

基于支持向量机的多因子风速预测

钱斌凯¹, 何彩芬^{2*}, 金 炜¹, 倪永森³, 龚 飞¹, 左 登¹

(1. 宁波大学 信息科学与工程学院, 浙江 宁波 315211; 2. 镇海区气象局, 浙江 宁波 315202;

3. 象山县气象局, 浙江 宁波 315700)

摘要: 针对目前数值模式预报的风速普遍存在小风预报偏大、大风预报偏小的问题, 本文基于支持向量机的方法, 结合 WRF 模式预报和自动观测站的实况数据资料, 建立多因子 SVM 风速预测模型, 对渔业避风港锚地风速进行预测修正。实验结果表明, 新模型预测的风速和实际风速基本一致, 相关性达到了 99%, 很好地表达了风速与 WRF 模式预报因子之间的非线性关系, 验证了该模型能改进 WRF 模式输出的风速数据。与仅利用历史风速的非数值 SVM 和 LSSVM 风速预报的结果进行比较, 进一步证实了多因子 SVM 风速预测的优越性。

关键词: 支持向量机; 风速预测; 避风港锚地; 多因子

中图分类号: TP391

文献标志码: A

文章编号: 1001-5132 (2018) 03-0014-06

我国海岸线漫长, 其间港湾曲折, 岛屿星罗密布。广袤的海域面积带来了丰富的海洋资源和水产养殖基础, 从而促进了以水产养殖和渔业捕捞为重点支柱性产业的沿海城市的经济发展。虽然海洋经济的蓬勃发展给当地经济带来了繁荣, 但同时也带来了诸多安全问题, 尤其是台风来袭时渔船的避风问题。为保障渔民的生命财产安全, 促进海洋经济的发展, 我国建设了多个渔船专用避风锚地。但这对气象部门的要求也大大提高, 其所面临的责任重大, 不准确的风力预报轻则劳民伤财, 重则人员伤亡。因此做好渔业避风港锚地的风力精细预测, 对保障渔民的生命财产安全, 发展海洋气象事业, 保证海上安全, 促进海洋经济的发展具有重大意义。

随着近年来科学技术的蓬勃发展, 越来越多的研究人员开始关注风速预测^[1-7]。根据预测模型的输入数据可分为两类: 一类不采用数值天气预报数据, 另一类采用数值天气预报数据。采用数值天气预报数据的风速预测, 由于考虑了各种各样比如气温、气压、湿度等的气象数据, 更加贴近实

际情况, 所以用数值模式来预报风速已经成为了一种趋势。但是各种数值模式预报的精度很大程度上取决于如何选取合适的预报因子和模型参数, 此外数值天气预报数据的风速预测普遍存在小风预报偏大、大风预报偏小的问题。因此, 本文主要采用支持向量机的方法对数值预报风速来进行改进, 利用 WRF 数值预报和自动观测站的实况数据进行风速预测, 并与仅利用历史风速的 SVM 方法和 LSSVM 方法预报的风速进行比较, 结果显示前者具有较大优越性。

1 资料和方法

在 SLT(Statistical Learning Theory)的基础上^[8], Vapnik 等发展了 SVM 算法, 是一种监督式学习的方法。主要的算法思想是利用核函数将输入样本映射到更高维度的特征空间中, 在其中建立最大间隔超平面, 将数据分隔成 2 个相互平行的超平面, 以使 2 个平行超平面的距离最大化, 最终得到输入变量和输出变量之间的非线性关系。SVM 在解决小样本、非线性及高纬度模式识别问题中都有着诸

收稿日期: 2017-06-30.

宁波大学学报(理工版)网址: <http://journalg.nbu.edu.cn/>

基金项目: 国家自然科学基金(61471212); 浙江省自然科学基金(LY16F010001); 宁波市自然科学基金(2016A610091, 2017A610297); 浙江省气象局科研项目(2016YB01); 宁波象山县社会发展科技项目(2016C6012)。

第一作者: 钱斌凯(1993-), 男, 浙江金华人, 在读硕士研究生, 主要研究方向: 气象业务。E-mail: qianbinkai2@163.com

*通信作者: 何彩芬(1974-), 女, 浙江宁海人, 高级工程师, 主要研究方向: 短临、短期天气预报。E-mail: 468176571@qq.com

多优异的表现, 并且能够推广到其他机器学习问题中.

