基于SVM的烟草销售量预测

刘 璐，丁福利，孙立民

（烟台大学 计算机与控制工程学院，山东 烟台 264005）

摘 要：烟草销售量预测能为烟草生产、运输、配送提供指导，使烟草行业能更好地适应市场需求。烟草销售量受众多因素的影响，具有季节性和周期性规律，传统的线性模型难以进行准确的预测。基于支持向量机建立烟草销售量的多维时间序列模型，实验结果表明，该模型具有较高的预测精度，能够准确地反映烟草销售量的变化趋势。对比实验也表明，所提出的方法比其它几种方法预测精度高，可以为烟草行业的销售管理提供科学依据，具有实用价值。

关键词：烟草销售量预测；支持向量机；多维时间序列

DOI：10.11907/rjdk.162026

中图分类号：TP319 **文献标识码**：A **文章编号**：1672-7800(2016)011-0134-03

0 引言

中国是世界上最大的烟草生产国和消费国[1]。烟草销售是烟草行业管理中最为关键的部分，准确的烟草销售预测能为烟草生产、运输、配送提供指导，而要进行准确的烟草销售预测必须找到合适的预测方法。因此，如何设计高精度的烟草销售预测方法是烟草行业管理的重要课题。

传统烟草销售量预测方法的研究主要集中在对烟草零售经营者订单的管理分析中，而且采用销售人员意见汇总法、德尔菲法（经理及员工的意见）等为主的人工预测方法[2]。这种人工预测方法业务流程较多，浪费大量的人力、物力，并且还可能引起烟草资源分配的不公平，难以满足市场需求。从机器学习的角度上看，烟草销售量的预测属于回归问题[3]，而回归包括线性回归和非线性回归。文献[4]在对烟草销售量数据进行分析的基础上，提出了一种线性预测模型，但由于烟草销售量受季节、人口、市场、节假日等一系列因素的共同影响，并不适合采用线性回归方法进行预测。在非线性回归方法中，较为常用的有神经网络和支持向量机（SVM）。文献[5]基于BP神经网络对烟草销售量进行建模并预测，而神经网络是基于经验风险最小化，不仅泛化能力较差，而且存在局部极小点问题[6]，因此神经网络虽然对原始数据的拟合能力较强，但对未来数据的推广能力较差，而对未来数据的推广能力往往更能反映学习机器的实用价值。支持向量机基于结构风险最小化，泛化能力强且预测精度高。因此，本文采用支持向量机方法对烟草销售量进行建模预测。

1 支持向量回归机

设已知训练集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_l,y_l)\}\in(X\times Y)^l$，其中 $x_i\in X=R^n$，$y_i\in Y=R$，$i=1,2,3,\cdots\cdots,l$。支持向量回归机首先通过引入核函数 $\Phi(x)$ 将输入空间中的 x_i 映射到高维空间中的 $\Phi(x_i)$[7]，并构造最优化问题：

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

$$s.t\begin{cases}[w\cdot\Phi(x_i)+b]-y_i\leqslant\varepsilon^*+\xi_i,i=1,2,\cdots,l\\ y_i-[w\cdot\Phi(x_i)+b]\leqslant\varepsilon^*+\xi_i^*,i=1,2,\cdots,l\\ \xi_i,\xi_i^*\geqslant 0,i=1,2,\cdots,l\end{cases} \tag{2}$$

其中，w 是高维空间中的权向量；C 是惩罚因子，用来调节置信区间和经验风险的权重[8]；ξ_i 和 ξ_i^* 是松弛变量；ε 为不敏感损失函数的参数；b 为回归函数的阈值[9]。

通过引入拉格朗日乘子，上述式子转化为原问题的对偶问题[7]：

$$min\ \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i^*-\alpha_i)(\alpha_j^*-\alpha_j)K(x_i,x_j)+\varepsilon\sum_{i=1}^{l}(\alpha_i^*+\alpha_i)-\sum_{i=1}^{l}y_i(\alpha_i^*-\alpha_i) \tag{3}$$

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

基金项目：山东省自然科学基金项目（ZR2014FQ016）

作者简介：刘璐（1992－），女，山东莱阳人，烟台大学计算机与控制工程学院硕士研究生，研究方向为机器学习；孙立民（1960－），男，山东莱西人，博士，烟台大学计算机与控制工程学院教授、硕士生导师，研究方向为机器学习、模式识别。本文通讯作者为孙立民。

