

文章编号:1001-9081(2009)09-2550-04

基于 ICA 和 SVM 的道路网短时交通流量预测方法

谢 宏, 刘 敏, 陈淑荣

(上海海事大学 信息工程学院, 上海 200135)

(hongxie@cie.shmtu.edu.cn)

摘要: 交通流量预测是智能交通系统(ITS)研究的一个重要课题。通过对多个观测点交通流量数据特点进行分析,采用一种基于独立成分分析(ICA)与支持向量机(SVM)相结合的短时交通流量预测方法。首先,通过独立成分分析得到同一条道路上各个观测点的交通流量的独立源信号;接着利用支持向量机预测模型对源信号进行建模和预测,并通过遗传算法(GA)优化参数;最后将其转换为交通流量数据,得到预测结果。实例分析结果显示,该算法优于直接利用支持向量机对交通流量进行预测的方法,并能去除同一条道路上多个观测点测量数据之间的相互影响。

关键词: 短时交通流量; 预测; 独立成分分析; 支持向量机; 遗传算法

中图分类号: U491.1; TP18 文献标志码:A

Forecasting model of short-term traffic flow for road network based on independent component analysis and support vector machine

XIE Hong, LIU Min, CHEN Shu-rong

(College of Information Engineering, Shanghai Maritime University, Shanghai 200135, China)

Abstract: Traffic flow forecasting is one of the important issues for the research of Intelligent Transportation System (ITS). Through analyzing the characteristics of data collected by different observation place on the same road, the authors proposed a new prediction method of short-term traffic flow in road network based on Independent Component Analysis (ICA) and Support Vector Machine (SVM). First, the traffic flow data of every observation point on the same road was turned into independent source signal through ICA method. Second, SVM model was used to train and predict the source signal, and through Genetic Algorithm (GA) parameters were optimized. At last, the traffic flow forecasting data were obtained by an inverse transform. Real traffic data were applied to test the proposed prediction model. The experimental results show that this method not only is more accurate than the method which uses SVM directly to predict traffic flow, but also can get rid of the data interaction of every observation points on the same road.

Key words: short-term traffic flow; forecasting; Independent Component Analysis (ICA); Support Vector Machine (SVM); Genetic Algorithm (GA)

0 引言

随着社会经济的快速发展,城市化的步伐加快,交通拥挤、交通事故、环境污染等问题使得城市的交通系统不堪重负。为了解决这些问题出现了智能交通系统(Intelligent Transportation System, ITS),它将先进的信息技术、电子通信技术、自动控制技术、计算机技术以及网络技术等有效、综合地运用于整个交通运输管理体系,其中对交通流量进行分析和预测是智能交通系统的研究的重要问题之一。

近年来,该领域的学者从各个方面提出了很多交通流量的分析和预测的模型,如时间序列预测模型、卡尔曼滤波模型、神经网络模型、基于混沌理论的模型和支持向量机预测模型^[1-5]等。其中支持向量机模型(Support Vector Machine, SVM)是一类基于结构风险最小化原则的新型机器学习算法。它解决了在神经网络等方法无法避免的局部最优问题,同时通过直接计算核矩阵巧妙地解决了“维数灾难”问题。这使得它成为近几年研究的热点。

城市中的交通系统都是由很多条道路共同组成的一个严密的道路网。在这个网络中,不同道路之间可以相互交叉形

成路口。车辆可以从任意一条道路进入道路网,通过路口进入另一条道路。因此同一条道路上相邻路口之间的交通流量是存在相互影响的。此时可以把同一条道路上不同路口之间的观测点看成一个整体来考虑,同时进行不同观测点的交通流量预测研究。国内外研究者已经进行了相关的研究,并取得了较好的效果^[6-8]。本文通过独立成分分析(Independent Component Analysis, ICA)从位于不同路口之间的观测点得到的混合信号中提取出源信号,再分别对各个观测点使用支持向量机模型进行预测。

