SVM 方法在乌鲁木齐机场跑道视程预测中的应用

朱蕾

（新疆空管局气象中心，乌鲁木齐 830016）

摘要

支持向量学习机(SVM)是基于统计学习理论的模式分类器,本文将SVM方法应用于乌鲁木齐机场跑道视程的预测。利用2004—2006年11月至次年二月的每半小时的自动观测资料，分别建立了跑道视程大于800米和小于800米2个SVM推理模型,并进行了跑道视程的预测试验,结果显示：对训练集的回报、对试验集的预报、对检验集预测的TS评分在48%-75%之间，较高TS评分表明所建SVM推理模型对跑道视程具有一的预测能力,展示了SVM的优越性和推广前景,可在跑道视程预测业务中参考应用。

关键词：支持向量机(SVM) ;推理模型;跑道视程;分类预测

引言

随着数值预报以及数值预报产品释用技术的不断发展,传统的基于天气图知识的经验预报在实际的天气预报业务中逐渐被淡化。基于统计理论的预报方法如相关分析、回归方法、判别方法、卡尔曼滤波等虽然在数值预报释用中都发挥了重要作用,但它们都是建立在线性相关基础上的,在实际应用中存在明显的局限,影响到预报效果。近年来,具有处理非线性问题能力的人工神经元模型方法在气象上获得了不少应用[1-3]。然而由于神经网络的过学习问题和对因子的过敏感问题，通过基于系数梯度的陡降方法来构造对预期函数的逼近解，通过有限给定数目的观测所得的解往往是局部最优[4]，预报效果仍然不理想。

在实际天气预报工作中,使用较多的仍然是基于不同天气的各种预报模型的知识积累和近似的记忆。可预报性问题的系统研究开始于20世纪60年代,研究的一个主要目的是对所谓决定论预报的可能时效作出科学的估计。在在90年代初期比较盛行的基于专家经验的专家预报系统,正是这种知识积累的客观表现,但因知识更新手段的繁琐而逐渐被预报员抛弃.随着计算机技术的发展和智能技术的进步,智能机器识别技能也得到了很好的发展。基于支持向量理论的支持向量机(Support Vector Machine 简称SVM)方法[5-7]是一种新颖的小样本机器学习方法,它具有从海量的信息中自动识别并提取关键信息的特点,适合处理本质上的非线性问题。SVM于20世纪90年代逐渐成熟发展起来,近年在气象领域已得到应用。陈永义、俞小鼎[8]等人2002年开始将SVM引入气象领域,冯汉中、陈永义[9]等介绍了处理非线性分类、回归等问题的SVM这一新方法的基本原理及在气象研究业务中的应用前景。

1 支持向量机分类方法的基本原理简介

SVM是Vapnik V N等人提出的基于统计学习理论的小样本学习方法,它依据结构风险最小化归纳原则,以特征空间中样本到最优划分超平面的最大间隔作为划分依据,建立线性学习机;通过在高维空间中构造较低VC维的函数集,从而获得好的推广能力;由于在线性SVM学习机中样本点仅以点积形式出现,依据Mercer定理引入核函数很容易得到非线性学习机,从而为解决非线性问题提供了较好的方法。

SVM的基本思想(如图1)是把低维样本空间中的难于线性划分解决的样本集通过非线性映射Φ映射到高维特征空间中,在高维特征空间寻求最优划分超平面,从而实现了样本空间的非线性分类。

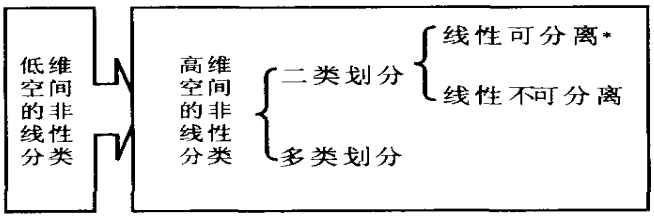

图 1 SVM基本思路

其原理主要是:给定1个样本集$\{(x_i, y_i), i=1,--,1\}$,可以把寻求最优划分决策函数的问题转化为如下的最优化问题:

目标函数: $\min(\frac{1}{2}\|w\|^2 + c\sum_i \xi_i)$

约束条件: $y_i((w^{\square}\cdot x_i)+b) \geq 1-\xi_i \qquad \xi_i \geq 0$

应用KKT(Karush2Kuhn2Tucker)条件可求得上述凸约束下的凸规划的全局最优超平面决策函数:

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

式中Sgn( )为符号函数; (x_i, y_i) 为样本集, $x_i \in R^N$, $y_i \in \{-1,1\}$; ξ_i 为松弛变量; a_i^*, b^* 为确定最优划分超平面的参数, (•)为2个向量的点积;w为超平面的方向向量。

为了解决非线性划分问题,将上述训练向量 x_i 通过函数φ被映射到一个高维以至于无穷维的特征空间中,在特征空间中寻求最优解过程中应用了Mercer定义的核函数展开定理[10]:

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

式中 $K(x_i, x_j)$ 称为核函数,所以在整个求解过程中根本不需要知道非线性映射φ的显式表达式。因在高维特征空间中采用的是线性分类方法,故在特征空间中的最优划分超平面分类函数的形式为:

$$\begin{aligned} M(x) &= Sgn((w^{\square}\cdot \phi(x)) + b^{\square})) \\ &= Sgn(\sum_{i=1}^{1} a_i y_i (\phi(x)\cdot\phi(x_i)) + b^{\square} \end{aligned} \qquad (2)$$

(2) 式用 $\phi(x)$ 和 $\phi(x_i)$ 分别代替了(1)式中的 x 和 x_i。根据Mercer定理, Mercer核展开表示成:

$$K(x, y) = \sum_i \lambda_i \phi_i(x)\phi_i(y)$$

如果把从样本空间到特征空间的非线性映射 ϕ 定义为:

$$\phi(x)=(\sqrt{\lambda_1}\phi_{1,}(x),\sqrt{\lambda_2}\phi_{2,}(x),\cdots\cdots,\sqrt{\lambda_k}\phi_{k,}(x),\cdots\cdots)$$

则显然有:

$$K(x,x_i)=(\phi(x)\cdot\phi(x_i)) \qquad (3)$$

将(3)式代入(2)式可得:

$$M(x)=Sgn((w^{\square}\cdot\phi(x))+b^{\square}))$$

$$=Sgn(\sum_{\text{支持向量}}a_iy_iK(x,x_i)+b^{\square}) \qquad (4)$$

(4)式就是SVM方法确定的最终非线性分类的决策函数。由于用Mercer核函数的计算代替了点积的计算,因此与线性方法相比几乎不增加多少计算量。但特征空间的线性划分却对应于原来样本空间的高度非线性划分,如图2。

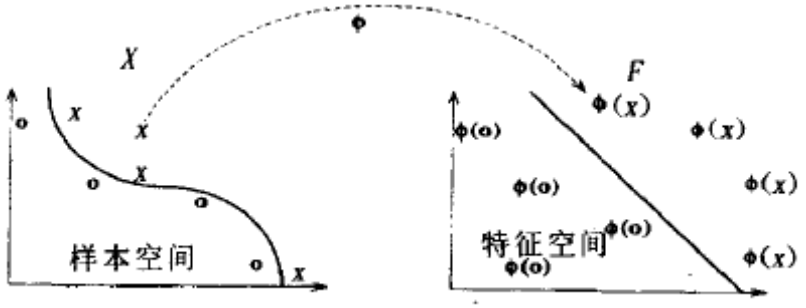

图 2 非线性映射图

2 建立跑道视程预测的SVM分类学习机

跑道视程（Runway Visual Range ,简称 RVR）指的是在跑道中线，航空器上的飞行员能看到的跑道面上的标志或跑道边界灯或中线灯的距离。一般来说 RVR 是指从飞机的接地点看到的能见距离，机场例行天气报告（METAR）中的 RVR 数值的高低是航空器能否安全起降的重要条件，按照《民用航空气象地面观测规范》，由于观测时所用目标物的不同，跑道视程和能见度是两个不同的概念，虽然如此，能见度和 RVR 有很大的相关性，统计资料表明，乌鲁木齐机场（以下称本场）能见度和 RVR 的相关系数达 0.8 以上，较低能见度天气出现时，RVR 也相应降低。机场预报（TAF）并不包含 RVR 的变化信息，但航空用户如管制员、飞行员、签派员等更关注 RVR 的变化趋势，RVR 的准确预测在低能见度出现时的对外服务中有着举足轻重的作用。

