粮食产量预测的支持向量机模型研究

向昌盛[a]，周子英[b]，武丽娜[b]
（湖南农业大学 a.东方科技学院；b.资源环境学院，湖南 长沙 410128）

摘　要：粮食产量的历史数据有限，是典型的小样本数据，又由于粮食产量受不确定性因素的影响，是一个复杂的非线性系统，以往的粮食产量时间序列预测模型的阶数采用经验方法或线性方法来确定，得到的预测精度不理想。针对这些问题，可将支持向量机引入到时间序列模型定阶的方法中，然后采用留一法交叉验证寻找最优参数，建立一个多输入、单输出的预测模型。通过对中国粮食产量进行仿真实验，并与一次滑动平均、ARIMA、LS_SVM 和 RBF 神经网络的预测模型作比较来验证模型的有效性，结果表明该模型具有较高的预测精度和较强的泛化能力，证明了该模型对近期粮食产量的预测是可靠的。

关　键　词：粮食；产量预测；支持向量机；时间序列
中图分类号：F323.5　　**文献标志码**：A　　**文章编号**：1009-2013(2010)01-0006-05

Study on the support vector machines model of grain production prediction

XIANG Chang-sheng[a], ZHOU Zi-ying[b], WU Li-na[b]
(a.College of Orient Science & Technology; b. College of Resources & Environment, Hunan Agricultural University, Changsha 410128,China)

Abstract: The grain production is complicated and unpredictable. In attempting to predict time series data, statistical methods are the major research stream in tradition, however, the factor of time is often overlooked in tradition model. In this paper, a model with multi-input and single output was proposed based on improved support vector machines(SVM) using leave one out for cross validation. China grain production data sets are predicted using this method; the results show that the SVM predictor has higher precision and greater generalization ability.

Key words: grain; production prediction; support vector machines; time series

一、问题的提出

农业是国民经济的基础，粮食生产是农业的关键。预测未来一段时间粮食产量的变化情况，可以为政府科学决策提供依据，也便于为粮农、企业提供决策参考。粮食产量数据是一种高度不稳定、复杂且难以预测的时间序列数据，这些数据往往既隐含大量的动态特征，又受自变量的影响，同时具有高度的非线性[1]。国内外的相关研究中，不少学者构建了许多很有价值的理论假说和预测模型，主要包括时间序列方法、回归分析法和人工神经网络方法等。时间序列方法中最具代表性的差分自回归移动平均(autore-gressive integrating moving average，ARIMA)，ARIMA是基于线性数据的模型，但粮食时间序列数据往往更多地表现为非线性且含有复杂的噪声，这样基于线性模型定阶获得的模型阶数和保留变量的ARIMA模型，往往并非最优，从而导致预测精度不高[2]。采用回归分析法[3]是基于固定模型的，而粮食产量是非线的，而且具有混沌性，所以采用回归分析法无法捕捉到粮食产量数据的非线性特征，预测结果失真。王启平[4]、禹建丽[5]等利用神经网络对粮食产量进行了预测，取得了不错的效果，但是神经网络要求数据样本大，而粮食产量数据属于小样本数据，所以在预测过程中常出现结果过拟合，泛化能力不强等现象。Vapnik等[6]根据统计理论提出了支持向量机(support vector

收稿日期：2009-11-16
基金项目："十一五"国家科技支撑计划(2008BADA 4B108)
作者简介：向昌盛(1971—)，男，湖南怀化人，副教授，博士，主要从事计算数学、智能预测研究。

machines，SVM)学习方法，SVM的最大特点是改变了传统的神经网络基于经验风险最小化原则，较好地解决了小样本、非线性、过拟合、维数灾和局极小等问题，SVM已经逐步应用于各个领域的预测中[7-9]，李晓东[10]等利用最小二乘支持向量机(least squares support vector machines，LS_SVM)对我国粮食产量进行了预测，得到的结果比较好。但LS_SVM通过构造损失函数将原支持向量机算法的二次优化问题转化为求解线性方程，牺牲了预测精度，失去了SVM稀疏性的优点，不能保证是全局最优解[11,12]。