1 独立成分分析

ICA 方法最早是由法国的 J. Herault 和 C. Jutten 于 20 世纪 80 年代中期提出来的。其后,Comon 在 1994 年发表的论文^[9],从理论上作了严密的讨论。ICA 成功解决了鸡尾酒会问题,给人们留下了深刻的印象。设 $s(t) = [s_1(t), \dots, s_m(t)]^T$ 是 m 维非高斯独立的源信号。 $x(t) = [x_1(t), \dots, x_n(t)]^T$ 是 n 维观测信号矢量,其每个观测信号分量都是 m 个独立源信号的线性组合,即:

收稿日期:2009-03-17;修回日期:2009-05-14。 基金项目:国家 863 计划项目(2007AA12Z152;2006AA09Z210)。

作者简介:谢宏(1962-),男,陕西汉中人,教授,博士,主要研究方向:人工智能及其应用系统; 刘敏(1985-),男,湖南常德人,主要研究方向:移动通信、无线接入; 陈淑荣(1972-),女,山西稷山人,副教授,硕士,主要研究方向:现代通信网络及控制、嵌入式系统应用。

$$\mathbf{x}(t) = \mathbf{A} \cdot \mathbf{s}(t) \quad (1)$$

其中, \mathbf{A} 是未知混合系数 a_{ij} 的 $n \times m$ 混合矩阵。这就是基本的 ICA 模型。

在这里, 独立成分 s_i 被称为隐变量, 是因为它们不能直接被观测到。另外模型中的混合系数 a_{ij} 也假设是未知的, 唯一能够观测到的是随机变量 \mathbf{x}_i 。ICA 的目标是找到一个对 \mathbf{x} 做线形变换的 $m \times n$ 变换矩阵 \mathbf{W} , 使得 \mathbf{x} 经过变换后得到的新矢量 $\mathbf{y}(t) = [y_1(t), \dots, y_m(t)]^T$ 的各分量之间尽可能的独立, 即:

$$\mathbf{y}(t) = \mathbf{W} \cdot \mathbf{x}(t) \quad (2)$$

$\mathbf{y}(t)$ 即为源信号矢量 $\mathbf{s}(t)$ 的估计值。从矩阵分析角度讲, ICA 的目的是寻找变换矩阵 \mathbf{W} 来实现多维观测信号的独立分量提取, 一旦求得 \mathbf{W} , 混合矩阵 \mathbf{A} 也就可以求出。

目前, 已有的 ICA 算法主要有: Herault-Jutten 算法、Cichocki-Unbehauen 算法、MLICA(最大似然)算法、非线性 PCA 算法、MMICA(互信息极小)算法、Infomax 算法和 FastICA 算法等^[10-11]。其中以 Infomax 和 FastICA 应用得最为广泛。

2 支持向量机

支持向量机是 AT&T Bell 实验室的 V. Vapnik 等人在 20 世纪 70 年代末提出并在 90 年代逐渐完善的一种针对分类和回归问题的统计学习理论, 是一类基于结构风险最小化原则的新型机器学习算法, 此处交通流量预测中要用到的是回归问题。

若 $f(\mathbf{x})$ 为线性模型, 在分类问题中, 即为线性可分的。根据结构风险最小化准则, 用于函数估计的支持向量机可以表示为:

$$\begin{aligned} & \min_{\omega \in \mathbb{R}^n, b \in \mathbb{R}} \frac{1}{2} \|\omega\|^2 \\ \text{s. t. } & \begin{cases} y_i - ((\omega \cdot \mathbf{x}_i) + b) \leq \varepsilon \\ ((\omega \cdot \mathbf{x}_i) + b) - y_i \leq \varepsilon \end{cases} \end{aligned} \quad (3)$$

其中 $i = 1, 2, \dots, l$; ε 是 ε -不敏感损失函数中参数, 是事先取定的一个正数。

求解式(3)一般采用对偶理论, 其对偶优化问题为:

$$\begin{aligned} & \min_{\alpha^{(*)} \in \mathbb{R}^{2l}} \frac{1}{2} \sum_{i,j=1}^l (\alpha_i^{*} - \alpha_i)(\alpha_j^{*} - \alpha_j)(\mathbf{x}_i \cdot \mathbf{x}_j) + \varepsilon \sum_{i=1}^l (\alpha_i^{*} + \alpha_i) - \sum_{i=1}^l y_i(\alpha_i^{*} - \alpha_i) \\ \text{s. t. } & \sum_{i=1}^l (\alpha_i - \alpha_i^{*}) = 0, \alpha_i^{(*)} \geq 0; i = 1, 2, \dots, l \end{aligned} \quad (4)$$

