双流机场低能见度天气预报方法研究*

冯汉中[1)2)]　陈永义[3)]　成永勤[4)]　罗可生[2)]

[1)](云南大学资环学院,昆明 650000)　[2)](四川省气象台,成都 610071)

[3)](中国气象局培训中心,北京 100081)　[4)](成都双流机场空管中心,成都 610000)

摘　要

在信息量较大,而预报对象与预报因子的关系又不清楚的状况下,智能机器学习方法是解决这类问题的较好手段。利用 1997—2001 年成都站的常规探空资料和双流机场的地面观测资料,使用支持向量机(Support Vector Machines,简称 SVM)方法,选取多种核函数进行双流机场低能见度天气的预报建模试验。测试结果表明:以径向基函数和拉普拉斯函数构造的 SVM 预报模型实验效果最好,Ts 评分分别为 0.287 和 0.292,远高于双流机场低能见度天气出现的频率(0.155)。试验结果还表明:以径向基函数构造的 SVM 预报模型空报较多,漏报较少;而以拉普拉斯函数构造的 SVM 预报模型空报较少,漏报较多。因此,如果强调模型对低能见度天气预报的准确性,则应采用以拉普拉斯函数构造的预报模型,如果强调对低能见度天气的预防性,则应采用以径向基函数构造的预报模型。

关键词:低能见度天气;支持向量机;预报方法

引　言

影响双流机场能见度的天气现象主要有雾、降水、霾、烟、浮尘等,雾是其中最主要的因素。能见度与飞行的关系十分密切,低能见度现象是引发飞行事故最常见的因素。《双流机场气候志》表明:地处四川盆地的双流机场各季均可出现低能见度天气,其中以冬、秋季居多,夏季最少,且具有显著的日变化规律。低能见度天气集中出现在 07:00—09:00(北京时,下同)。双流机场低能见度日较多,且持续时间长,常导致飞行航班延误。由于导致低能见度天气出现的因素较多,影响系统复杂,预报难度较大,因而如何尽可能地预报出低能见度天气的出现,是承担机场天气预报服务保障的预报工作者最为关注的问题。在此,我们以目前最有理论保证且具有处理非线性问题能力的机器学习方法——支持向量机[1],运用成都常规探空资料和双流机场的地面观测资料对低能见度天气加以研究,以期对其预报有一定的指示意义。

本文是利用 SVM 方法对双流机场低能见度天气进行研究的结果介绍,包含 3 部分内容:首先简单介绍 SVM 分类的基本原理,然后介绍所使用的资料及建立 SVM 分类学习机的过程,最后对所建立的 SVM 分类学习机推广能力进行检验的结果分析。

1　SVM 分类方法基本原理简介

机器学习问题可概括的表述为[1]:给定训练样本$(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l)$,其中 $x_i \in \mathbf{R}^N$,为 N 维向量,$y_i \in \{-1, 1\}$ 或 $y_i \in \{1, 2, \cdots, k\}$,给出预报数据集:$x_{l+1}, x_{l+2}, \cdots, x_m$,通过训练学习建立分类模式 $M(x)$,使其不但对训练样本能够正确分类,而且具有较强的推广能力。即可以由模式对于输入的预报数据 x_i 得到正确的对应输出值 y_i。

对于训练样本集的线性二类划分问题,就是寻求函数:

$$y = f(x) = \mathrm{Sgn}((w \cdot x) + b)) \tag{1}$$

使对于 $i = 1, 2, \cdots, l$ 满足条件:

$$y_i = f(x_i) = \mathrm{Sgn}((w \cdot x_i) + b)) \tag{2}$$

其中 $w, x, x_i \in \mathbf{R}^N$, $b \in \mathbf{R}$, w, b 为待确定的参数,Sgn 为符号函数。显然,$(w \cdot x) + b = 0$ 为划分超平面,w 为法方向向量。

