文章编号:1005-0523(2010)05-0067-05

基于多输出支持向量机的物流量预测研究

骆世广[1],叶 赛[2],胡 蓉[1]

(广东金融学院 1.应用数学系; 2.广东金融学院 计算机科学与技术系,广东 广州 510521)

摘要:物流量预测问题受众多因素影响,而已有的方法都是用多输入单输出模型进行预测,因此难以获得满意的预测效果。一种多输出支持向量机的方法用于广州市的物流量的预测中,为了与单输出预测相比,选取自适应迭代支持向量机方法进行预测。结果表明,多输出支持向量机的预测是有效的。

关 键 词:物流量预测;自适应迭代支持向量机;多输出支持向量机

中图分类号:U491.14 **文献标识码**:A

随着经济一体化步伐的加大,区域物流系统规划、管理等面临着更大的挑战。高效的物流系统规划依赖于准确的区域物流量预测。传统的区域物流量预测方法有移动平均预测法[1]、回归分析预测法、时间序列分析法等。这些方法大都首先假定一个模型,然后用已有的数据进行数据分析。由于假设的模型受众多因素的影响,往往一定程度上不能刻画真实的模型,甚至有些模型为了处理方便,进行了一些线性的假设,这些因素都会导致预测的效果较差。针对物流量不稳定、波动性较大的特点,不少学者[2-4]将灰色预测法与马尔柯夫模型结合起来用来预测区域物流量,从而推出未来物流量的一个区间及相应的概率;这些方法对影响物流量的因素考虑欠充分,而且结果受主观性影响过多。随着机器学习、人工智能等计算机科学的发展,近年来,人们提出用神经网络[5-7]、支持向量机[8-10]等来预测区域物流量。

BP 神经网络直接从观测数据出发,简单有效,易于实现,获得了广泛的应用。由于神经网络的设计是否成功与设计者的相关经验与足够的先验知识密切相关,因此缺乏通用性。对它的改进的研究又遇到了一些重要的困难,譬如如何确定网络结构问题、过学习与欠学习问题、陷入局部极值问题等。唐伟鸿[8]提出基于时间序列的支持向量机模型,对公路货运量与客运量进行预测。他们的预测是直接基于历年的货运量与客运量数据,对历史物流量的直接依赖过大。众所周知,一个区域的物流量与该区域的人口,经济发展,基建投入等有很大的关系,单纯地根据历年物流量去推测未来的物流量显得过于直观。庞明宝[9-10]等人考虑到物流量与一个地区的总人口、GDP、消费品零售额、工业生产总值、农业生产总值有直接的联系,因此以这些因素作为因子,用非线性支持向量回归和偏最小二乘支持向量机来预测货运量。一方面,随着近年来经济规模、经济发展轨迹的转变,固定资产投资对货运量也有直接的影响关系,不考虑这个因素,会使得预测结果有所偏差;另外一方面,货运量只说明一个区域一段时间的货物运输总量,即使知道了总量,而不考虑运输距离,物流系统规划仍然无法高效运行,因此,单纯的用货运量作为因变量进行预测分析,使得模型过于简化。

首先自适应迭代支持向量机用于对广州市物流量进行预测分析,考虑因素为广州市总人口、GDP、消费品零售额、工业总产值、农业总产值、固定资产投资额等 6 个因素。然后在此基础上,用多输出支持向量机,对广州市货运量、货物周转量、吞吐量进行预测分析。

1 自适应迭代支持向量机回归算法

1992—1995 年,在统计学习理论的 VC 维理论和结构风险最小化(SRM)准则的基础上,Vapnik 等人[11]提出了一种新的机器学习算法—支持向量机(Support Vector Machine,SVM)方法。由于其坚实的理论基础,

收稿日期:2010-06-02

作者简介:骆世广(1981—),男,讲师,硕士,研究方向为支持向量机算法的设计与分析。

良好的泛化性能，简洁的数学形式，直观的几何解释等特点，它在许多实际问题的应用中取得了成功。目前，SVM 已成功地应用于手写体识别、人脸识别、图像处理、三维物体识别、金融分析等问题。

为了提高 SVM 的求解效率，Suykens 等[12]创造性地把标准 SVM 的线性不等式约束转化成了线性等式约束，从而使得 SVM 的求解问题等价于一组线性方程组的求解。这种回归方法被称为最小二乘支持向量机回归(least squares support vector machine，LSSVM)。模型如下。