与 $\alpha_i \neq 0$ 和 $\alpha_i^{*} \neq 0$ 相对应的样本 \mathbf{x}_i , 即在不灵敏区边界上或外面的样本, 称为支持向量。从而有:

$$\omega = \sum_{i=1}^l (\alpha_i - \alpha_i^{*}) \mathbf{x}_i = \sum_{i \in SV_s} (\alpha_i - \alpha_i^{*}) \mathbf{x}_i \quad (5)$$

其中 SV_s 表示支持向量集。这样就得到决策函数的表达式:

$$f(x) = \sum_{i \in SV_s} (\alpha_i^{*} - \alpha_i) (\mathbf{x}_i \cdot \mathbf{x}) + b \quad (6)$$

实际问题并不一定是线性可分的, 如果继续坚持用超平面进行划分, 那必须“软化”对间隔的要求, 引入松弛变量和 ξ , 惩罚参数 C 对原始问题进行改写。或者用超曲面代替上面所研究的超平面, 这时候可以通过核函数: $K(\mathbf{x}, \mathbf{x}') = (\varphi(\mathbf{x}) \cdot \varphi(\mathbf{x}'))$ 将低维空间的非线性变换映射到高维特征空间, 然后通过高维空间中构造线性判别函数来代替输入空间的非线性判别函数, 详细介绍参看文献[12]。

3 相关参数的选取

3.1 核函数

常见的核函数有线性核函数、多项式核函数、高斯核函数、Sigmoid 核函数等。对于多项式核函数(线性核函数是多项式核函数的一个特例), 当特征空间维数很高时, 其计算量将大大增加, 而高斯核函数不存在这个问题。此外, 高斯核函数是正定核, Sigmoid 核函数是非正定的, 一般来说选取正定核作为映射的核函数会比非正定核效果要好, 而且高斯核函数需要调整的参数只有一个, 所以本文中选择高斯核函数:

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \cdot \|\mathbf{x} - \mathbf{x}'\|^2) \quad (7)$$

3.2 C 和 γ 的选取

在模型建立时, 惩罚参数 C 和高斯核函数中的参数 γ 的选取对模型的精度具有很大影响, 且目前尚无统一选择标准。本文中利用遗传算法进行最优参数的选择。

对支持向量的 C 和 γ 参数进行编码以搜索最优的参数。同时选取进化代数、停滞代数和停滞时间作为遗传算法的进化停止准则。经过多次迭代, 最终获得参数的优化结果。遗传算法虽然不一定得到最优解, 但可以在较短时间内得到满意解。

4 交通流量预测模型

通过上面的分析介绍, 本文提出的预测模型具体步骤如下。

1) 将同一条道路上的不同观测点在各个时刻观测得到的交通流量数据时间序列 $\{x(n)\}$ 作为混合信号, 通过 ICA 得到独立源信号 $\{s_1, s_2, \dots, s_n\}$ 和混合矩阵 \mathbf{A} 。

2) 分别将各个观测点的源信号 s_i 作为输入样本, 使用支持向量机算法训练模型。利用遗传算法选择最优的惩罚参数 C 和高斯核函数中的参数 γ , 建立预测模型。