对于线性可分离的问题,满足条件形如式(1)的

* 国家自然科学基金项目(60072006)资助。
2004-12-28 收到,2005-04-13 收到再改稿。

线性决策函数是不唯一的。图 1 给出二维情况下满足条件的划分直线的分布区域图。落在虚线区域内的任一直线都可作为决策函数。谁是最优的决策函数,就要对其进行判断。

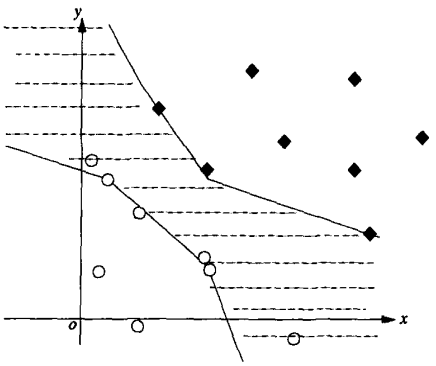

图 1　划分直线的分布区域图
(图中 ◆表示有,○表示无,下同)

V. N. Vapnik 提出一个间隔最大化原则①:寻求使间隔达到最大的划分为最优,即是对 w, b 寻优,求得最大间隔: $\max(\min(\| x - x_i \| : x \in R^N, (w \cdot x) + b = 0, i = 1, \cdots, l))$,对应最大间隔的划分超平面称为最优划分超平面,简称为最优超平面,如图 2 中的平面 L。图 2 中两条平行虚线 l_1, l_2(称为边界)距离之半就是最大间隔。

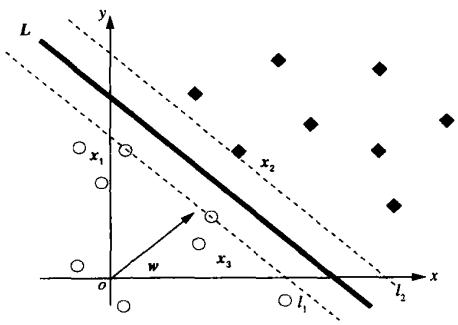

图 2　最优划分超平面示意图

最大间隔和最优超平面只由落在边界上的样本点完全确定,而不依赖于所有点,称这样的样本点为支持向量,如图 2 中的 x_1, x_2, x_3 样本点。

对于给定的训练样本集,根据相关的理论和算法,最终获得的线性支持向量机为:

$$M(x) = \mathrm{Sgn}((w^* \cdot x) + b^*) = \mathrm{Sgn}(\sum_{\text{支持向量}} a_i^* y_i (x \cdot x_i) + b^*) \quad (3)$$

式(3)中 Sgn( )为符号函数; a_i^*, b^* 为确定最优划分超平面的参数; $(x \cdot x_i)$ 为两个向量的点积。

对于线性不可分的情况,通过非线性映射 φ,将样本集映射入一个高维的特征空间,称为 Hilbert 空间,使在样本空间中的高度非线性问题在高维空间中应用线性分类的方法得以实现。

由于在特征空间中采用的是线性分类方法,所以在特征空间中的最优划分超平面分类函数的形式为:

$$M(x) = \mathrm{Sgn}((w^* \cdot \varphi(x)) + b^*) = \mathrm{Sgn}(\sum_{i=1}^{l} a_i^* y_i (\varphi(x) \cdot \varphi(x_i)) + b^*) \quad (4)$$

与式(3)相比,式(4)只是用 $\varphi(x)$ 和 $\varphi(x_i)$ 代替了 x 和 x_i。

根据 Mercer 定理[2],式(4)最终转变为:

$$M(x) = \mathrm{Sgn}((w^* \cdot \varphi(x)) + b^*) = \mathrm{Sgn}(\sum_{\text{支持向量}} a_i^* y_i K(x \cdot x_i) + b^*) \quad (5)$$

式(5)就是 SVM 方法确定的最终非线性分类的决策函数。与式(3)相比,这里只是用 Mercer 核函数的计算代替了点积的计算,在整个求解过程中不需要知道非线性映射的显式表达式。