1.1 支持向量机分类

支持向量机分为线性可分、非线性可分和核函数映射 3 种情况. 对于线性可分的情况, 已知样本数据集为 $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, 样本特征向量 $x \in R^D$, 类标签 $y \in \{-1, 1\}$. 要对此两类样本进行分类, 目标就是寻找最优分类超平面.

最优化问题的目标函数如下:

$$\min_{w, b, \varepsilon} \left(\frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \varepsilon_i \right), \quad (1)$$

$$\text{s.t. } \alpha_i \geq 0, \sum_{i=1}^N \alpha_i y_i = 0, i = 1, 2, \dots, N,$$

式中, C 为最大间隔和最小分类误差的惩罚系数; ε_i 为松弛变量.

该优化问题的对偶形式如下:

$$\max_{\alpha} \left(\sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \right), \quad (2)$$

$$\text{s.t. } 0 \leq \alpha_i \leq C, \sum_{i=1}^N \alpha_i y_i = 0, i = 1, 2, \dots, N.$$

为避免复杂的高维计算, 由此引入核函数, 目标函数则转变为:

$$\max_{\alpha} \left(\sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j) \right), \quad (3)$$

最优分类函数为:

$$h(x) = \text{sign}(\alpha_i y_i K(x_i \cdot x_j) + b), \quad (4)$$

其中, $K(x_i \cdot x_j)$ 为核函数, 而常用的核函数有线性核函数、多项式核函数、径向基核函数和 Sigmoid 核函数等.

1.2 支持向量机回归

Vapnik 等在原有 SVM 的基础上定义了 ε 不敏感函数, 提出了 ε 支持向量回归. 对于支持向量回归^[9], 其就是利用非线性映射将样本数据映射到一个高维特征空间, 然后在此高维特征空间上进行线性回归分析.

已知 N 个独立同分布样本数据集为 $\{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\} \subseteq R^D \times R$, ε 支持向量回归算法在定义了 ε 不敏感损失函数的基础上, 寻找目标函数:

$$f(x) = \langle w, \varphi(x) \rangle + b, \quad (5)$$

式中, $w, x \in R^D, b \in R$, 使得预测期望风险 $R(w)$ 最

小:

$$R(w) = \frac{1}{2} \|w\|^2 + C \cdot R_{\text{emp}}^{\varepsilon}(w), \quad (6)$$

式中, ε 为不敏感损失函数; $\|w\|^2$ 为结构风险, 代表回归模型的复杂程度; $R_{\text{emp}}^{\varepsilon}$ 为经验风险, 代表回归模型的误差情况; C 为惩罚因子, 用于平衡结构误差和经验误差, 控制回归模型的鲁棒性.

最小化(6)式等价于寻找以下的目标函数:

$$\min_{w, \xi, \xi^*} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N (\xi_i + \xi_i^*),$$

$$\text{s.t. } f(x_i) - y_i \leq \xi_i^* + \varepsilon, i = 1, 2, \dots, N, \quad (7)$$

$$y_i - f(x_i) \leq \xi_i + \varepsilon, i = 1, 2, \dots, N,$$

$$\xi_i, \xi_i^* \geq 0, i = 1, 2, \dots, N,$$

式中, ξ 和 ξ^* 为松弛变量的上下限.

引入 Lagrange 系数 a_i 和 a_i^* 后, 将目标函数转化为其对偶形式:

$$\max_{a, a^*} \left(-\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (a_i - a_i^*)(a_j - a_j^*) \langle \varphi(x_i), \varphi(x_j) \rangle + \sum_{i=1}^N (a_i - a_i^*) y_i - \sum_{i=1}^N (a_i + a_i^*) \varepsilon \right), \quad (8)$$

$$\text{s.t. } \sum_{i=1}^N (a_i - a_i^*) = 0, 0 \leq a_j, a_j^* \leq C,$$

$$i = 1, 2, \dots, N.$$

w 表达式如下:

$$w = \sum_{i=1}^N (a_i - a_i^*) \varphi(x_i), \quad (9)$$

得出回归目标函数为:

$$f(x) = \sum_{i=1}^N (a_i - a_i^*) \langle \varphi(x_i), \varphi(x_j) \rangle + b. \quad (10)$$

为避免复杂的高维计算, 引入核函数. 因此最终的回归目标函数形式为:

$$f(x) = \sum_{i=1}^N (a_i - a_i^*) K(x, x_i^*) + b. \quad (11)$$

1.3 基于 SVM 的风速预测模型构建

本文构建基于 SVM 的风速预测模型用于预测某渔业避风港锚地的风速, 风速预测模型的工作流程如图 1 所示. 风速预测具体步骤如下:

步骤 1 主要特征因子选取. 通过对 WRF 模式预报数据和自动观测站数据进行相关性分析, 并结合相关风速气象资料确定主要特征参数因子.