总之，粮食产量的历史数据有限，是典型的小样本数据，同时由于粮食产量受不确定性因素影响的，是一个复杂的非线性系统，以往的粮食产量时间序列预测模型的阶数采用经验方法或线性方法来确定，得到的预测精度不理想，针对这些问题，笔者将SVM引入到粮食产量的模型定阶和预测中，建立一种基于支持向量机的预测模型，并与一次滑动平均、ARIMA、LS_SVM 和 RBF 神经网络的预测模型作了比较来验证模型的有效性。

二、粮食产量预测的支持向量机模型

1. 时间序列模型的定阶

建立时间序列模型最为关键的、最困难的是确定模型的阶数。近年来，许多专家和学者对如何确定时间序列的时滞阶数进行了深入研究[13]，模型的阶数选取方法在实践中有两种：一是靠经验选择，二是先设定模型其他参数方法，然后对滞后阶数按照一定的标准进行优化。第一种方法过分依赖研究者的水平和经验，不能客观选择最佳滞后阶数。而第二种方法则忽略了一个重要问题：滞后阶数与其他参数对模型好坏的影响是相互的，如果先人为设定其他参数选择滞后阶数，而后使用该滞后阶数值优化参数，很有可能只是在该滞后阶数下的参数局部最优，而不是全局最优。同时这两种方法都是针对线性模型的，得到的模型对线性时间序列比较理想，但都忽略了时间序列的非线性关系，这样导致对非线性模型，其预测准确性受到很大的限制。针对此问题，笔者将支持向量机引到非线性时间序列模型定阶中，具体步骤如下：

(1) 假定一个原始时间序列 n 个样本，首先强制性的拓一阶，用留一法进行交叉验证进行支持向量机测试，得到模型的方根误差(Root Mean Square Error，RMSE)；

(2) 继续进行下一轮的拓阶，利用相同的方法进行测试，得到本轮的RMSE，对待比较的相邻两模型 SVM(n) 和 SVM(n+1) 的 RMSE，如果 $RMSE_{SVM(n)} > RMSE_{SVM(n+1)}$($RMSE_{SVM(n)}$为SVM(n)的均方误差，$RMSE_{SVM(n+1)}$为SVM(n+1)的均方误差)，重复第(2)步，否则转到下一步；

(3) 拓阶停止，此时取 SVM(n)模型的阶数为时间序列模型的滞后阶数。

2. 时间序列预测模型的建立

假设观察到的时间序列为 $(z_1, z_2, \cdots, z_t)$，时间序列的 SVM 预测数学关系表达式为：

$$z_{t+1} = F(z_t, z_{t-1}, \cdots, z_{t-m}) \tag{1}$$

式中：$z_t, z_{t-1}, \cdots, z_{t-m}$ 为历史数据，m 为滞后阶数，z_{t+1} 为 $t+1$ 时刻的预测目标数值。具体的建模步骤如下：

步骤 1：数据预处理。为了提高运算速度和预测精度，对样本进行归一化处理，预测处理变换关系式如下：

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{2}$$

式中：x'为归一化后的值，$x_{\max}$，$x_{\min}$ 分别为原数据序列中的最大值和最小值。

步骤2：模型的定阶。根据本文提出的基于SVM的非线性定阶方法确定时间模型的滞后阶数，即SVM 时间序列模型的输入向量的个数，这样就可以建立一个多输入、单输出支持向量机预测模型。按照式(3)得到支持向量机输入向量和输出向量。

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \\ x_2 & x_3 & \cdots & x_{m+1} \\ \cdots & \cdots & \cdots & \\ x_{n-m} & x_{n-(m-1)} & \cdots & x_{n-1} \end{bmatrix}, Y = \begin{bmatrix} x_{m+1} \\ x_{m+2} \\ \cdots \\ x_n \end{bmatrix} \tag{3}$$

步骤 3：核函数选择。据泛函的有关理论，只要一种核函数K（x,x_i）满足mercer条件的对称函数均可作为核函数，但是对于特定的问题，如何选择最合适的核，一直是困扰研究者的一个难点，针对