影响本场RVR的天气现象主要有雾、烟、沙尘、强降水等,其中大雾和浓烟是最主要天气的天气现象。本场冬季的烟雾日较多,且持续时间长,较低的RVR常导致航班延误、备降或取消。新疆空管局航班正常办的统计资料表明：2004、2005、2006年冬季（每年11月到次年2月）因低标跑道视程准造成的航班不正常分别为566、518、718个架次，占不正常总班次的百分比达89.3%。由于导致低RVR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RVR气的出现,是航空气象预报工作者颇为关注的问题。本文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,运用本场的地面自动观测系统的资料对对跑道视程进行研究，以期对跑道视程的预测有一定的指示和参考意义。

2.1 资料

2004年—2006年的11-3月本场每半小时自动观测资料，包括风向、风速、温度、露点、气压、相对湿度、光学能见度。

2.2 预报对象

按照机场服务的需要,以本场跑道视程小于800米作为低RVR天气,以此为预报对象,进行低RVR的研究。

2.3 构建预报因子

本场冬季的低跑道视程天气主要由雾造成,形成雾的大气主要特征有:地面气温在0℃～－10℃,特别是在－5℃－－10℃,大气相对湿度较大,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与大雾天气形成有关的气象要素,利用本场自动观测每半小时的风向、风速、温度、露点、气压、相对湿度、光学能见度等地面观测资料来构造样本空间,每一个样本由78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[8] 。

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13000余个。

2.5 建立SVM预报模型

在建立预报模型使用的是中国气象局培训中心推出的CMSVM 软件，CMSVM采用Thorsten Joachims的快速SVM算法，在保留SVM^{light}内核的基础上，通过减少人为参与，实现了计算机自动选取最优模型的功能，提高选择最优模型的效率。

2.5.1 建立数据集

按程序要求按规定格式建立用于确定预报模型的训练样本集，确定可用于预报的最优模型参数的试验样本集，以及用于检验预报效果的检验样本集。

将预测对象进行分类，即将乌鲁木齐国际机场07、25方向以及中间的RVR小于等于800米的归为+1类，大于800米的归为-1类。

使用的原始数据为本机场2004年1月1日至2006年6月30日的每分钟自动观测系统资料，包括机场07、25方向风向风速，机场温度、露点温度、相对湿度、修正海平面气压以及光学能见度。通过程序提取出11月到2月每半小时的气象要素值，剔出问题数据，跑道每个方向都得到13000多个样本。

将各气象要素的观测值，以及计算得到的各气象要素随时间的变化趋势等作为因子，共构建得到78个因子。由于各因子由不同的的气象要素值组成，为了避免各个因子之间的量级差异，对全部样本的每一因子分别作归一化处理，使每一个因子的数据落入区间[-1，1]。本文数据处理使用了LIBSVM中附带的归一化处理工具。具体算法如下是：

$$\overline{x} = \frac{x_i - \min\limits_k(x_k)}{\max\limits_k(x_k) - \min\limits_k(x_k)}$$

其中$\min\limits_k(x_k)$和$\max\limits_k(x_k)$分别为第k个因子数据的最大值和最小值。这里的最大、最小值不仅是针对训练集，还应充分考虑检验集、试验集以及未来可能的数据。

建模使用的最终数据格式如下所示：

分类	因子1	因子2	因子3	因子4	因子5	… …
-1	1:0.818182	2:-0.73913	3:1	4:0.142857	5:-0.777778	… …
1	1:-1	2:-0.391304	3:1	4:-0.885714	5:-0.555556	… …
-1	1:-0.818182	2:0.217391	3:1	4:0.0857143	5:-1	… …
-1	1:1	2:0.913043	3:1	4:0.0857143	5:-0.777778	… …

