SVM方法在霾识别和能见度预报中的应用

郑朝霞　周梅　季致建　方桃妮　刘学华
（金华市气象局，金华 321000）

摘要：选用2013—2014年地面自动站资料、探空气象资料以及大气污染物浓度的数据，采用支持向量机（Support Vector Machine，SVM）方法分别建立金华SVM霾识别预报模型和14时能见度SVM回归预报模型来进行实证研究。通过预报结果检验发现：1）金华地区SVM霾识别预报模型的TS评分均在0.65以上，且8个最优模型判断完全错误的天数只有3d，占2.7%，表明模型分类结果较好，可在实际业务预报中推广应用；2）金华地区14时能见度SVM回归预报模型得到的预报值集中在6～16km，预报值较为集中，而实况值波动较大，即模型对极值预报能力较弱，表明模型对中度霾和重度霾天气预报的指导意义不大。

关键字：SVM方法，霾识别预报，14时能见度预报

DOI：10.3969/j.issn.2095-1973.2016.06.005

Application of SVM Method to Identification of Haze and Prediction of Visibility

Zheng Zhaoxia, Zhou Mei, Ji Zhijian, Fang Taoni, Liu Xuehua
(Jinhua Meteorological Bureau, Jinhua 321000)

Abstract: Based on the data in 2013-2014 at automatic weather station, radiosonde and the concentrations of air pollutants, the identification models of haze and the prediction models of visibility at 1400 BT were respectively carried out by using the Support Vector Machine (SVM) at Jinhua Meteorological Bureau. The results show that: 1) The identification models of haze may be used in the actual business forecast, because the satisfied TS scores were all over 0.65, except for that only three days were judged completely wrong by eight optimal models; 2) The prediction of the visibility forecast models for haze-day at 1400 BT in Jinhua were concentrated in 6-16km, which is much small than the actural range, however. That means the models have little cability to forecast the extreme values, so have a little guidance to distinguish the moderate and severe haze.

Keywords: SVM method, the identification of haze, visibility forecast at 1400 (BT)

0　引言

近年来，霾是民众关注的热点问题之一且其在我国的出现频率越来越高，特别是长三角和京津冀等工业化发达的城市尤为突出。金华地处金衢盆地东段，为浙中丘陵盆地地区，地形南北高、中部低，呈马鞍形，位于长三角经济发达地区的下游，污染源较多，加上特殊的地形有利于由北向南移动的污染物堆积滞留，有利于霾的形成。金华的气候特点为除连阴雨时期、梅雨汛期和台风汛期等三个时段降水较为丰沛外，其他时间（尤其是每年10月至次年2月）降水常年较少，霾的影响较为严重。霾与人民的身体健康息息相关，影响较为严重，故霾的形成机理、变化和预测预报方法是目前气象和环境部门研究的重点。在业务中，霾常用的预报方法主要有经验预报法、统计预报法和数值预报法，其中经验预报和统计预报法需要结合有利于霾形成的气象条件和变化规律等，但考虑到影响霾形成的因素较多、影响因子复杂，且其与雾形成的气象条件相类似等特点，目前尚未得到较好的研究结果。而数值预报方法则是先要了解当地污染物浓度和变化规律，再计算能见度。但由于影响污染物浓度变化的因素较多，变化规律较难掌握，加上计算量较大，这项方法在实际业务应用中开展较为缓慢。本文采用一种对预报因子与预报对象是否线性相关无明显的依赖关系，且对因子的数量没有明显的限制的[1-2]，并基于统计学习理论发展而来的方法，即支持向量机（Support Vector Machine，SVM）方法对霾的识别和能见度预报进行了较为系统的研究。

统计学习理论是一种专门解决有限样本学习问题的理论，SVM方法就是在这一理论基础上发展演化而来的。陈永义等[1]和冯汉中等[2]指出，SVM方法的

收稿日期：2015年6月23日；修回日期：2016年3月14日
第一作者：郑朝霞（1987—），Email:zzx19870310@163.com
资助信息：金华市气象局青年项目（2014QN01）