此问题，很多研究和实验表明[14]，当缺少过程的先验知识时，选择径向基函数比选择其它核函数效果好。本文核函数采用径向基函数核函数(RBF)。RBF 函数定义如下：

$$k(x,x_i)=\exp(\frac{-\|x-x_i\|^2}{\delta}) \quad (4)$$

这样其回归函数可以表示为：

$$f(x)=\sum_{i=1}^{l}(a_i-a_i^*)\exp(\frac{-\|x-x_i\|^2}{\delta})+b \quad (5)$$

式中：其中 l 为支持向量的个数，x_i 为作为支持向量的样本因子向量，x 为待预报因子向量，a_i、a_i^* 和 b 为建立 SVM 模型待确定的系数，δ 为核参数。

步骤 4：模型参数寻优及验证。利用 SVM 对训练样本集进行学习，通过交叉验证结果找出 SVM 模型的最优参数，利用最优参数对验证集进行检验，检验模型的泛化能力。

3. 参比模型及评价指标

为了考察 SVM 模型的优劣，同时采用一次滑动平均、ARIMA 模型、LS_SVM 和 RBF 神经网络进行对比实验，所有模型均采用一步预测法。一次滑动平均和 ARIMA 模型由 DPS6.5 给出，RBF 神经网络和 LS_SVM 自编程序通过 MATLAB7.0 调用神经网络工具箱和 LS_SVM1.5 工具箱；SVM 采用 LIBSVM2.86 版本在 MATLAB7.0 平台下实现，LIBSVM 算法是一种将序贯最小优化算法(Sequential Minimal Optimization，SMO)和 Svmlight 算法相结合的优化方法，对工作集的选择策略有所改进，各方面性能优于标准 SVM。为了评价模型预测性能的优劣，RMSE 和 MAPE 作为模型的评价指标，RMSE 和 MAPE 分别定义如下：

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{i}(y_i-\hat{y}_i)^2} \quad (7)$$

$$MAPE=\frac{1}{n}\sum_{i=1}^{i}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \quad (8)$$

其中，y_i 为真值，$\hat{y}_i$ 为预测值，n 为预测样本数。RMSE 仅适用于同一数据集不同模型间的比较，MAPE 可用于不同数据集间的比较。但对同一数据集，如 A 模型与 B 模型相比虽 MAPE 较大而 RMSE 较小，则 A 模型预测更为稳健，因此，RMSE 为主要评价指标。

三、模型的仿真实验

1. 数据来源

分析粮食产量的重要性不但体现在对历史数据的拟合上，对未来短期预测也同样重要，准确的预测可以帮助调节粮食需求的市场平衡性，防止出现因为粮食供应不足或过多产生的不良的市场效应。笔者选用的数据为 1978—2007 年我国的粮食产量，数据来自于 2008 年《中国统计年鉴》，如表 1 所示。将数据分为两部分，1978—2000 年的实际产量作为训练样本来拟合和建模，2001—2007 年的实际产量作为检验样本来检验模型的泛化能力。

表 1 1978—2007 年我国粮食产量

年份	粮食产量/万吨	年份	粮食产量/万吨
1978	30 476.5	1993	45 648.8
1979	33 121.2	1994	44 510.1
1980	32 055.5	1995	46 661.8
1981	32 150.2	1996	50 453.5
1982	35 145.0	1997	49 417.1
1983	38 172.8	1998	51 229.53
1984	40 173.1	1999	50 838.58
1985	37 910.8	2000	46 217.52
1986	39 115.1	2001	45 263.67
1987	40 129.8	2002	45 705.75
1988	39 140.8	2003	43 069.53
1989	40 175.5	2004	46 946.95
1990	44 624.3	2005	48 402.19
1991	43 529.3	2006	49 804.23
1992	44 265.8	2007	50 160.28

2. 模型建立与验证

根据模型定阶得到滞后阶数 m=3，这就意味着可以将前三年的粮食产量 SVM 的输入来预测当年的粮食产量。按照式(3)，表格 1 中的数据可构造样本 27 组，将前 20 组作为训练样本，后 7 组作为检验样本，选用 RBF 核函数，利用 gridregression.py 和留一法交叉验证进行模型参数自动寻优，确定最优模型的核函数参数，利用建立的最优模型对 1987—2000 年的数据进行拟合，然后模型对 2001—2007 年的粮食产量进行验证，检验模型的泛化能力，结果分别见表 2 和表 3。