3) 利用上一步建立的交通流量预测模型, 得到各个观测点源信号数据的 h 个时刻的预测值 (h 为大于 0 的整数)。

4) 将各个观测点源信号数据的预测值, 乘以混合矩阵 \mathbf{A} 进行还原, 就可以得到各个观测点交通流量的预测结果。

如上所述, 可以看出该种预测模型可以同时对多个观测点同时进行多步预测, 总结上述算法的计算流程如图 1 所示。

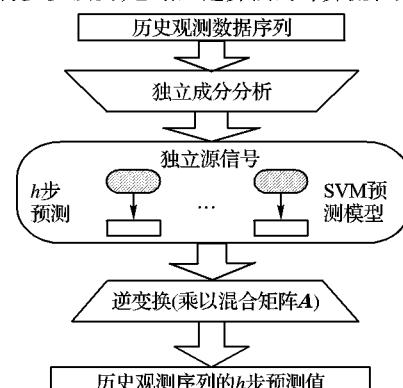


图 1 基于 ICA 和 SVM 的短时交通流量预测流程

5 实例分析

为检验以上所述预测算法用于实际交通流量预测的效果, 本文以 UMD CATT Lab 采集到的某城市环路上的四个观测点在 2008 年 11 月的交通流量为例进行计算^[13]。把 2008 年 11 月 1 日 0:00 至 30 日 24:00 时间段中的历史数据作为样本数据, 采样的时间间隔都是 1 小时。四个观测点每个都有 720 个样本, 从中随机选取 400 个样本建立模型, 剩下的

样本对建立的模型进行测试。

首先对观测样本 X 进行独立成分分析, 从各个观测点记录的混合信号中提取出独立源信号 S , 并得到混合矩阵 A 。结果如图 2 所示。

接着, 使用采用了高斯核函数的支持向量机模型训练模型。利用遗传算法选择合适的参数, 建立模型。再利用上述

模型, 对随机选取的测试样本进行预测。最后进行 ICA 反变换, 得到实际的交通流量数据。

另一方面, 作为与本文方法的对比, 对同一组历史观测数据的各个分量直接采用支持向量机预测模型进行建模和预测。在计算的结果中随机选取 24 个样本点的预测结果如图 3 所示。