最终决策函数只由少数的支持向量所决定，计算的复杂性大大降低，而且根本不需要知道自变量和因变量之间的显式表达式，对研究预报对象与预报因子间关系不明确的情况十分有利。目前，SVM方法在气象预测预报领域，如暴雨预报[3]、温度预报[4-5]、大雾预报[6]、天空云量预报[7-8]等方面已经取得了一些进展。本文以金华国家基准气候站（简称金华站，下同）为例，将SVM分类和回归方法应用到霾识别和14时（北京时，下同）能见度预报中。

1 SVM霾识别预报模型

1.1 确定预报对象

随着气象观测业务现代化的发展，自2014年1月1日起金华站能见度观测改为自动观测，同时，国家气象中心对霾判识标准进行了相应修订，即当降水量为0mm、风速＜4m/s、能见度＜7500m、相对湿度＜80%时，判识为霾，且规定霾日的确定应以台站的自动观测记录为准。故本文霾日的确定均以金华站地面观测上传的长Z文件的记录为准，若天气现象栏中出现霾记录则确定当日为霾日，记为“1”；否则记为“－1”，即无霾日。

1.2 构建预报因子

选取2013年1月至2014年12月地面、探空和污染物浓度资料，共71个预报因子构建模型，具体包括：1）金华站逐日08、14和20时的温度、气压、相对湿度、露点和风速5类地面资料；2）衢州站①的1000、925、850、700和500hPa各层的位势高度（其中1000 hPa位势高度由于缺测太多故剔除），温度、露点和风速等探空资料；3）逐日08、14和20时的SO_2、CO、O_3、NO_2、$PM_{2.5}$和PM_{10}6类大气污染物浓度②。

1.3 选取建模方式

CMSVM2.0系统中分类问题的模型择优标准有三种③，本文选用适合正样本发生频率较小且相对较严格的标准——正样本TS评分。用此标准分别选取8种核函数的最优模型，并进行试验对比。

1.4 建立识别预报模型

剔除缺测样本后，模型的有效样本均为726个，由于样本长度有限，且试验样本的数据不参与建模过程，故可用试验样本代替检验样本。将样本资料按时间顺序分为两个部分：1）训练样本，约占85%，共617个；2）试验和检验样本，约占15%，共109个。

1.5 模型结果和分析

采用逐步筛选方法确定最优模型参数，得到SVM霾识别预报最优模型，结果如下：

通过表1的分析得到：1）8类核函数最优模型的正样本分类TS评分均在0.65以上，最大为0.68，分类正确率均大于73.39%，最大为77.06%，分类结果较为满意；2）除线性核函数外，其余核函数空报次数少于漏报，但8类核函数最优模型的预报结果总体相近，故需要对最优模型分类情况和错误的样本进行逐个比对和交叉分析。

表1 SVM霾识别最优模型和分类结果

Table 1 The optimal models and test results of the identification models by using the SVM method

核函数	c	g/u/d	漏(空)报	正确率/%	成功率/%	TS评分	正样本的TS评分
线性	160	—	12(16)	74.31	78.67	0.68	0.43
多项式	50	2	18(9)	75.23	85.48	0.66	0.55
径向基	50	1.7	18(11)	73.39	82.81	0.65	1.0
对称三角形	50	0.3	18(7)	77.06	88.33	0.68	1.0
柯西	50	1.3	16(11)	75.23	83.33	0.67	1.0
拉普拉斯	50	0.35	18(7)	77.06	88.33	0.68	1.0
双曲正割	990	0.2	17(8)	77.06	87.10	0.68	0.85
平方正弦	900	0.25	19(8)	75.23	86.67	0.66	0.82

通过对8类核函数最优模型的分类结果逐个比对、交叉分析以及对分类错误样本的分析（图1）知：1）检验样本共有109个，其中有44个样本分类完全正确，38个样本有大于等于5个最优模型分类正确，24个样本有小于等于4个最优模型分类正确，只有3个样本分类完全错误；2）65个分类错误样本中，除3个完全分类错误的样本外，最优模型对其余样本的判断错误结果只有一种，未出现既有空报又有漏报的样本，减少了分歧，增强了实际业务中预报员通过结合当日实况和气象条件的主观分析进行正确预报的把握。