3. 2008—2010粮食产量预测

用上述训练好的支持向量机模型对我国2008—2010年的粮食产量进行预测，结果如表4。

4. 结果分析与讨论

由表 2，表 3 及表 5 可知，线性模型一次滑动平均和 ARIMA 预测精度不高，主要是不能捕捉到粮食数据的非线性特征；RBF 神经网络精度也不理想，是因为粮食数据属于小样本数据，容易出现局部最优，导致泛化能力差。在所有的模型中，SVM 的拟合和泛化预测能力是最好的，这是由于 SVM 基于结构风险最小，较好地解决了小样本、非线性、过拟合、维数灾和局极小等问题，泛化推广能力优异。同时说明了基于支持向量机的非线性定阶方法挖掘了时间序列数据之间隐含的非线性关系，使非线性时间序列模型的定阶得到了很好的解决。

利用 SVM 对 2008—2010 年我国粮食产量的预测，从表 4 得知：我国粮食产量在这三年内将稳定在 5.1~5.4 亿吨之间，波动不大，基本满足我国人口的增长需要。近年来我国农业生产结构调整力度继续加大，减少了粮食作物的种植面积；西部省份继续落实国家退耕还林、还草的政策，劳动力外出打工等导致粮食种植面积有所减少；但由于国内外粮油价格、科技技术进步和农业生产力发展，粮食单产有所提高，这样粮食总产量会维持在 5 亿吨以上，这证明本模型的预测结果符合我国粮食产业发展趋势。

表 2　各种模型对 1978－2000 年粮食产量的拟合值　　万吨

年份	真值	一次滑动平均	ARIMA	BP 神经网络	LS_SVM	SVM
1978	30 476.5	31 757.4	28 396.6	30 361.5	30 333.41	30 255.7
1979	33 121.2	32 202.5	31 126.4	31 177.4	31 667.65	32 115.3
1980	32 055.5	33 823.5	34 142.6	32 236.4	32 246.87	32 045.5
1981	32 150.2	33 877.4	34 184.2	33 141.8	32 494.95	32 133.39
1982	35 145	33 600.0	33 159.9	35 342.5	34 753.44	35 184.49
1983	38 172.8	35 107.9	36 451.9	38 160.8	38 095.15	38 203.71
1984	40 173.1	37 920.4	39 347.8	39 182.7	40 022.46	40 061.08
1985	37 910.8	40 823.6	41 177.3	39 176.4	38 679.72	38 226.61
1986	39 115.1	40 812.5	38 183.5	39 134.4	39 017.74	38 911.36
1987	40 129.8	41 004.6	40 151.3	39 113.7	39 842.16	39 988.63
1988	39 140.8	41 389.7	41 435.6	39 165.9	41 433.08	39 126.3
1989	40 175.5	40 701.1	40 235.4	41 054.8	40 743.43	40 057.28
1990	44 624.3	40 681.5	42 171.4	43 181.7	41 513.46	44 559.06
1991	43 529.3	43 471.1	43 771.6	44 257.2	43 441.34	43 327.98
1992	44 265.8	44 524.4	44 143.5	45 193.4	44 972.13	44 113.76
1993	45 648.8	45 382.8	45 171.9	46 167.2	44 939.06	44 857.17
1994	44 510.1	46 531.1	46 185.5	47 234.9	45 592.41	45 023.97
1995	46 661.8	46 241.0	48 130.1	48 110.9	45 229.51	46 136.39
1996	50 453.5	47 135.0	49 321.2	48 161.3	47 875.46	50 164.22
1997	49 417.1	50 004.9	50 123.4	49 214.2	49 557.19	49 044.43
1998	51 229.53	50 997.9	51 148.7	49 558.7	50 544.32	50 922.67
1999	50 838.58	52 406.6	53 147.1	49 694.8	50 321.3	50 668.61
2000	46 217.52	52 687.9	54 348.7	50 122.7	46 717.09	46 127.24