步骤 2 样本数据提取. 从自动站观测数据和 WRF 模式预报数据中提取样本, 将样本分为训练

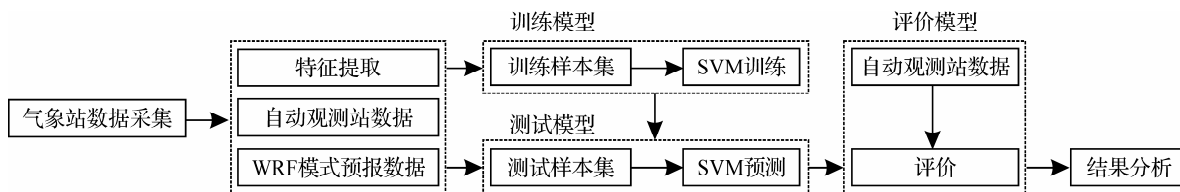


图1 锚地风速预测流程

集和测试集。其中,训练集由特征分量和类别标签两部分组成,WRF模式预报数据参数特征作为模型的特征分量,类别标签是与特征分量一一对应的风速真值。

步骤3 训练SVM。用训练集样本训练SVM回归预测模型。

步骤4 预测风速。利用训练好的回归预测模型对测试集样本进行风速预测。

步骤5 评价模型。将风速预测模型预测的风速值与风速实际值进行比较,评价模型性能和参数因子选择的合理性。

2 结果与分析

2.1 样本数据预处理和预测模型的参数确定

本文选取的样本数据来源为2015年7月份和8月份(台风活跃时期)某渔业避风港锚地的WRF模式预报数据和自动站观测数据(采样间隔为1h)。选取该锚地7月1日~8月4日的840个样本作为测试集;按同种方法选取8月5日~8月7日的40个样本作为非台风时期测试集,8月9日的24个样本作为台风时期测试集。

通过对WRF模式预报数据和自动站观测数据进行相关性分析,按相关系数从WRF模式预报数据中选取的特征参数因子为1h降水、温度、风向、相对湿度和露点温度,其中1h降水、温度呈负相关,风向、相对湿度和露点温度呈正相关;自动站观测数据中选取的是实际最大风速。最大风速是指在某个时间段内出现的最大10min平均风速。为了平衡参数特征作用、减小误差,需要对数据进行预处理归一化到[0,1]。本文选用径向基核函数(RBF)作为建立风速预测模型的核函数。为了获取最优的参数 C 、 ε 和RBF核函数参数 γ ,需要固定其中2个参数,调试另1个参数,直到误差最小。

2.2 风速预测模型评价标准

本文对预测结果的评价指标为平均百分比误

差(M_{MAPE})和均方根误差(R_{RMSE}),计算公式如下:

$$R_{RE} = \frac{v - \tilde{v}}{v} \times 100\%, \quad (12)$$

$$M_{MAPE} = \frac{1}{N} \sum_{i=1}^N \frac{|v - \tilde{v}|}{|v|} \times 100\%, \quad (13)$$

$$R_{RMSE} = \sqrt{\frac{\sum_{i=1}^N (R_{RE} - M_{MAPE})^2}{N}} \times 100\%, \quad (14)$$

式中, v 为实际真实风速; $\tilde{v}$ 为模型预测风速。

2.3 风速预测模型结果分析和评价

如图2所示,对WRF模式预报数据和实际真实风速进行分析对照,可以发现WRF模式预报结果和实际风场存在着一定偏差,总体偏小,同时也存在着小风天气预报偏大、大风预报偏小的问题。