在SVM建模程序中选取不同的参数训练，可得到不同的预测模型；具体哪一种参数值训练得到的预报模型最好，就要在试验集中对其进行测试；最后将训练得到的最优模型用于检验集，以检验模型的预报效果。故本文将总样本分为三部分，从2004年1月、2月、11月、12月以及2005年1月、2月的数据作为训练集，共8123个；试验集以及检验集是从剩余的数据中互不重复地提取出1777个。

2.5.2 确定支持向量机的核函数类型

对于核函数的选取是SVM建模的又一核心内容，合适的核函数，可提高预测精度。凡是满足Mercer定理的对称正定连续核都可以作为支持向量机的核函数，因此核函数的类型有很多，常用的有多项式、径向基、对称三角形、柯西、拉普拉斯、双曲正割、平方正弦基。由于不同的核函数构造出来的支持向量机是不同的，具体哪种核函数得到的结果最优，需要多次试验来选定。本文以径向基核函数为基础建立RVR的SVM预测模型。

2.5.3 支持向量机的参数寻优

确定核函数后，选取不同的参数可以得到不同的预测模型，但对于一个实际问题，只能有一组推广能力最强的参数，也就是参数选优的问题，但是目前并没有很好的方法解决这个问题。

CMSVM建模程序依据设定的评优标准，在所有循环生成的模型中确定一个最优者，作为最后的模型。该模型对应的参数称为最优参数。

CMSVM模型择优的标准有TS评分、预报准确率、预报概括率，具体定义如下：

$$TS\text{评分}=\frac{\text{预测正确的正样本数}}{\text{所有正样本数+预报错误的正样本数}}$$

$$\text{预报准确率}=\frac{\text{预报正确的正样本数}}{\text{预报为正样本的数目}}$$

在分类运算中，根据应用实际，采用了TS评分机制，选用试验文件TS评分标准当作评选最优模型标准，选取试验文件TS评分值达到最大的模型为最优模型。最终确定最优参数C=100，g=0.0085。

2.6 预测试验结果分析

通过试验集和检验集的参数逼近，最终获得较为理想的预测模型，通过对训练集的回报以及试验集和检验集预测，得到统计数据（见表2和表3）：

跑道07方向：通过训练和参数逼近,使得所建立的低RVR事件的SVM预测模型的支持向量为1111个,占训练样本的13.68%,模型对训练集、试验集、检验集预测的综合正确率高达96.54%，综合TS评分达到66.64%，综合准确率为87.69%。

跑道中间方向：通过训练和参数逼近,使得所建立的低RVR事件的SVM预测模型的支持向量为1384个,占训练样本的17.04%,模型对训练集、试验集、检验集预测的综合正确率高达

95.97%，综合TS评分达到65.82%，综合准确率87.52%。

跑道25方向：通过训练和参数逼近,使得所建立的低RVR事件的SVM预测模型的支持向量为1607个,占训练样本的19.78%,模型对训练集、试验集、检验集预测的综合正确率高达96.03%，综合TS评分达到69.61%，综合准确率为82.79%。

表2　跑道各方向低RVR事件SVM预测模型的试验结果Ⅰ

跑道方向	检验项目	正确分类样本	非正确分类样本	正确率
07 方向	训练集	7929	194	97.61%
	试验集	1675	102	94.26%
	检验集	1669	108	93.92%
	综合	11273	404	96.54%
跑道中间	训练集	7896	227	97.21%
	试验集	1655	122	93.13%
	检验集	1656	121	93.19%
	综合	11207	470	95.97%
25 方向	训练集	7867	256	96.85%
	试验集	1670	107	93.98%
	检验集	1676	101	94.32%
	综合	11213	464	96.03%

表3　跑道各方向低RVR事件SVM预测模型的试验结果Ⅱ

跑道方向	检验项目	正样本数	正样本比例	预测正样本数目	预测正确	空报	漏报	TS 评分	准确率
07 方向	训练集	684	8.42%	578	534	44	150	73.35%	92.38%
	试验集	179	10.07%	137	107	30	72	51.20%	78.10%
	检验集	178	10.02%	130	100	30	78	48.08%	76.92%
	综合评分	1041	8.91%	845	741	104	300	66.64%	87.69%
跑道中间	训练集	789	9.71%	704	633	71	156	73.60%	89.91%
	试验集	230	12.94%	160	134	26	96	52.34%	83.75%
	检验集	227	12.77%	170	138	32	89	53.28%	81.18%