表 3　各种模型对 2001－2007 年粮食产量泛化值　　万吨

年份	真值	一次滑动平均	ARIMA	BP 神经网络	LS_SVM	SVM
2001	45 263.67	49 452.9	45 032.85	43 539.93	45 336.14	45 174.92
2002	45 705.75	46 396.1	45 497.13	44 157.82	45 562.03	45 361.39
2003	43 069.53	44 774.0	44 117.25	42 219.76	45 408.49	42 772.73
2004	46 946.95	42 350.4	46 693.86	46 037.01	46 938.61	46 667.54
2005	48 402.19	43 693.0	48 941.51	48 603.85	49 001.91	48 350.91
2006	49 804.23	46 040.0	50 026.67	51 354.47	49 749.47	49 402.81
2007	50 160.28	48 724.0	50 409.5	50 713.25	50 868.37	49 949.57

表 4 2008－2010 我国粮食产量的预测值

	2008年	2009年	2010年
粮食产量/万吨	50 651.23	51 932.15	53 213.08

表 5 各种模型 RMSE 和 MAPE 比较

预测模型	RMSE	MAPE/%
一次滑动平均	4 180.4782	8.380 68
ARIMA	3 111.663	3.369 88
RBF神经网络	1 886.705	3.455 35
LS_SVM	953.2484	1.240 51
SVM	269.6404	0.537 31

四、结论

粮食产量序列数据具有非线性特征，存在弱混沌特性，因此采用精确的数学模型预测比较困难。支持向量机适合于小样本的粮食产量建模与预测，而对模型的定阶方法进行改进后，可以较好地拟合粮食产量时间序列模型，获得较高的回归与建模精度，其外推预测也具有较高的泛化能力。由于粮食生产是自然再生产与经济再生产的统一，粮食产量预测是受政策、自然环境、资源投入等多因素的影响，本文没有考虑到这些因素的影响，只考虑粮食产量历史数据，这是今后要进一步研究的方向。

参考文献：

[1] 陈成忠，林振山．山东省粮食产量波动的多时间尺度分析[J]．农机化研究，2007(10)：1-4.

[2] 张晓杰，张希良．时间序列分析模型在山东省粮食总产量预测中的应用[J]．水土保持研究，2007，14(3)：309-311.

[3] 吕效国，王晓燕，孙建平，吴梅君．用自回归预测的新方法预测粮食产量[J]．安徽农业科学，2008，36(33)：14 359-14 374.

[4] 王启平．BP 神经网络在我国粮食产量预测中的应用[J]．预测，2002，21(3)：79-80.

[5] 禹建丽，黎 娅．基于人工神经网络的粮食产量预测模型[J]．河南农业科学，2005(7)：44-46.

[6] Vapnik V．The nature of statistical learning theory [M]．New York：Springer-Verlag，2000.

[7] 刘 涵，刘 丁，李 琦．基于支持向量机的混沌时间序列非线性预测[J]．系统工程理论与实践，2005(9)：94-99.

[8] 曲文龙，樊广佺，杨炳儒．基于支持向量机的复杂时间序列预测研究[J]．计算机工程，2005，31(23)：1-3.

[9] 罗正根．机器学习新方法：支持向量机[J]．湖南农业大学学报：社会科学版，2009(2)：111-114.

[10] 李晓东，席升阳，潘 立．基于最小二乘支持向量机的中国粮食产量预测模型研究[J]．水土保持研究．2007，14(6)：321-324.

[11] 任勋益，王汝传，谢永娟．基于支持向量机和最小二乘支持向量机的入侵检测比较[J]．计算机科学，2008，35(11)：83-85.

[12] 阎威武，邵惠鹤．支持向量机和最小二乘支持向量机的比较及应用研究[J]．控制与决策[J]，2003，18(3)：358-360.

[13] 袁哲明，张永生，熊洁仪．基于 SVR 的多维时间序列分析及其在农业科学中的应用[J]．中国农业科学，2008，41(8)：2 485-2 492.

[14] Kim H S，Eykholt R J D．Salas．Nonlinear dynamics，delay times and embedding windows[J]．Physica D，1999(127)：48-60.

责任编辑：李东辉