假设训练样本集为 $S=\{s_i \mid s_i=(\boldsymbol{x}_i, y_i), \boldsymbol{x}_i\in R^n, y_i\in R, i=1,2,\cdots,l\}$，基于等式约束，Suykens 和 Vandewalle 给出了下面的二次优化问题

$$\min_{w,b,\mathbf{e}} J(\boldsymbol{w}, b, \mathbf{e})=\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}+\gamma\frac{1}{2}\sum_{k=1}^{l}e_k^2 \tag{1}$$

s·t·

$$y_k=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{\varphi}(\boldsymbol{x}_k)+b+e_k,\quad k=1,\cdots,l \tag{2}$$

其相应的拉格朗日函数为

$$L(\boldsymbol{w}, b, \mathbf{e}, \boldsymbol{\alpha})=J(\boldsymbol{w}, b, \mathbf{e})-\sum_{k=1}^{l}\alpha_k\{\boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_k)+b+e_k-y_k\} \tag{3}$$

其中：α_k 是拉格朗日乘子，对应于 $\alpha_k\neq 0$ 的拉格朗日乘子称为支持向量。

经过数学处理，可以写成下列线性方程组的形式

$$\begin{bmatrix}0 & \vec{1}^{\mathrm{T}}\\ \vec{1} & \boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}+\gamma^{-1}\boldsymbol{I}\end{bmatrix}\begin{bmatrix}b\\ \boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}0\\ \boldsymbol{Y}\end{bmatrix} \tag{5}$$

式中：$\boldsymbol{Z}=[\varphi(x_1),\wedge,\boldsymbol{\varphi}(x_l)]^{\mathrm{T}}$，$\boldsymbol{Y}=[y_1,\wedge,y_l]^{\mathrm{T}}$，$\vec{1}=[1,\wedge,1]^{\mathrm{T}}$，$\boldsymbol{\alpha}=[\alpha_1,\wedge,\alpha_l]^{\mathrm{T}}$。

结合 Mercer 条件可知

$$\varphi(\boldsymbol{x}_i)\cdot\varphi(\boldsymbol{x}_j)=k(\boldsymbol{x}_i,\boldsymbol{x}_j)\equiv\Omega_{ij},\ i,j=1,2,\wedge,l \tag{6}$$

其中 $k(\cdot,\cdot)$是核函数，通常取为高斯核 $k(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/(2\sigma^2))$。

本文中，$\boldsymbol{x}_i$ 代表的是第 i 个输入样本，是一个 6 维变量，分别代表广州市总人口，GDP，消费品零售额，工业总产值、农业总产值、固定资产投资额 6 个因素。y_i 代表的是第 i 个输入样本的目标值，即广州市物流量。

LSSVM 仅仅使用增量学习，工作集中的元素数将会变得很大，这将引起训练和测试的困难，从而逆学习将是必须的。基于增量学习和逆学习策略，杨晓伟等人[13]给出自适应迭代算法(Adaptive and Iterative Support Vector Machine Regression，AISVR)。出发点是：在给定的样本集中，应该存在一个支持向量集的近似集，它覆盖了样本集的大部分信息。通过对 LSSVM 算法的重新设计，使得机器能够自动并且高效地找到这个集合，然后利用这个集合构造学习机。

区域物流量受当时经济环境、生态环境等的波动影响比较大，为了减少类似 1998 金融危机、2008 南方雪灾这种危机对物流量预测的干扰，AISVR 算法被用来做单输出情形下的预测。AISVR 不仅能够完成大样本回归问题，而且能够降低不正常数据带来的影响。

2 多输出支持向量机回归算法

多输出支持向量机回归(Multi-Output Support Vector Regression，MOSVR)[14]算法是针对模型的输出变量 y 是一个向量(即 $y\in\boldsymbol{R}^k$，$k>1$)而提出一种新的 SVM 回归算法。它主要是对单输出函数回归算法中的损失函数进行了改进，用定义在超球上的损失函数代替了定义在超立方体上的损失函数，将一般支持向量机回归模型中的损失函数

$$L(|y-f(x,\alpha)|)=\begin{cases}0, & |y-f(x,\alpha)|\leqslant\varepsilon\\ |y-f(x,\alpha)|-\varepsilon, & |y-f(x,\alpha)|>\varepsilon\end{cases} \tag{7}$$