求解该最优化问题，得到回归函数表达式：

$$f(x) = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)K(x_i, x) + b \tag{5}$$

2　预测方法

2.1　数据预处理

本文收集到了云烟品牌一个品类 2006 年 1 月～2011 年 10 月共 6 年的销售数据，销售数据信息中包括销售量、销售日期（年月日）、仓库编号、发票信息、审核人信息等，其中对销售量预测影响最大的是销售日期及对应的销售量。由于中国的香烟销售对阴历呈现出更强的规律性，因此将销售统计数据转换为以阴历月为标准。

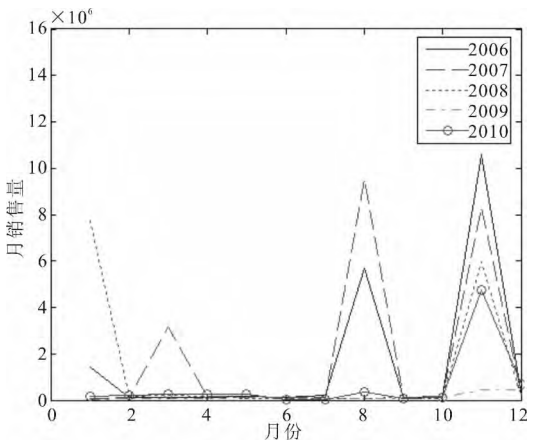

图 1　烟草销售量

图 1 是 2006 年－2010 年各月份的销售量，由图 1 可以看出：烟草销售与日期有着较大关系，并呈现出较强的周期性，在 8 月份和年末时期，烟草销售量最大。导致这一现象的原因是中秋节在 8 月份，春节在年末时期，这两个节日是中国重要的传统节日，由于香烟为逢年过节的代表性礼品，消费者购买量会大幅度增加。因此在这两个时间段，烟草的销售量将会大幅增加。

2006 年 1 月－2011 年 10 月的烟草销售量信息如表 1 所示。表 1 给出的是该品类月销售量数据。

表 1　月销售量

年份	月份	月销售量
2006	1	143 602
2006	2	11 354
2006	3	12 678
2006	4	12 876
2006	5	16 350
2006	6	10 297
2006	7	21 010
……	……	……
2011	9	103 190
2011	10	127 534

2.2　数据归一化处理

由表 1 可以看出，各列数据属性不同，数值范围相差较大。为避免数值范围较大的属性控制数值范围较小的属性，使数据具有统一性和可比性，将属性值都归一化[10]为[0，1]之间。归一化所用公式为：

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{6}$$

其中，x 表示原始数据，x' 表示归一化后的数据，$x_{\max}$ 为该属性的最大取值，$x_{\min}$ 为该属性的最小取值。

2.3　模型定阶

由于烟草销售量预测属于经济预测，因此它不仅与当前日期有关，更与之前的销售信息有关。为确定当前销售量与前多少个月的销售信息关系最大，需要通过拓阶[11]的方法来确定。

设烟草销售量数据的一个样本为 $\{y_i, year_i, month_i\}$，y_i 为第 i 个样本中的烟草销售量，$year_i$ 为当前年份，$month_i$ 为当前月份。其中，$year_i$ 和 $month_i$ 为样本的自变量，y_i 为样本的因变量。通过拓阶能够更为准确地得到自变量和因变量的函数依赖关系。当阶数为 n 时，表示将前 n 个样本中的信息添加到当前样本中的自变量中。即用前 n 个月的销售信息和当前年月来预测当前销售量。此时，自变量总数为 $(3\times n+2)$，其中 n 为阶数。通过 SVM 由低阶到高阶逐步进行拓阶，模型每拓一阶，自变量相应地增加 3 个。对于每一次的拓阶，以 MSE 最小为标准决定是否接受拓阶。设 $SVM(n)$ 为拓阶 n 次后的模型，$SVM(n+1)$ 为拓阶 $n+1$ 次后的模型，比较两者的 MSE 大小，如果 $SVM(n+1)$ 的 MSE 小于 $SVM(n)$ 的 MSE，表示接受本次拓阶，并进行下一步拓阶；如果 $SVM(n+1)$ 的 MSE 大于 $SVM(n)$ 的 MSE，表示不接受本次拓阶，并停止拓阶，最终得到最优阶数 n。通过对烟草数据的拓阶，得到拓阶结果如图 2 所示。