2　建立低能见度天气的 SVM 分类学习机

2.1　资　料

1997 年 5 月—2001 年 12 月成都双流机场本站的逐小时温度、露点、气压、相对湿度、能见度,以及成都探空站逐日 08:00,20:00 的位势高度、温度、露点、风向、风速。

2.2　预报对象

按照机场服务的需要,以双流机场测站能见度小于 800 m 作为低能见度天气,以此为预报对象,进行低能见度天气的研究。

2.3　构建预报因子

低能见度天气主要由雾造成,在四川盆地,一般形成雾的大气主要特征有:近地面至 925 hPa,大气层结呈中性状态,而 925 ~ 500 hPa 高度常有逆温层,呈现出稳定状态;从地面至 400 hPa,大气相对湿

① 陈永义.支持向量机方法及其在气象中的应用.中国气象局培训中心,2004:61-64.

度较大,接近准饱和状态,近地层无风。因此,为了尽可能地描述有可能形成低能见度天气的大气空间结构,利用成都站每天 08:00 和 20:00 的 400,500,700,850,925 hPa 层次的位势高度、温度、露点、风向、风速等探空资料和双流机场 00:00—21:00 逐小时的温度、露点、气压、相对湿度等地面观测资料来构造样本空间,每一个样本由 138 个变量构成。选取这些变量,既保证了现有资料的充分使用,也能反映各个要素随时间的动态变化过程,进而构造一个比较完备的相空间。需要指出的是 SVM 是通过支持向量构造推理模型,对因子的数量没有限制[3]。

2.4 构建样本序列

因使用的是常规观测资料,故以经典统计预报方式建立样本序列,即以 t 时刻的预报因子对应 $t+1$ 时刻的预报对象。除去有问题的样本资料,获得有效样本 1108 个。

2.5 建立 **SVM** 预报模型

从总样本中以随机抽样方式获取用于建立预报模型的训练样本 936 个,剩余的 172 个样本作为测试模型优劣的实验样本。低能见度样本在训练样本中所占的比率为 15.52%,在实验样本中所占的比率为 15.51%,即低能见度样本在训练样本和实验样本中出现的几率一致。

虽然支持向量机方法具有较好的鲁棒性,但是对于具体的应用问题来说,选取不同的核,或相同核函数下选取不同的参数都会对结果产生明显的影响[3],而对于一个实际问题来说,只能有一个推广能力最强的解。因此,应用 SVM 方法解决实际问题时,存在一个在这些不同条件下的最优解中选优的问题,即使选不出最优,也要找出可以实际应用的较优结果,这就是 SVM 的参数选优问题。但至今在理论上还没有对此有好的解决方法。在建立预报模型时,使用的是中国气象局培训中心推出的 CMSVM 软件。在建立具有推广能力的预报模型时,要对核函数进行选取,然后通过训练和测试确定核函数的参数,进而确立具有较好推广能力的预报模型。研究中选取 7 种核函数(满足 Mercer 定理)进行相关的试验,这 7 种核函数为:

① 多项式核函数:

$$K(x,y) = [s \cdot \sum_{i=1}^{n}(x_i \cdot y_i) + r]^d,$$

② 径向基核函数(RBF):

$$K(x,y) = e^{-g \cdot \sum_{i=1}^{n}(x_i - y_i)^2},$$

③ 对称三角形核函数:

$$K(x,y) = \prod_{i=1}^{n} \max(1 - u|x_i - y_i|, 0),$$

④ 柯西核函数:

$$K(x,y) = \prod_{i=1}^{n} \frac{1}{1 + u(x_i - y_i)^2},$$

⑤ 拉普拉斯核函数:

$$K(x,y) = \prod_{i=1}^{n} e^{-u|x_i - y_i|},$$

⑥ 双曲正割核函数:

$$K(x,y) = \prod_{i=1}^{n} \frac{2}{e^{u(x_i - y_i)} + e^{-u(x_i - y_i)}},$$

⑦ 平方正弦基数核函数:

$$K(x,y) = \prod_{i=1}^{n} \frac{\sin^2[u(x_j - y_i)]}{u^2(x_i - y_i)^2}, u > 0。$$

