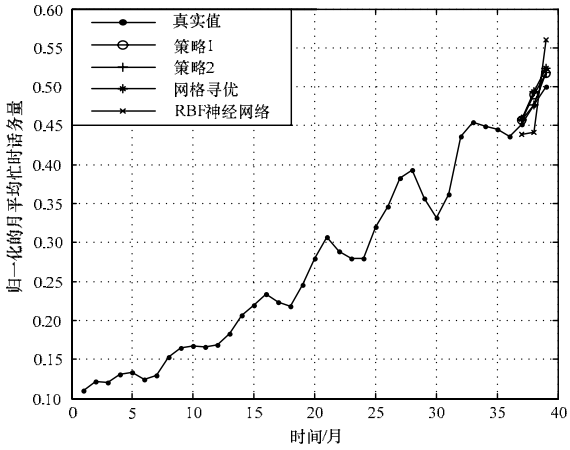

图 2　B 地区预测结果比较

可以看出，以微分进化策略为基础的 2 种搜索方法的效果明显好于网格寻优算法以及 RBF 神经网络。每种方法的平均相对误差由表 2 给出，因为微分进化策略初始化种群的随机性，所以每次的实验结果会略有不同。策略 1 和策略 2 的平均相对误差由程序单独运行 10 次后的平均值得出。

表 2　月平均忙时话务量预测的平均相对误差

方法	A 地区平均相对误差	B 地区平均相对误差
策略 1	4.37	2.74
策略 2	4.61	3.07
网格寻优	5.44	3.28
RBF 神经网	11.22	7.47

(下转第 182 页)