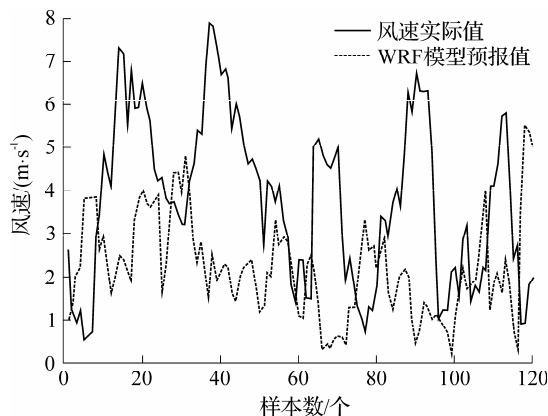


图2 WRF模式预报和真实风速的比较

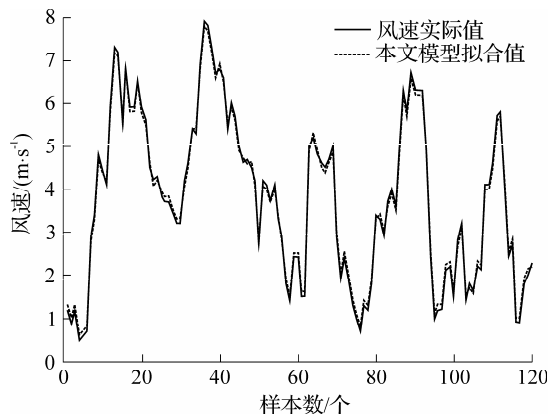


图3 模型拟合结果

本文首先对训练集 840 个样本进行训练,训练结果如图 3 所示. 然后对测试集样本进行非台风期间的 0~24 h 逐小时预报与 24~72 h 每 3 h 预报和台风期间的 0~24 h 逐小时预报,具体结果如图 4~图 6 所示.

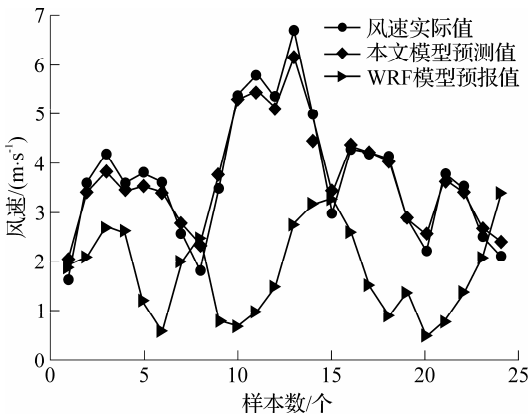


图 4 非台风期间的 24 h 模型预测和真实值

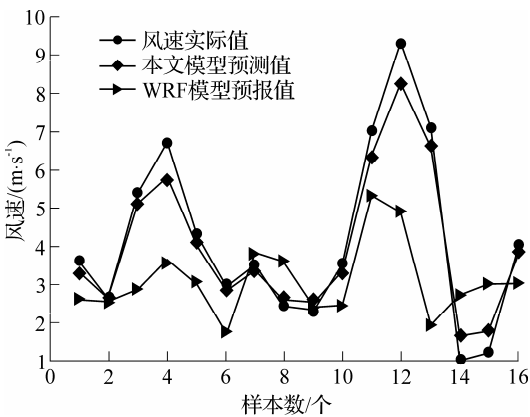


图 5 非台风期间的 24~72 h 模型预测和真实值

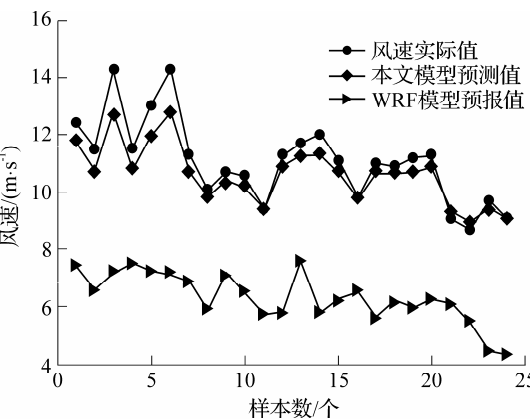


图 6 台风期间的 24 h 模型预测和真实值

从图 3 可以看出,本模型的拟合效果非常好,预测曲线与实际风速曲线在总体上保持一致,说明所选预报特征参数因子能很好地刻画风速.从

图 4、图 5 和图 6 可以看出,与 WRF 模式预报的结果相比,预测结果和真实风速非常接近,只是在风速较大时出现较大的差值,说明本文模型具有较高的精度.

将非台风期间 24 h 内逐小时与 24~72 h 每 3 h 的和台风期间 24 h 内逐小时的预测结果和真实风速进行比较,具体结果见表 1~表 3.