试验结果显示出(表 1),以对称三角形核函数和柯西核函数分别建立的预报模型对训练样本的回报最好,Ts 评分高达 0.784,但将对应的模型用于实验样本的预报结果(表 2)表明,这两种核函数构造的预报模型漏报次数较多,尽管假警报率较低,但概括率较小,实际使用的价值不大。

表 1 由不同函数构造的 **SVM** 分类学习机对训练样本进行回报的评价

样本	核函数	构造模型的参数		准确率/%	概括率/%	Ts 评分
训练样本	多项式	$c=10$	$d=1$	55.43	44.74	0.329
	径向基	$c=420$	$g=0.07$	51.77	83.33	0.469
	对称三角形	$c=2$	$u=0.2$	98.37	79.39	0.784
	柯西	$c=2$	$u=0.2$	98.37	79.39	0.784
	拉普拉斯	$c=4$	$u=0.1$	95.50	83.77	0.806
	双曲正割	$c=11$	$u=1.1$	92.02	75.88	0.712
	平方正弦基	$c=190$	$u=0.5$	83.70	83.33	0.717

注:c 为容错参数。

从表 2 可以发现,以径向基核函数和拉普拉斯核函数分别建立的预报模型,尽管对训练样本的回报水平略差于前述的两种预报模型,但其对实验样本的预报效果最好,Ts 评分分别为 0.287 和 0.292,远高于

低能见度天气在样本中出现的频率(15 .51 %) ,具有显著的正预报技巧。

从表 2 还可以看出 ,以径向基核函数构造的预报模型 ,空报次数较多(52 次) ,漏报次数较少(25 次) ;而以拉普拉斯核函数构造的预报模型 ,空报次数较少(9 次) ,漏报次数较多(37 次) 。因此 ,如果强调模型对低能见度天气预报的准确性 ,则应采用拉普拉斯核函数构造的预报模型 ,如果要强调对低能见度天气的预防性 ,则应采用径向基核函数构造的预报模型。

表 2　由不同函数构造的 SVM 分类学习机对实验样本进行试报的评价

样本	核函数	正确次数	空报次数	漏报次数	准确率/ %	概括率/ %	假警报率/ %	Ts 评分
实验样本	多项式	21	24	35	46 .67	37 .50	7 .87	1 .263
	径向基	31	52	25	37 .35	55 .36	17 .05	0 .287
	对称三角形	9	3	47	75 .00	16 .07	0 .98	0 .153
	柯西	4	1	52	80 .00	7 .14	0 .33	0 .070
	拉普拉斯	19	9	37	67 .86	33 .93	2 .95	0 .292
	双曲正割	16	22	40	42 .11	28 .57	7 .21	0 .205
	平方正弦基	8	16	48	33 .33	14 .29	5 .25	0 .111

3　以径向基核函数为基础建立的 SVM 低能见度天气预测模型

研究低能见度天气的预测方法 ,目的在于给出有警示性提示的预报结果 ,提示相关的技术人员仔细分析资料 ,最终对提示结果进行肯定或否认。依据前面的分析 ,选取以径向基核函数为基础进行训练建立的 SVM 预测模型作为业务试用模型 ,构成模型的核函数的参数值为 $g = 0 .07$, $c = 420$,支持向量个数为 601 。表 3 给出了构成预测模型的部分支持向量 , a_i 为每一个支持向量对应的系数值 , x_i 为构成支持向量的每一个预报因子(对各因子值进行了归一化处理) 。模型的最终预测结果 ,由预测样本和支持向量通过式(5) 计算得出 ,若决策函数 $M(x) < 0$,则预测无雾 ;若 $M(x) \geqslant 0$,则预测结果为有雾。