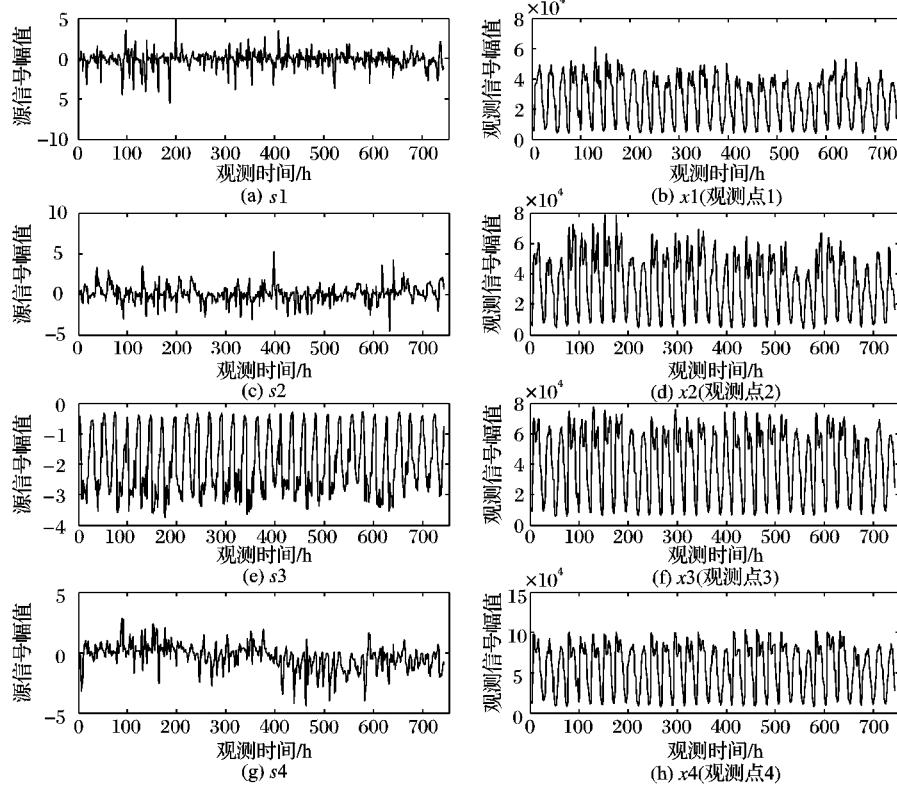


图 2 四个观测点的样本真实值和其通过 ICA 后得到源信号

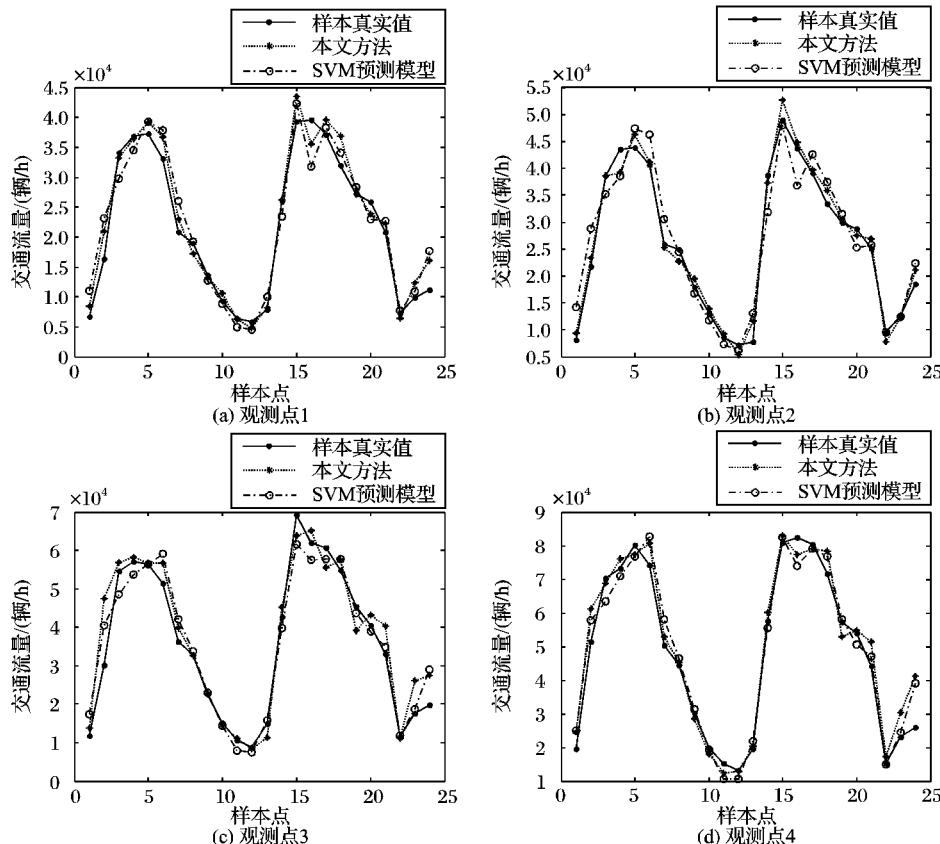


图 3 四个观测点随机选取的 24 个样本数据点的真实值和预测值

本文采用了均方根误差(Root Mean Square Error, RMSE)来检验该算法的预测效果。分别比较两种模型测试样本的预测RMSE、建模RMSE(将训练样本代入模型计算其预测值,再和真实值计算RMSE)。结果列在表1中。

表1 两种模型的均方根误差对比

观测点	预测 RMSE		建模 RMSE	
	本文方法	对比方法	本文方法	对比方法
观测点1	3 702.3	4 498.3	3 570.5	4 287.6
观测点2	4 236.2	5 954.5	4 218.2	5 315.0
观测点3	4 779.7	6 453.6	4 178.2	6 162.1
观测点4	6 498.7	8 597.9	5 875.6	8 347.8

从表1可得,本文提出的交通流量预测方法对同一条道路上四个观测点数据进行建模和预测得的预测误差、建模误差都小于直接使用SVM模型建模和预测,而且本文提出的模型的建模误差和预测误差非常接近,证明该方法相对于直接使用SVM模型对同一条道路上的多个观测点进行短时交通流量预测有着一定的优势和可行性。

6 结语

本文中首先使用独立成分分析对原始的交通流量观测数据进行预处理,再利用支持向量机模型进行预测,并利用遗传算法选择合适的参数。使用此方法对同一条道路上的多个观测点的交通流量进行预测时,不仅能够通过独立成分分析去除了各个观测点之间交通流量的相互影响,提高了模型的训练和预测精度。而且实现多个观测点交通流量的同时预测。但是任何模型都有自己的缺点。此预测模型的主要缺点是没有考虑其他影响交通流量的因素,如天气、交通事故等,并需要使用大量的历史数据去训练模型。

参考文献:

- [1] LU JIAN-CHANG, NIU DONG-XIAO, JIA ZHENG-YUAN. A study of short-term load forecasting based on ARIMA-ANN [C]// Proceedings of the 3rd International Conference on Machine Learning and Cybemetics. [S. L.]: IEEE Press, 2004: 3183–3187.
- [2] WANG YI-BING, PAPAGEORGIOU M, MESSMER A. Real-time freeway traffic state estimation based on extended Kalman filter: A general approach [J]. Transportation Research, 2007, 41(2): 167–181.
- [3] 尚宁,覃明贵,王亚琴,等.基于BP神经网络的路口短时交通流量预测方法[J].计算机应用与软件,2006,23(2): 32–33.
- [4] XIE HONG, LIU ZHONG-HUA. Short-term traffic flow prediction based on embedding phase-space and blind signal separation [C]// Proceedings of the 2008 IEEE Conference on Cybernetics and Intelligent Systems. [S. L.]: IEEE Press, 2008: 760–764.
- [5] 刘艳忠,邵小健,李旭宏.基于Lagrange支持向量回归机的短时交通流量预测模型的研究[J].交通与计算机,2007,25(5): 46–49.
- [6] STATHOPOULOS A, KARLAFTIS M G. A multivariate state space approach for urban traffic flow modeling and prediction [J]. Transportation Research, 2003, 11(2): 121–135.
- [7] KAMARIANAKIS Y, PRASTACOS P. Forecasting traffic flow conditions in an urban network: Comparison of multivariate and univariate approaches [J]. Transportation Research Record, 2003, 1857: 74–84.
- [8] 姚智胜,邵春福,熊志华,等.基于主成分分析和支持向量机的道路网短时交通流量预测[J].吉林大学学报:工学版,2008,38(1): 48–52.
- [9] COMON P. Independent component analysis – A new concept? [J]. Signal Processing, 1994, 36(3): 287–314.
- [10] BELL A J, SEJNOWSKI T J. An information-maximization approach to blind separation and blind deconvolution [J]. Neural Computation, 1995, 7(6): 1129–1159.
- [11] HYVÄRINEN A, OJA E. A fast fixed-point algorithm for independent component analysis [J]. Neural Computation, 1997, 9(7): 1483–1492.
- [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: 224–273.
- [13] CATT. CATT Laboratory [EB/OL]. [2009-01-10]. <http://www.cattlab.umd.edu>.

(上接第2549页)

此方法编程简单易于实现,计算量小,且不需要相空间重构。两种经典离散混沌系统的仿真实验证明此方法是有效可行的。

参考文献:

- [1] CHIARALUCE F, CICCARELLI L, GAMBI E, et al. A new chaotic algorithm for video encryption [J]. IEEE Transactions on Consumer Electronics, 2002, 48(4): 838–844.
- [2] 权安静,蒋国平,左涛,等.基于超混沌序列的分组密码算法及其应用[J].南京邮电学院学报,2005,25(4): 80–84.
- [3] 徐全生,李震,杜旭强.一种基于混沌序列的图像加密算法[J].小型微型计算机系统,2006,27(9): 1754–1756.
- [4] BENNETT C H, PETER G S, LI M, et al. Information distance [J]. IEEE Transactions on Information Theory, 1998, 44(4): 1407–1423.
- [5] 陈式刚.映象与混沌[M].北京:国防工业出版社,1992.
- [6] ECKMANN J P, RUELLE D. Ergodic theory of chaos and strange attractors [J]. Reviews of Modern Physics, 1985, 57(3): 617–656.
- [7] YANG ZHI-JIA, ZHAO GUANG-ZHOU. Application of symbolic techniques in detecting determinism in time series [C]// Proceedings of 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Washington, DC: IEEE Computer Society, 1998: 2670–2673.
- [8] 张雨.时间序列的混沌和符号分析及实践[M].长沙:国防科技大学出版社,2007.
- [9] HU J, TUNG W W, GAO J B, et al. Reliability of the 0-1 test for chaos [J]. Physical Review, 2005, 72(5): 1–5.
- [10] 丘水生,陈艳峰,吴敏,等.混沌保密通信的若干问题及混沌加密新方案[J].华南理工大学学报,2002,30(11): 75–80.