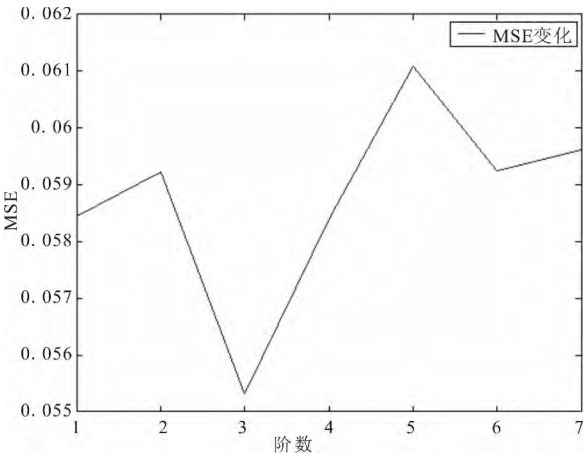

图 2　拓阶结果

由图 2 可以看出，当阶数为 3 时，MSE 最小，此时模型的效果最佳。因此，选择阶数为 3，并根据最优阶数 3 可得此时模型的自变量个数 $N=11$。

2.4　回归模型的参数选择

当训练模型确定后，通过支持向量回归机进行预测。由于径向基核函数的准确率较高，并且大多数 SVM 默认的核函数也是径向基核函数[12]，本文亦采用径向基核函数。

$$K(x,y)=\exp(-\frac{\|x-y\|^2}{2\sigma^2}) \tag{7}$$

模型中涉及的参数有惩罚因子 C，核函数中的参数 g（$g=\frac{1}{2\sigma^2}$），以及不敏感损失函数中的参数 ε^*。

SVM 的参数选择算法中较为常用的有 V. Cherkassky 公式法[13]、遗传算法[14]、网格搜索法[15]。这 3 种算法的训练时间依次增大，训练精度也随之增高。由于烟草预测系统对时间复杂度要求并不高，因此，本文采用网格搜索法获取参数。

设惩罚因子 C 的取值为 $\{2^{-10},2^{-9},2^{-8},\cdots\cdots,2^{0},\cdots\cdots2^{8},2^{9},2^{10}\}$，参数 g 的取值为 $\{2^{-10},2^{-9},2^{-8},\cdots\cdots,2^{0},\cdots\cdots,,2^{8},2^{9},2^{10}\}$，$\varepsilon^*$ 的取值为 $\{2^{-10},2^{-9},2^{-8},\cdots\cdots,2^{0},\cdots\cdots,2^{8},2^{9},2^{10}\}$。每个参数的取值有 21 种可能，一共有 21^3 组参数。验证参数优劣的方法采用 k－fold 交叉验证法[16]。该方法首先将样本集随机地分成 k 个互不相交的子集 $S_1,S_2,S_3,\cdots\cdots,S_k$，对于每一组参数，用 $k-1$ 个子集做训练样本，剩下一个子集做测试样本，一共进行 k 次实验，并求取 k 次试验的平均 MSE。最后，根据 MSE 最小原则得到最佳参数组。

根据上述方法，求得最佳参数组为 $C=0.25$，$g=32$，$\varepsilon^*=0.031\ 25$。根据该参数组获得最佳回归模型，并进行预测。

3 实验结果与分析

以云烟数据集为例，选择 2006 年 1 月－2010 年 12 月的销售量数据为训练样本，以 2011 年 1－10 月的销售量数据为测试样本。在本文算法实现过程中，实验环境配置如表 2 所示。

表 2 实验环境

主要硬件配置		主要软件配置	
中央处理器 CPU	内存 RAM	系统软件	应用软件
Intel(R) Pentium(R) 2.8GHZ2.8GHZ	2.00GB	Windows XP	LIBSVM MATLAB