表 3　构成双流机场低能见度天气预测模型的部分支持向量

	a_i	$x1$	$x2$	$x3$	$x4$	$x5$	$x6$	. . .	$x138$
支持向量	- 225 .28	0 .2952	0 .3419	0 .4079	0 .8460	0 .2628	0 .3930	. . .	0 .4420
	- 93 .25	0 .3119	0 .3584	0 .5728	0 .8460	0 .2897	0 .4207	. . .	0 .3470
	- 2 .60	0 .5265	0 .3721	0 .6079	0 .3517	0 .4665	0 .4710	. . .	0 .3787
	0 .43	0 .2366	0 .2814	0 .4167	0 .8319	0 .2039	0 .3351	. . .	0 .2205
	116 .31	0 .7912	0 .8369	0 .1588	0 .8743	0 .7398	0 .8487	. . .	0 .2627
	249 .55	0 .5962	0 .6251	0 .3956	0 .8319	0 .5362	0 .6448	. . .	0 .3365
	480 .96	0 .3063	0 .3556	0 .4430	0 .8602	0 .2682	0 .4031	. . .	0 .2521
	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

支持向量学习机方法是在看似杂乱无章的资料中寻求关键信息 ,记忆关键样本 ,通过对历史资料的学习 ,积累知识 ,类似于人脑的学习过程。对部分关键样本的本站气温、露点、气压、相对湿度进行分析后发现 ,出现雾与否 ,不在于温度的高低 ,而在于相对湿度的大小 ,温度高可以有雾发生 ,温度低也可以产生雾 ,关键看湿度的演变趋势。如图 3 所示 ,样本 a ,b ,c 为无雾的关键样本 ,样本 A ,B ,C ,D 为有雾的关键样本 ,有雾出现之前 ,00 :00 —21 :00 的相对湿度都较大 ,且在 15 :00 以后 ,均呈现增加的趋势 ;无雾出现之前 ,10 :00 —16 :00 相对湿度较小 ,或20 :00 —21 :00相对湿度呈下降趋势。如果对支持向量的各要素进行仔细分析 ,则可以认识形成雾与否的要素空间分布或随时间的演变趋势 ,给出有无雾出现的基本特征 ,为实际预报积累知识。如果对每一个个例进行分析 ,就类似于经验总结。但由于雾的形成机理比较复杂 ,且构成预报模型的支持向量较多 ,逐一分析不太可能 ,所以借助于机器学习方法也是解决问题的有效方式 ,业务

试验结果也表明:利用支持向量机方法建立的预测模型具有较好的稳定性和推广能力。

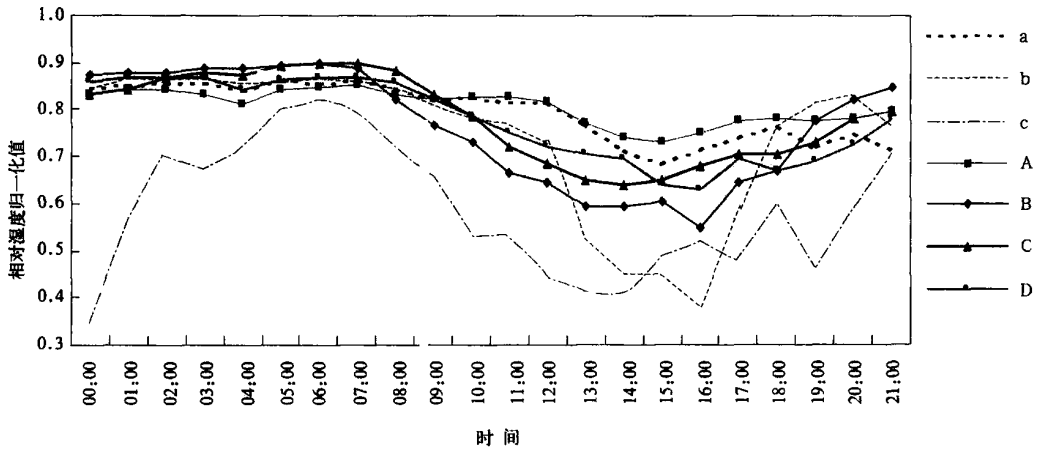

图 3 作为支持向量的关键样本中相对湿度的逐小时演变情况(虚线:无雾出现,实线:有雾出现)