通过对8个最优模型都错误归类的3个样本（2014年10月13日、11月11和27日）的分析可知：三天均为霾日和无霾日转折天气，主客观预报难度都相应增加。其中，10月12日有霾，属轻度污染；随着冷空气南下，13日金华处地面锋区，风力增大，有利于污染物的扩散，转为无霾；14日受地面冷高控制，层结稳定，且没有输入性污染物，仍无霾。11月10日冷空气南下补充，输入性污染物增多，但8—10日金华

① 金华站不是探空站，故选用距离较近的探空站——衢州站（58633）资料代替。

② 金华站未设环境质量监测点，环保部门在金华市本级共设4个监测站点，其中焦岩背景站2013年8月开始投入使用，故文中大气污染物浓度数据用金华监测站（29.10°N，119.68°E）、十五中（29.08°N，119，65°E）和四中（29.11°N，119.65°E）三个监测站点的平均值代替。

③ 中国气象局气象干部培训学院《SVM2.0用户使用手册》和陈永义，冯汉中，王泳等的《SVM讲义》。

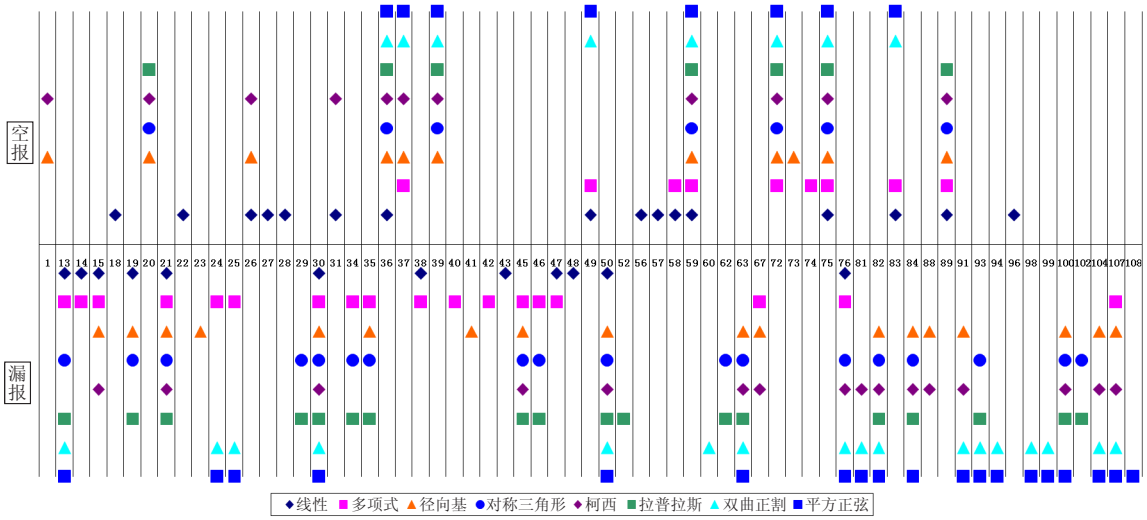

图1 8个最优模型中分类错误样本分布

Fig.1 The error classification samples of the eight optimal models

站及周边地区都有弱降水，无霾；11日冷空气继续渗透，弱降水发生前或发生时对颗粒物的洗涤作用较好[9-10]，能见度短暂增大，无霾；12日冷空气主体南下，输入性污染物持续的增加，转为霾日；26日夜里高空有下滑槽东移影响，27日低层有弱切变东移，金华站周边有弱降水，无霾；28日受弱冷空气影响，输入性污染物增加，转为有霾。综上可以看出，前期弱降水对空气中颗粒物或污染物有一定的洗涤作用，而强降水又可以使能见度显著减低[9-10]，说明降水天气现象可以影响能见度的变化，进而影响霾的判识，增加了霾天气识别的难度，尤其是有霾和无霾转折性天气预报的难度。

从这三个例子也可以看出，由于金华特殊的地形和地理位置影响，冷空气对金华站霾的形成是一把双刃剑：一方面受冷空气影响，风力将会增大，将有利于污染物的扩散，有利于霾的消散，如2014年10月13日；另一方面由于冷空气路径一般为自北向南，因而冷空气会携带上游地区的污染物南下，受盆地效应的影响，有利于输入性污染物在金华的堆积，加重污染，如2014年11月27日。