以上计算及分析显示:以径向基函数作为核函数,采用参数逐步逼近法建立的本场低RVR事件的SVM预测模型，支持向量占训练样本的13%-19%,是对训练样本的浓缩，建立的预测模型只需通过少量的支持向量，实现对本场低RVR事件的描述。对跑道三个方向的训练集回报的TS评分都在85%以上，TS评分均远远高于正样本的出现频率即低RVR事件出现的频率，具有正的预报技巧，说明我们建立的模型对跑道方向的低RVR事件都有较好的预测推理能力[9]；对跑道三个方向的试验集以及检验集预测，TS评分在48%-59%之间，同样比正样本比例的频

率高，具有一定的预测技巧[9]，表明所建模型得到的预测结论具有一定的稳定性和推广能力。

三 小结与讨论

影响乌鲁木齐机场低RVR的天气主要是烟雾。烟雾的预报仍以经验预报和统计预报方法为主,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。

支持向量机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。

本文采用径向基核函数所建立的SVM预测模型对训练集的回报、对试验集的预报、对检验集预测的TS评分在48%-75%之间,较高TS评分表明所建SVM推理模型对跑道视程具有一定的预测能力,展示了SVM的优越性和推广前景,可在跑道视程预测业务中参考应用。

以上工作的不足之处：首先未对因子在模型中所起的作用进行深入研究。尽管SVM方法对因子的数量无明显的限制,但无用(分辨能力不强)因子的引入有增加杂噪的可能,从而对推理效果产生影响。对因子进行分析,剔除一些意义不大的因子,从理论上说可改善方法的推理效果。其次对于核函数的选取是SVM建模的又一核心内容，合适的核函数，可提高预测精度，采用径向基核函数所建立的SVM低RVR预测模型是否是最优模型还有待于进一步研究。

未来研究中将要改进的方面：

（1） 由于低RVR的天气出现与否与天气形势有密切关系，SVM模型建立过程中因子的选取应考虑天气形势。

（2） 冬季出现较低RVR时，近地面至900hPa附近大气层结呈中性弱不稳定状态,而900hpa～800hPa高度为逆温层，地面至逆温层下部湿度大。为了尽可能地描述有可能形成低RVR天气的大气空间结构,将增加乌鲁木齐站400、500、700、850、925hPa层次的位势高度、温度、露点、风向、风速等探空资料作为建模因子。

（3） 采用不同的核函数和参数建模，进行比较，选取最优SVM低RVR模型为业务工作中RVR的预测提供指导。

参考文献：

[1]胡江林,张礼平. 神经网络模型预报湖北汛期降水量的应用研究. 气象学报,2001 ,59 (6) :776～782.

[2]张韧,蒋国荣,余志豪,等. 利用神经网络计算方法建立太平洋副高活动的预报模型. 应用气象学报,2000 ,11 (4) :474～477.

[3]熊秋芬等. 神经网络方法在静止气象卫星多通道资料估算降水中的应用. 气象,2002 ,28 (9) :17～21.

[4]Vapnik V N. Statistical Learning Theory. John Wiley & Sons , Inc., New York , 1998.

[5]VAPN IK V N. The nature of statistical learning theory[M]. Sp ringerVerlag, New York, 2000. (中译本:张学工译. 统计学习理论的本质.北京:清华大学出版社,2000:10～45.

[6]BURGES C J. A tutorial on support vector machines for pattern recognition[J]. DataMining and Knowledge Discovery, 1998 (2):127～167.

[7]CR ISTIAN IN IN , SHAWA2TATLOR J.An Introduction of Support Vector Machines and other kernel_based learning methods [M]. Cambridge: Cambridge University Press,2000: 23～66.

[8]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[J] 应用气象学报,

2004, 15 (3) : 345- 353
[9]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[J] 应用气象学报, 2004,15(3):356～365
[10]COURANTR, H ILBERT D. Method of mathematical physics[J].Sp ringerVerlag,1953,1:5 ～40.