根据上述方法建立模型，并进行预测。为了对比分析本文所采用方法的优劣，采用 AR、神经网络，基于 SVM 的一维时间序列模型对上述数据集进行训练和测试，所有预测结果都已列入表 3。表中第一列为真实销售数据，第二列为线性 AR 模型所预测的结果，第三列为 RBF 神经网络所预测的结果，第四列为基于 SVM 的一维时间序列模型所预测的结果，即仅仅通过以前的销售量来预测未来的销售量。第五列为基于 SVM 的多维时间序列模型，即本文所描述的方法。

采用 Libsvm[17] 中的两个评价标准（MSE 和相关系数）比较各模型的结果好坏。MSE 和相关系数 R 的计算方法如下：

令真实序列 U 为 $\{u_1,u_2,u_3,\cdots,u_n\}$，所预测得到的序列 V 为 $\{v_1,v_2,v_3,\cdots,v_n\}$。

表 3 实验结果

真实值	预测值			
	AR 模型	神经网络	基于 SVM 的一维时间序列模型	本文方法
4.488 5	2.596	6.008 3	6.544	3.863 4
4.662 9	1.767 9	7.896 2	7.399 9	7.651 4
8.283 2	1.825 3	10.839	7.668 2	7.339 1
8.815 9	3.018 1	7.951	7.737	6.805 7
8.212	3.1936	7.564 1	7.866 8	7.646 4
9.516 4	2.994 7	8.078 9	7.909 7	8.843 3
9.478 1	3.424 4	8.535 6	7.926 5	9.7682
10.71	3.4117	9.577 9	7.942	10.683
10.319	3.817 7	9.214 6	7.967 1	11.699
12.753	3.688 8	12.258	7.972 6	12.709

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(u_i-v_i)^2 \tag{8}$$

$$R=\frac{\sum_{i=1}^{n}(u_i-\bar{u})(v_i-\bar{v})}{\sqrt{\sum_{i=1}^{n}(u_i-\bar{u})^2}\sqrt{\sum_{i=1}^{n}(v_i-\bar{v})^2}} \tag{9}$$

MSE 为均方误差，反映了真实值与预测值之间的数值差异；相关系数反映了真实序列和预测序列变化趋势上的差异。相关系数的取值在[－1，1]之间。MSE 越小，相关系数越大，说明实验效果越好，反之亦然。

表 4 实验结果对比

评价标准	AR 模型	神经网络模型	基于 SVM 的一维时间序列模型	本文方法
MSE	36.934	2.616 5	5.441 2	1.701 8
相关系数	0.751 95	0.747 51	0.825 55	0.858 42

表 4 是 4 种模型的实验结果对比。由表 3 和表 4 可以看出，基于 SVM 的多维时间序列模型预测效果要好于其它模型。烟草销售量受多种因素的共同影响，不适合采用线性模型；神经网络模型是基于经验风险最小化的学习机器，对原始数据的拟合能力较强，但推广能力相对较差；基于 SVM 的一维时间序列模型仅仅通过之前的销售量的信息进行预测，但忽略了时间因素对销售量的影响，因而基于 SVM 的一维时间序列模型与多维时间序列模型相比，效果也较差。而本文方法所提到的基于 SVM 的多维时间序列模型，基于结构风险最小化，不仅选定了样本的最佳维数，更充分考虑了对预测有利的信息进行预测，取得较好的预测效果。因此，本文建立的模型是合理有效的。

4 结语

通过预测烟草销售量可以提前了解烟草的销售动态，为烟草物流、仓储等部门提供决策依据。本文基于支持向量机建立烟草销售预测的多维时间序列模型。实验证明，根据本文方法建立的模型所预测的结果与实际结果基本一致，能够比较准确地反映烟草销售量的变化趋势。对比实验也证明，与其它几种方法相比，本文方法预测误差最小。综上，本文所述方法是合理有效的，可以应用到实际烟草销售量预测中。

智慧城市电子政务云平台构建

周 冰,张志刚

(焦作大学 信息工程学院,河南 焦作 454000)

摘 要:电子政务发展不仅有利于社会资源配置,而且对社会政治、经济、文化发展,推动社会文明进步起到举足轻重的作用。电子政务不仅仅是政务的电子化,还是智慧城市的一个重要方面。探讨云计算技术环境下,电子政务云平台的服务模式、顶层设计思想、云平台网络结构及系统架构。