表 1 24 h 风速预测结果(非台风期间)

t/h	实际风速/(m·s ⁻¹)	预测风速/(m·s ⁻¹)	相对误差/%
1	3.6	3.399	5.58
2	4.2	3.836	8.67
3	3.6	3.446	4.28
4	3.8	3.547	6.66
5	3.6	3.422	4.94
6	2.6	2.774	-6.69
7	1.8	2.386	-32.56
8	3.5	3.786	-8.17
9	5.4	5.289	2.06
10	5.8	5.473	5.64
11	5.3	5.120	3.40
12	6.7	6.143	8.31
13	5.0	4.888	2.24
14	3.0	3.405	-13.50
15	4.3	4.368	-1.58
16	4.2	4.247	-1.12
17	4.1	4.046	1.32
18	2.9	2.884	0.55
19	2.2	2.574	-17.00
20	3.8	3.672	3.37
21	3.5	3.422	2.23
22	2.5	2.690	-7.60
23	2.1	2.419	-15.19
24	2.4	2.646	-10.25

注:平均百分比误差 $M_{MAPE}=7.21\%$,均方根误差 $R_{RMSE}=6.86\%$.

表 2 24~72 h 中每 3 h 风速预测结果(非台风期间)

t/h	实际风速/(m·s ⁻¹)	预测风速/(m·s ⁻¹)	相对误差/%
1	3.6	3.304	8.22
2	2.6	2.634	1.31
3	5.4	5.115	5.28
4	6.7	5.805	13.36
5	4.3	4.068	5.40
6	2.9	2.791	3.76

续表

t/h	实际风速/($m \cdot s^{-1}$)	预测风速/($m \cdot s^{-1}$)	相对误差/%
7	3.5	3.349	4.31
8	2.4	2.567	-6.96
9	2.3	2.567	-11.61
10	3.5	3.327	4.94
11	7.0	6.339	9.44
12	9.3	8.221	11.6
13	7.1	6.607	6.94
14	1.0	1.630	-63.00
15	1.2	1.754	46.17
16	4.0	3.813	4.68

注: 平均百分比误差 $M_{MAPE}=12.94\%$, 均方根误差 $R_{RMSE}=16.4\%$.

表3 24 h 风速预测结果(台风期间)

t/h	实际风速/($m \cdot s^{-1}$)	预测风速/($m \cdot s^{-1}$)	相对误差/%
1	12.4	11.804	4.81
2	11.5	10.713	6.84
3	14.3	12.728	11.00
4	11.5	10.780	6.26
5	13.0	11.954	8.05
6	14.3	12.833	10.3
7	11.3	10.720	5.13
8	10.1	9.820	2.77
9	10.7	10.355	3.22
10	10.6	10.208	3.70
11	9.4	9.361	0.41
12	11.3	10.890	3.63
13	11.7	11.233	3.99
14	12.0	11.329	5.59
15	11.1	10.693	3.67
16	9.8	9.696	1.06
17	11.0	10.710	2.64
18	10.9	10.652	2.28
19	11.2	10.719	4.29
20	11.3	10.876	3.75
21	9.1	9.302	-2.20
22	8.7	8.954	-2.90
23	9.7	9.478	2.29
24	9.1	9.109	-0.10

注: 平均百分比误差 $M_{MAPE}=4.21\%$, 均方根误差 $R_{RMSE}=2.68\%$.

从表1和表2可以看出, 对于非台风期间0~24 h 内的逐小时预测, 模型预测结果和真实风速的误差小, 平均百分比误差 M_{MAPE} 为 7.21%, 均方根误

差 R_{RMSE} 为 6.86%. 而对于 24~72 h 内每 3 h 预测, 模型预测结果和真实风速的误差增大, 平均百分比误差 M_{MAPE} 为 12.94%, 均方根误差 R_{RMSE} 为 16.4%, 说明对于长时间的预测, 精度还不够理想. 其次, 相比较于只利用历史风速的非数值 SVM 和 LSSVM 风速预报^[10-11], 平均百分比误差 M_{MAPE} 从 10.07%和 8.22%减小到了 7.21%, 说明多因子数值天气预报有一定优势, 性能更稳定. 另外, 从表3可以看出, 在台风期间0~24 h 内的逐小时预测, 模型预测结果和真实风速的误差较小, 平均百分比误差 M_{MAPE} 为 4.21%, 均方根误差 R_{RMSE} 为 2.68%, 说明本模型在台风风速的预测上效果突出, 对于台风期间风速的实时预报具有重要实践意义. 此外, 也可以看出模型还是存在较小程度的小风天气预报偏大、大风预报偏小的问题, 表明本文模型对风速预测还有提升空间.