2 14时能见度SVM回归预报模型

业务中，霾预报的重点和最终目的是为了做霾的强度预报，依据霾的强度发布相应级别的预警信号等，而根据国家气候中心对霾等级的划分标准可以看出，大气能见度是其判断的主要依据。其中，当能见度大于2000m且小于3000m定义为中度霾，当能见度小于2000m时定义为重度霾。下文中将利用SVM回归预报方法建立14时能见度预报模型。

2.1 确定预报对象和预报因子

预报对象为2013年1月1日—2014年12月31日14时能见度，预报因子同SVM霾识别预报模型。

2.2 选取建模方式

SVM回归模型的择优标准选用均方根，分别得到8种核函数的最优模型，并进行对比。

2.3 建立回归预报模型

剔除缺测样本后，模型的有效样本均为726个，样本分类同SVM霾识别预报模型。

2.4 模型结果和分析

采用逐步筛选的方法确定最优模型参数，得到SVM回归预报最优模型，结果如表2和图2所示。

表2 能见度预报最优模型和预报结果

Table 2 The optimal models and forecast results of the prediction models of visibility

核函数	c	g/u/d	w	绝对差	均方差	准确率（≤3km）	准确率（≤2km）
线性	10	–	0.5	3.99	5.00	43%	35%
多项式	10	1	0.5	3.99	5 .00	43%	35%
径向基	10	0.05	1.5	3.84	4.66	45%	33%
对称三角形	10	0.05	0.1	3.97	4.84	44%	33%
柯西	10	0.05	1.5	3.84	4.66	45%	33%
拉普拉斯	10	0.1	0.05	3.97	4.84	44%	33%
双曲正割	10	0.5	0.45	3.81	4.67	46%	31%
平方正弦	10	0.9	0.45	3.81	4.65	45%	35%