改为

$$L(\boldsymbol{x})=\begin{cases}0 & \|\boldsymbol{x}\|<\varepsilon\\ (\|\boldsymbol{x}\|-\varepsilon)^2 & \|\boldsymbol{x}\|\geqslant\varepsilon\end{cases} \tag{8}$$

式(8)定义的损失函数优势在于它能将输入变量各分量的拟合误差综合考虑进来，使目标函数与各分量的误差都有关，从而达到整体优化的目的。另一方面这样定义的损失函数可弱化噪声数据对结果的影响，提高算法的抗噪性能。这一特点尤其适合区域物流量预测这类非线性多因素复杂系统。MOSVR 主要模型如下。

对于 M 维输入，N 维输出的函数拟合问题，假定给定的学习样本集为

$$S=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\mid i=1,2,K,L\},x\in\boldsymbol{R}^M,y\in\boldsymbol{R}^N$$

第 j 个输出的函数模型为 $G_j:f_j(\boldsymbol{x}_i,\boldsymbol{w}_j,b_j)=\boldsymbol{w}_j\cdot\boldsymbol{\varphi}(\boldsymbol{x}_i)+b_j,b_j\in\boldsymbol{R}$。可以将函数表达为 $\boldsymbol{F}(\mathrm{x})=\boldsymbol{\varphi}(\boldsymbol{x}_i)^T\boldsymbol{W}+\boldsymbol{B}$，其中，$\boldsymbol{\varphi}(\cdot)$是高维空间的非线性映射，$\boldsymbol{W}=[\boldsymbol{w}^1,\boldsymbol{w}^2,\wedge,\boldsymbol{w}^N]$，$\boldsymbol{B}=[\mathrm{b}^1,b^2,\wedge b^N]$。因此要解决多维回归问题就是要对每一个输出求出回归量 $\boldsymbol{w}^j$ 和 $b^j(j=1,2,\wedge,N)$。其目标函数如式(11)。

$$\mathrm{Min}L(\mathrm{W},\mathrm{B})=\frac{1}{2}\sum_{j=1}^{N}\|\boldsymbol{w}^j\|^2+C\sum_{i=1}^{L}L(u_i) \tag{9}$$

其中：$L(u)$是在超球上定义的损失函数，$u_i=\|\boldsymbol{e}_i\|=\sqrt{\boldsymbol{e}_i\boldsymbol{e}_i}$，$\boldsymbol{e}_i=y_i-\boldsymbol{\varphi}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{W}-\boldsymbol{B}$。当 $\varepsilon=0$ 时，该问题就是对每一个输出分量做最小二乘回归；当 $\varepsilon\neq0$ 时，在求解每一个输出函数的回归量 $\boldsymbol{w}^j$ 和 b^j 时会兼顾到其他输出分量的拟合效果，这样得到的解将会是一个整体拟合最优的解。

其中：$\boldsymbol{x}_i$ 的含义同上，y_i 是第 i 个输入样本的输出值，是一个 3 维变量，分别代表广州市货运量、货物周转量、吞吐量。

3　AISVR 与 MOSVR 在广州市物流量预测的应用

实验数据选自广州市统计年鉴[15]1985－2007 年间，共 9 个指标，分别是：总人口、GDP、消费品零售额、工业总产值、农业总产值、固定资产投资额、货运量、货物周转量、货物吞吐量。考虑到广州作为一个外来人口较多的城市，人口流动性比较大，数据中选取的总人口为年度平均人口。另外，由于各个指标的单位的不一致，数量级别差别也较大，原始数据都被进行了归一化，如表 1。