4 结 语

目前,对低能见度天气的预报仍以统计预报方法为主,尽管随着数值预报的发展,现在也有数值释用和雾模式预报,但许多试验结果表明:雾模式仅有一定的机理分析用途,难以实际预报,而数值释用方法必须对模式因子做适当有效的处理,否则难以有效,因而除了经验预报以外,统计法仍是目前较为实用的方法。严格说来,SVM 方法也是一种统计预报方法,只不过它是从样本资料中通过自学习的方式获取预报知识,建立预报模型,且 SVM 方法是通过核函数实现从低维到高维空间的非线性映射,以隐式方式间接地表述预报对象与预报因子之间的高度非线性关系,最终通过支持向量来刻化因子与对象之间的依赖关系。在预报因子与预报对象之间的关系并不明了之前,基于机器学习的"黑箱"输出,仍是目前比较实用的方法。

试验表明:双流机场低能见度天气的 SVM 分类预报模型,对核函数的选取有较大的依赖关系,如果强调对低能见度天气预报的准确性,则应采用拉普拉斯核函数构造的 SVM 预报模型,如果强调对低能见度天气的预防性,则应采用径向基核函数构造的 SVM 预报模型。业务试验结果表明:利用支持向量机方法建立的预测模型具有较好的推广能力。

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] Courant R, Hilbert D. Method of mathematical physics, Volume I. Springer Verlag,1953.

[3] 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):355-364.

A Study on the Forecast Method of the Low Visibility Weather of Shuangliu Airport

Feng Hanzhong[1)2)]　Chen Yongyi[3)]　Cheng Yongqin[4)]　Luo Kesheng[2)]

[1)] (*Institute of Resource and Circumstances of Yunnan University, Kunming* 650000)

[2)] (*Sichuan Provincial Meteorological Observatory, Chengdu* 610072)

[3)] (*China Meteorological Administration Training Center, Beijing* 100081)

[4)] (*Chengdu Shuangliu Airport Management Center, Chengdu* 610000)

Abstract

At Shuangliu International Airport the frequent, long-time low visibility often leads to the delay of scheduled flight and affects the safety of the planes. With the development of social economy and the air transport, the low visibility weather of Shuangliu Airport arouses great concern of the air control authority of the airport. Because a number of the factors can result in the low visibility weather, the factor mechanism is complex, and meanwhile the correlativity between the factors and the low visibility weather is insignificant, meteorologists are concerned most about how to forecast such weather as possible as they can. While there are huge amounts of information, the relation between the forecast object and the factors is unclear. Given this, intelligent machine learning technique is a good method to solve this sort of problems. In order to improve the forecast, Support Vector Machines (SVM), an intelligent machine learning technique that can solve the nonlinear problems is employed in the research to study the forecast method of the low visibility weather of the Shuangliu Airport. SVM method nonlinear is mapped by kernel function from lower dimensional space to higher. The higher nonlinear correlativity between the factors and the forecast object is indirectly expressed in implicit expression. Finally, the dependence of the factor and the object is depicted, and the model is set up by support vector. Hence, the model is not only related to the forecasting factor, but the kernel function as well. By using the data from Chengdu radiosound observatory and Shuangliu Airport surface observatory through 1997 to 2001, the forecast models of low visibility weather of Shuangliu Airport are built by mean SVM method with several kernel functions. Test results show the forecast model constructed with the radial base kernel function and the forecast model constructed with Laplace kernel function are better than the others, in which Threat Score are 0.287 and 0.292 separately, far above the frequency(0.155) of low visibility weather occurring at Shuangliu Airport. Test results also show, in the SVM model constructed with the radial base kernel function, the false alarm is higher and the omitted alarm is lower; in the SVM model constructed with Laplace kernel function, it is opposite. Therefore, if the accuracy of low visibility weather forecast is emphasized, the model constructed with Laplace kernel function should adopted; but if the precaution is emphasized, the model constructed in radial base function should be selected.

Key words: low visibility weather; Support Vector Machines; forecasting method