通过分析表2可知：8个核函数的最优模型对14时能见度预报值和预报效果较为接近，其中绝对差

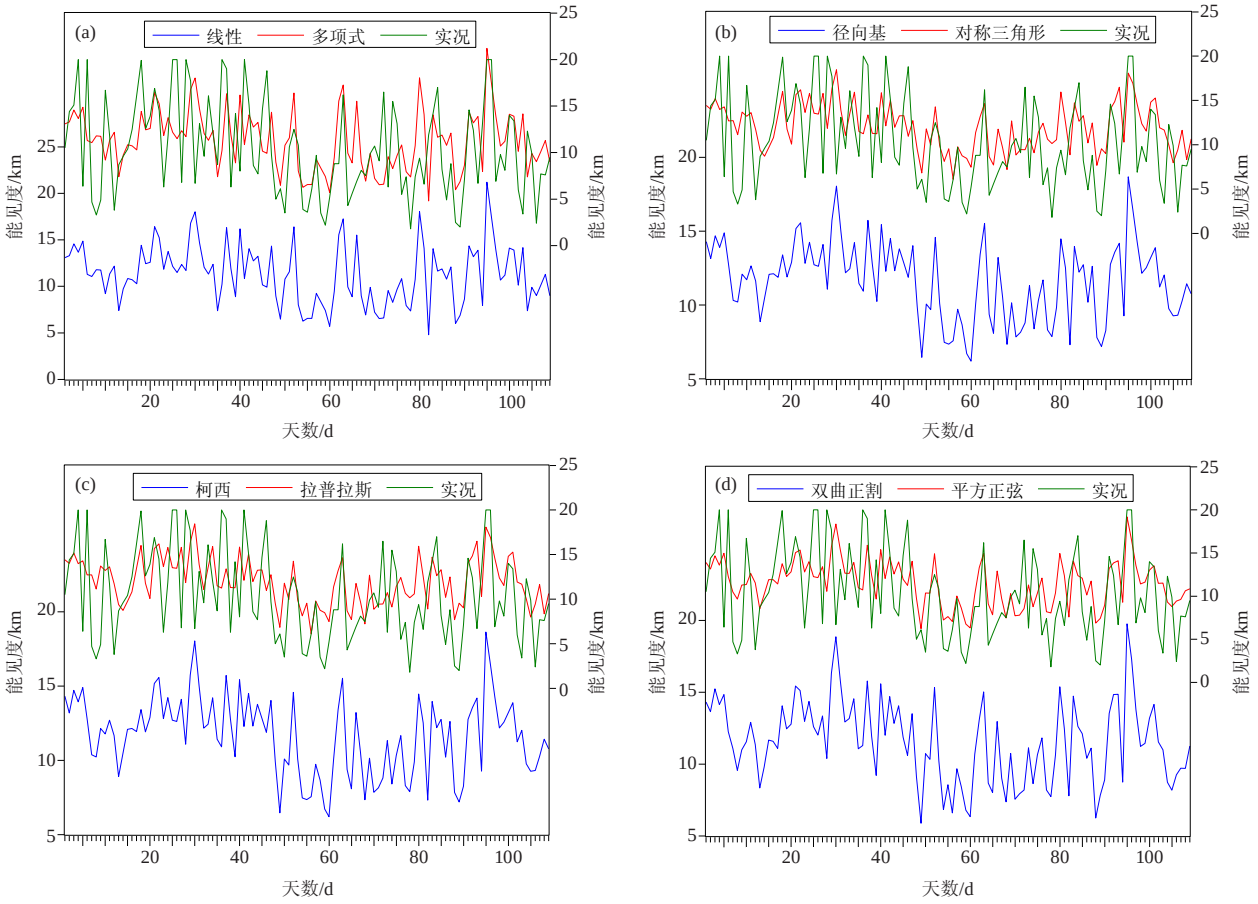

图2 14时能见度实况值和8个最优模型14时能见度预报值分布
(a)线性函数、多项式函数预报值；（b）径向基、对称三角形预报值；（c）柯西函数、拉普拉斯函数预报值；（d）双曲正割函数、平方正弦函数预报值

Fig.2 The visibility at 1400 of the real and the forecast of the eight optimal models
(a) The forecast of linear and polynomial; (b) The forecast of the radial basis function and symmetrical triangle function; (c) The forecast of the Cauchy function and Laplace function; (d) The forecast of Hyperbolic secant function and square sine function

为3.81～3.99，均方差为4.65～5.00，误差小于2km的预报准确率为31%～35%，小于3km的准确率为43%～46%。

通过预报值和真实值对比（图2）可知：14时能见度实际值波动较大，而8个最优模型的预报值均集中在6～16km，且波动较为平缓，说明SVM回归预报模型对14时能见度极值预报效果较差，尤其是对能见度小于6km的样本预报准确率几乎为0，即对重度霾和中度霾天气的指导意义不大。

应用SVM回归方法建立14时能见度预报模型的预报结果不是很理想，究其原因分为客观和主观两方面。从客观上看，一方面由于探空站设置的局限性，用衢州站探空信息代替，大大增加了模型的误差；另一方面由于只有近两年的大气污染物浓度数据资料，样本长度较短，使得模型构建时样本包含的信息不够全面，并且检验样本以冬季为主，而训练集中冬季样本较少，这就会引起建立的模型不足以预报检验样本。主观方面的原因在于影响能见度的因素众多且十分复杂，众多研究表明气温、空气湿度、地面风速和24h内变压等气象因素以及太阳辐射、紫外辐射和大气成分等环境因素都与能见度有很好的相关性。而且晴天、阴天或雨天等不同类型天气下能见度与污染物之间的相互作用也不同。从长期角度出发来看，能见度变化趋势与当地人为排放污染物浓度密切相关，化工厂等污染物排放较为严重企业的建立、拆除，以及污染物处理设备的更新等都将影响当地能见度的变化。然而，本文在建立模型时，仅考虑了气压、相对湿度、风速、探空和6类大气污染物，对当日太阳辐射、紫外辐射和天气类型未分析，尤其是未考虑雨水对空气中颗粒物的沉降冲刷作用，所选因子过于局

限，不够全面。另一部分的误差可能是由于引入了影响能见度相关性不大的因素引起的，如引入了700和500hPa两层的探空资料，资料距离地面较高，而污染物平流扩散的主要层次在200～500m[11]。故在实际业务中，需要在不断增加样本长度、样本的多样性和综合性的同时，引入更多相关性较好的或删除相关性不大的预报因子，不断完善模型。