表 1　广州市 1985－2007 年间部分指标统计经归一化随机化后数据

总人口	GDP	消费品零售额	工业总产值	农业总产值	固定资产投资额	货运量	货物周转量	货物吞吐量
−0.768420	−0.825800	−0.764220	−1.045200	−0.883020	−0.798000	−0.975010	−0.759270	−0.811300
−0.373010	−0.227110	−0.268520	−0.253100	0.148100	−0.309650	−0.277110	−0.239010	0.099833
−0.708150	−0.679710	−0.727190	−0.764980	−0.519520	−0.408770	−0.662950	−0.689320	−0.654500
0.734550	0.407210	−0.135010	0.705550	0.274870	0.472230	0.253910	0.810430	0.616600
1.137100	0.984460	1.292800	1.445700	1.390500	1.397900	1.329300	1.441800	1.484900
−1.085200	−0.905690	−0.840710	−1.283300	−0.776660	−0.838130	−0.772820	−1.081000	−0.899250
−0.550070	−0.600800	−0.507690	−0.456990	−0.593310	−0.374080	0.007554	−0.431710	−0.474000
0.203560	0.099212	0.694250	0.227950	−0.169820	0.600710	0.150420	0.341880	0.151790
−1.607300	−0.786950	−0.861050	−0.781540	−1.097100	−0.910410	−0.857470	−1.486300	−0.783960
0.144750	0.855870	0.453980	0.635590	0.376700	0.847720	0.696210	0.264240	1.012100
1.436300	1.802200	1.690200	1.781300	1.572300	1.740200	1.847500	1.566800	2.268100
−0.856350	−0.779020	−1.094900	−0.910120	−0.841560	−1.373400	−0.779630	−0.846020	−0.776660
−0.645750	−0.721040	−0.755800	−0.717670	−0.871980	−0.820740	−0.578670	−0.524000	−0.687160
2.075500	2.296100	1.535000	2.050900	2.438400	1.696700	2.792500	2.634500	2.971400
−0.521230	−0.215500	−0.612900	−0.654230	−0.588500	−0.583940	−0.741270	−0.461640	−0.113880
−0.001492	−0.198540	0.278590	−0.170900	−0.139060	−0.150810	0.393170	0.160760	−0.175160
−0.835260	−1.162000	−0.772230	−0.832690	−0.769480	−1.045600	−0.886310	−0.810380	−1.070000

总人口	GDP	消费品零售额	工业总产值	农业总产值	固定资产投资额	货运量	货物周转量	货物吞吐量
0.386320	−0.065098	0.046034	−0.018277	0.595120	0.197020	−0.157000	0.498280	0.041308
0.630720	0.808040	0.488960	0.897960	0.791830	0.379990	1.125000	0.814260	0.974040
−0.682410	−0.759520	−0.520480	−0.310580	−0.646880	−0.684530	−0.640980	−0.646350	−0.764320
−0.756350	−1.017700	−0.870690	−0.798530	−0.871680	−0.745780	−0.783260	−0.742270	−0.971420
0.640500	0.998400	0.846240	0.725200	1.196300	1.095200	1.141700	0.991650	1.068700
−0.846680	−0.670610	−0.759520	−0.738090	−0.776180	−0.736830	−0.937060	−0.831140	−0.643490

对于单输出支持向量机回归模型，建立如下模型

$$y=f(x_1,x_2,x_3,x_4,x_5,x_6)$$

其中：y 代表货运量；$x_i(i=1,2,\cdots,6)$分别代表总人口、GDP、消费品零售额、工业总产值、农业总产值、固定资产投资额 6 个因素。对于多输出支持向量机回归模型，模型设置为

$$\begin{cases} y_1=f(x_1,x_2,x_3,x_4,x_5,x_6) \\ y_2=f(x_1,x_2,x_3,x_4,x_5,x_6) \\ y_3=f(x_1,x_2,x_3,x_4,x_5,x_6) \end{cases}$$

其中：$y_i(i=1,2,3)$分别代表货运量、货物周转量、货物吞吐量；$x_i(i=1,2,\cdots,6)$同上。

用 VC++ 6.0 编写了相关程序，并在内存为 512 MB、CPU 为 1.8 GHz 的 PC 机上训练并测试了上述数据。学习过程中把表 1 数据随机打散，取 18 个进行训练，5 个进行测试。

表 2 AISVR 与 MOSVR 的结果比较

输出	训练平均误差		训练正确率		测试平均误差		测试正确率	
	AISVR	MOSVR	AISVR	MOSVR	AISVR	MOSVR	AISVR	MOSVR
y_1	0.006326	0.008522	0.944444	0.944444	0.015332	0.071180	0.600000	0.600000
y_2	0.054686	0.054663	0.277778	0.777778	0.063372	0.018326	0.400000	0.600000
y_3	3.764770	0.096726	0.388889	0.722222	9.452045	0.258894	0.200000	0.800000