总体来说, 本文模型预测风速和实际风速之间的相关性达到了 99%, 表明本文模型风速预测特征参数因子的选择是合理的, 同时也说明了支持向量机可以较好地刻画风速的非线性特征.

3 结论

在多因子特征中引入支持向量机后, 较好地解决了 WRF 模式预报中风速的小风预报偏大、大风预报偏小的问题. 由模型预测的风速与实际真实风速非常接近, 相关性达到了 99%, 预测效果比较理想, 说明支持向量机可以改进数值预报风速. 与仅利用历史风速的非数值 SVM 和 LSSVM 风速预报的结果进行比较, 证实了多因子 SVM 风速预测的优越性, 为进一步提高数值天气预报的精度提供了理论基础. 台风期间的风速预测结果说明本模型在台风风速的预测上效果突出, 对于台风期间风速的实时预报具有重要的实践意义. 但本文模型没有彻底解决上述问题, 有待进一步研究, 以提高预测的准确性. 另外本文模型支持向量机参数寻优上还有待加强, 对于样本噪声的处理还有待优化.

参考文献:

- [1] Mohandes M A, Halawani T O, Rehman S, et al. Support vector machines for wind speed prediction[J]. Renewable Energy, 2004, 29(6):939-947.

- [2] Lei M, Luan S, Jiang C, et al. A review on the forecasting of wind speed and generated power[J]. *Renewable & Sustainable Energy Reviews*, 2009, 13(4):915-920.
- [3] Santamaria-Bonfil G, Reyes-Ballesteros A, Gershenson C. Wind speed forecasting for wind farms: A method based on support vector regression[J]. *Renewable Energy*, 2016, 85:790-809.
- [4] Wang S X, Zhang N, Wu L, et al. Wind speed forecasting based on the hybrid ensemble empirical mode decomposition and GA-BP neural network method[J]. *Renewable Energy*, 2016, 94:629-636.
- [5] Meng A B, Ge J F, Hao Y, et al. Wind speed forecasting approach using wavelet packet decomposition, artificial neural networks trained by crisscross optimization algorithm[J]. *Energy Conversion and Management*, 2016, 114(2):75-88.
- [6] Wang C, Wu J, Wang J, et al. Short-Term wind speed forecasting using the data processing approach and the support vector machine model optimized by the improved Cuckoo search parameter estimation algorithm[J]. *Mathematical Problems in Engineering*, 2016(2):1-17.
- [7] Chen J, Xue X, Ha M, et al. Support vector regression method for wind speed prediction incorporating probability prior knowledge[J]. *Mathematical Problems in Engineering*, 2017(4):1-10.
- [8] Adankon M M, Cheriet M. Support vector machine[J]. *Computer Science*, 2002, 1(4):1-28.
- [9] Awad M, Khanna R. Support vector regression[J]. *Neural Information Processing Letters & Reviews*, 2007, 11(10): 203-224.
- [10] 张华, 曾杰. 基于支持向量机的风速预测模型研究[J]. *太阳能学报*, 2010, 31(7):928-932.
- [11] 刘忠宝. 新型支持向量机在风速预测模型中的应用研究[J]. *电子科技大学学报*, 2014(5):754-757.

Multi-factors wind speed forecasting based on support vector machine

QIAN Bin-kai¹, HE Cai-fen^{2*}, JIN Wei¹, NI Yong-sen³, GONG Fei¹, ZUO Deng¹

(1.Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China;

2.Zhenhai Observatory, Ningbo 315202, China; 3.Xiangshan Observatory, Ningbo 315700, China)

Abstract: The problem of deviation has long been identified in the wind speed forecasting with numerical model when wind speed turns too high or low. To tackle this problem, multi-factors SVM model to predict and update the wind speed of fishery harbor anchorage is established, which combines the data of WRF model forecast and those from automatic observation stations. The experimental results suggest that the forecasting wind speed and the actual one is basically in agreement with each other, and the correlation is found to reach up to 99%. The presented model also acts as a good expression for the nonlinear relationship between wind speed and WRF model forecasting factors, which verifies the model's ability for updating the WRF wind speed forecasting. By comparing forecast results obtained in this work with other forecast models such as SVN and LSSVM only using historical wind speed, it proves the advantage of the proposed model.

Key words: support vector machine (SVM); wind speed forecasting; harbor anchorage; multi-factors

(责任编辑 章践立)