关键词:智慧城市;电子政务;云平台;云计算;架构

DOI:10.11907/rjdk.162183

中图分类号:TP319 **文献标识码**:A **文章编号**:1672-7800(2016)011-0137-03

1 电子政务概述

我国城市发展经历了工业型、经济型两个阶段,现在正经历着"数字化"城市发展阶段。智慧城市发展的关键在于政府的顶层设计,其首要任务是建设智慧型政府,智慧型政府是带动经济建设、社会发展的关键因素。做好电子政务平台架构,充分收集数字资源并发挥资源优势对实现智慧城市具有重要的意义。

智慧政府是信息技术发展的必然,它将改变传统政府的服务模式,不论是主导思想、工作机制、流程、方法还是办公模式,都将发生巨大变化,这些变化随着近些年各地电子政务的推进已逐渐呈现。电子政务使传统的政府自上而下的政务工作结构变成同一层面点对点的网络架构。实现电子政务后,公众和政府的信息交互可以直接通过电子政务平台进行,政府也由管理型向服务管理型转化。电子政务数字化平台成为人民群众和政府之间沟通的桥梁,政府政策信息发布以及人民群众意见建议反馈等都可以通过数字化平台进行。电子政务是提高政务办公效率的重要工具,也是信息化社会发展的基础。

参考文献:

[1] 蒋德珺.我国烟草业国际化战略研究[J].北方经济,2012(14):94-95.
[2] 利普·科特勒,洪瑞云,梁绍明,等.市场营销管理[M].亚洲版·2版.北京:中国人民大学出版社,2001.
[3] 郑逢德,张鸿宾.拉格朗日支持向量回归的有限牛顿算法[J].计算机应用,2012,32(9):2504-2507.
[4] 张素平.基于乘法模型的内蒙古乌兰察布市卷烟总销量预测研究[J].内蒙古科技与经济,2012(21):33-35.
[5] 仲东亭,张玥.BP神经网络对烟草销售量预测方法的改进研究[J].工业技术经济,2007,26(9):115-118.
[6] 刘苏苏,孙立民.支持向量机与RBF神经网络回归性能比较研究[J].计算机工程与设计,2011,32(12):4202-4205.
[7] 邓乃扬,田英杰.数据挖掘的新方法——支持向量机[M].北京:科学出版社,2004
[8] 肖建,于龙,白裔峰.支持向量回归中核函数和超参数选择方法综述[J].西南交通大学学报,2008,43(3):297-303.
[9] 单黎黎,张宏军,张睿,等.基于主导因子法的装备维修保障人员调度值预测[J].计算机应用,2012,32(8):2364-2368.
[10] 彭丽芳,孟志青,姜华,等.基于时间序列的支持向量机在股票预测中的应用[J].计算技术与自动化,2006,25(3):88-91.
[11] 向昌盛,周子英.基于支持向量机的害虫多维时间序列预测[J].计算机应用研究,2010,27(10):3694-3697.
[12] 谭征,孙红霞,王立宏,等.中文评教文本分类模型的研究[J].烟台大学学报:自然科学与工程版,2012,25(2):122-126.
[13] CHERKASSKY V,MULIER F. Learning from data: concepts, theory and methods[M]. NY:JohnViley&Sons,1997.
[14] YONG M,XIAO-BO Z,DAO-YING P,et al. Parameters selection in gene selection using Gaussian kernel support vector machines by genetic algorithm[J]. Journal of zhejiang university science B,2005,6(10):961-973.
[15] 王兴玲,李占斌.基于网格搜索的支持向量机核函数参数的确定[J].中国海洋大学学报:自然科学版,2005,35(5):859-862.
[16] ITO K,NAKANO R. Optimizing support vector regression hyperparameters based on cross-validation[C]. Proceedings of the International Joint Conference on Neural Networks, 2003:2077-2082.
[17] HSU C W,CHANG C CLIN C J. LIBSVM:a library for support vector machines[EB/OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

(责任编辑:孙 娟)

基金项目:河南省软科学研究项目(152400410355)

作者简介:周冰(1981-),男,河南温县人,硕士,焦作大学信息工程学院讲师,研究方向为物联网技术、智慧城市;张志刚(1976-),男,河北张家口人,硕士,焦作大学信息工程学院讲师,研究方向为智慧城市、计算机应用。