3 结论和讨论

主要结论如下：

（1）利用SVM分类方法建立的霾识别模型正样本分类TS评分均在0.65以上，最大为0.68，且109个样本中，3个样本的预报分类完全错误，占2.7%，分类预报结果基本达到业务应用水平。

（2）应用SVM回归方法建立的霾日14时能见度预报模型中，8类最优模型的预报效果和预报值较为接近，误差小于3km的准确率仅为43%～46%，尤其是对极值预报能力较弱，对霾强度的预报指导意义不大。

（3）冷空气对金华霾的形成和变化具有双重作用，一是随着冷空气南下，风力有所增大，有利于污染物扩散，可以缓解霾污染；二是由于冷空气南下携带大量的北方污染物，尤其是持续性补充的冷空气，加上盆地效应，使得污染物不断往金华输送并在此堆积，将会加重污染。

（4）弱降水发生前或发生时可以对空气中颗粒物有一定的洗涤作用，使大气水平能见度增大，缓解霾污染。但由于降水本身强度较弱，对空气的洗涤作用不明显，弱降水停止后反而可使大气中气溶胶粒子膨胀，能见度降低，加剧霾的形成。

通过对本文的分析可以发现，冷空气、降水对金华霾的形成和变化具有双重作用，既可以缓解霾污染，也会加重污染，如降水强度、冷空气本身强度以及冷空气带来降水的强度等的不同，对霾形成的作用完全不同。并且由于环流形势的每日变化，在地面要素或污染条件相近的条件下，霾是否出现也不尽相同。稳定的天气系统配置，如500hPa环流呈纬向型、锋区偏北、冷空气活动偏弱、南支系统不活跃，或地面形势场稳定、多均压场控制等都有利于霾形成或持续。

综上所述，在建立模型时，不仅要对引入因子的相关性进行讨论，而且要更注重对其物理意义的分析。在今后的研究中，可以增加一些数值预报中预报效果较好、较稳定且对霾形成有明确影响的要素，增加预报因子从而改善预报效果。同时，在大量个例积累的基础上，可以尝试根据每个因子的影响程度不同来设置权重，或可在实际业务中明确判定阈值。

致谢：感谢中国气象局气象干部培训学院SVM应用研究小组提供CMSVM2.0应用软件。

参考文献

[1] 陈永义, 俞小鼎, 高学浩, 等. 处理非线性分类和回归问题的一种新方法(Ⅰ)——支持向量机方法简介. 应用气象学报, 2004, 15(3): 345-354.
[2] 冯汉中, 陈永义. 处理非线性分类和回归问题的一种新方法(Ⅱ)——支持向量机方法在天气预报中的应用. 应用气象学报, 2004, 15(3): 355-365.
[3] 韦惠红, 李才媛, 邓红, 等. SVM方法在武汉区域夏季暴雨预报业务中的应用. 气象科技, 2009, 37(2): 145-148.
[4] 陈晓燕, 赵玉金, 孙文英, 等. 支持向量机方法作温度预报试验. 贵州气象, 2006 , 30(1): 31-33.
[5] 常军, 李祯, 朱业玉, 等. 基于支持向量机(SVM)方法的冬季温度预测. 气象科技, 2005, 33(s1): 100-104.
[6] 贺皓, 罗慧. 基于支持向量机模式识别的大雾预报方法. 气象科技, 2009, 37(2): 149-151.
[7] 胡邦辉, 刘丹军, 王学忠, 等. 最小二乘支持向量机在云量预报中的应用. 气象科学, 2011, 31(2): 187-193.
[8] 熊秋芬, 胡江林, 陈永义. 天空云量预报及支持向量机和神经网络方法比较研究. 热带气象学报, 2007, 23(3): 255-260.
[9] 刘西川, 高太长, 刘磊, 等. 降水现象对大气消光系数和能见度的影响. 应用气象学报, 2010, 21(4): 433-441.
[10] 赵胡笳, 马雁军, 赵明, 等. 沈阳两次降水过程能见度变化特征. 气象与环境学报, 2014, 30(2): 60-66.
[11] 俞剑蔚, 孙燕, 张备, 等. 江苏沿江一次重霾天气成因分析. 气象科学, 2009, 29(5): 664-669.