多输出算法训练的整体误差为 0.119 705，测试的整体误差为 0.370 778。

从表 2 可以看出，MOSVR 取得了较好的预测效果。与 AISVR 相比，有效避免了过学习，并具有更高的预测精度和抗噪能力。尤其是针对个别的指标预测上。

4 结论

通过多输出支持向量机回归模型，揭示了货运量、货物周转量、货物吞吐量与总人口、GDP、消费品零售额、工业总产值、农业总产值、固定资产投资额之间的关系。从而为准确的预测未来的物流量提供依据；如果能对总人口、GDP、消费品零售额、工业总产值、农业总产值、固定资产投资额这些量进行预测，从而可以得到物流量的一个预测值。

这样做似乎是增加了问题的不确定性，实际上，上述 9 个指标都是受很多因素影响的，在一定程度上受随机因素的影响，而这些随机因素的影响任何算法都很难考虑完全，而将它们放在一起考虑总体性质时，却会存在稳定性。未来的工作是，希望找到影响物流量的更多的因素，使用在线支持向量机进行学习预测，不断的更新历史值，以期获得更加准确的预测结果。

参考文献：

[1] 杨荣英，张辉，苗张木. 物流预测技术中的移动平均线方法[J]. 武汉理工大学学报：交通科学与工程版，2001，25(3)：353-355.

[2] 王奋奎，黄艳. 基于灰色马尔可夫模型的物流园区物流量预测研究[J]. 物流科技，2007(2)：1-4

[3] 孙卫华，王成林，经维. 邯郸国际物流园区物流量预测[J]. 物流技术，2009，28(7)：121-123

[4] 吴玉朝，蔡启明，李斌．基于灰色-马尔柯夫模型的逆向物流量预测[J]．物流科技，10(2008)：19-22.
[5] 魏连雨，庞明宝．基于神经网络的物流量预测[J]．长安大学学报：自然科学版，2004，24(6)：55-59
[6] 林连，林桦．改进的 BP 神经网络在港口物流预测中的应用[J]．交通信息与安全，2009，27(5)：161-165.
[7] 杨峰，牛惠民，邵晓彤．基于 GA-BP 算法的模糊神经网络模型在港口物流量预测中的应用[J]．物流科技，12(2009)：102-105.
[8] 唐伟鸿，李文锋．基于时间序列的支持向量机在物流预测中的应用[J]．物流科技，2005，28(3)：8-11.
[9] 庞明宝，常振华，刘娟．基于非线性支持向量机区域物流量预测[J]．物流科技，2007(9)：20-23
[10] 庞明宝，谢玲，郝然，马宁．基于偏最小二乘支持向量机回归区域物流量预测[J]．河北工业大学学报，2008，37(2)：91-96.
[11] VAPNIK V．The Nature of Statistical Learning Theory[M]．New York：Spring Verlag，1995.
[12] SUYKENS J A K，VANDWALLE J．Least squares support vector machine classifiers[J]．Neural Processing Letters，1999(9)：293-300.
[13] 杨晓伟，骆世广，余舒，等，基于支持向量机的大样本回归算法比较研究[J]．计算机工程与应用，2006，42(6)：36-38.
[14] 胡蓉．多输出支持向量回归算法[J]．华东交通大学学报，2007，24(1)：129-132.
[15] 广州统计信息网，http：//www．gzstats．gov．cn/．

A Research of Forecasing the Logistics Amount Based on Multi-output Support Vector Regression

Luo Shiguang[1]，Ye Sai[2]，Hu Rong[1]

(1．Department of Applied Mathematics，Guangdong University of Finance，Guangzhou 510521，China；2．Department of Computer Science and Technology，Guangdong University of Finance，Guangzhou 510521，China)

Abstract：The forcasting of logistics quantity is affected by many factors，The traditional method of forecasting is conducted by single-input model，which cannot achieve satisfactory forecasting results．A Multi-output Support Vector Regression (MOSVR) is used for forecasting Guangzhou's logistics amount，and Adaptive and Iterative Support Vector Machine (AISVM) is used for a contrast in this paper．The result showsthat the MOSVR is Effective．

Key words：logistics forecast；adaptive and iterative support vector machine；multi-output support vector regression

（责任编辑　王建华）