施萧，徐幼平，胡邦辉，等. 支持向量机在雷暴预报中的应用[J]. 气象，2012，38(9)：1115-1120.

支持向量机在雷暴预报中的应用*

施 萧[1] 徐幼平[2] 胡邦辉[3] 成 巍[2]

1 中国人民解放军 63796 部队气象室，西昌 615000

2 北京应用气象研究所，北京 100089

3 解放军理工大学气象学院，南京 211101

提 要：论文利用 2002—2006 年 AREM 模式产品和常规观测报文资料，综合运用改进的 K 平均聚类和主成分分析等方法，基于 MOS 原理逐月建立了最小二乘支持向量机和线性规划支持向量机的单站雷暴释用预报模型，并针对海口站 2007 年 5—8 月进行了具体的预报。结果表明：支持向量机结合 AREM 模式产品进行雷暴的释用预报是合适、有效的，而且主成分分析对预报结果的提高也起到了积极的作用。

关键词：雷暴，AREM，释用预报，支持向量机，主成分分析

Application of Support Vector Machine to Thunderstorm Forecasting

SHI Xiao[1] XU Youping[2] HU Banghui[3] CHENG Wei[2]

1 Meteorological Division of PLA 63796 Troops, Xichang 615000

2 Beijing Institute of Applied Meteorology, Beijing 100089

3 Institute of Meteorology, PLA University of Science and Technology, Nanjing 211101

Abstract: In the paper the K-means clustering of the improved algorithm, the principal component analysis (PCA) and other methods are used to establish the interpretation forecasting model of thunderstorm by the least squares support vector machine (LS_SVM) and linear programming support vector machine (LP_SVM) based on MOS theory monthly in terms of AREM prediction products and conventional observation data during 2002 to 2006. And use the data at Haikou Station for testing from May to August 2007. The results show that, combining with SVM and AREM products to interpret the forecast products is feasible. The PCA also plays a positive role in improving the forecast accuracy.

Key words: thunderstorm, AREM model, interpretation forecast, support vector machine, principal component analysis

引 言

雷暴是一种灾害性天气，由于雷暴的时空尺度以及发生概率都较小，因此雷暴一直是预报中的重点和难点。现阶段，对雷暴常见而有效的预报手段主要有卫星、雷达等实况资料的外推以及天气学概念模型等方法，即雷暴预报的预报时效及有效性仅仅适用于雷暴的临近预报[1-4]，而雷暴的短期预报却仍属于较难预报的范畴。随着数值模式的发展和模式产品精度相应的提高，模式产品的释用为雷暴短期预报准确率的提高提供了一种很好的解决途径。

模式产品的释用主要是建立预报对象与预报因子之间的关系。在模式既定的情况下，预报效果的提高与预报因子的选择和对象，以及因子之间关系的形式有关。近年来，因子分析技术在释用预报中

* 2011 年 6 月 14 日收稿； 2012 年 2 月 5 日收修定稿

第一作者：施萧，主要从事航天气象保障. Email：shi_xo@163.com

已经有了较为成熟的应用；同时一些较好的建模新方法也在解释应用领域得到引进，比如逻辑回归判别、神经网络和支持向量机等方法。其中支持向量机方法在被冯汉中等[5]引入到气象领域之后，在暴雨、大雾等的预报中均有较为成功的应用[6-7]。

本文以在暴雨预报表现良好的 AREM 模式产品和常规观测报文为基础，运用支持向量机方法以及主成分分析技术，建立单站的雷暴 MOS 预报模型。

1 预报流程的前处理

1.1 资料介绍

AREM 模式基于中国科学院大气物理研究所大气科学和地球流体力学国家重点实验室发展的 REM η 坐标暴雨数值预报模式[8]。本文所用版本为 AREM V 2.4，模式分辨率是 37 km，垂直层次 22 层，模式顶高 20 hPa，积分区域为 14°～51°N、74°～136°E。模式的初始场为 NCEP 再分析场与常规报文资料的融合。模式起报时间为当日世界时 00 UTC(下同)，积分时间为 36 h，模式产品输出间隔为 6 h。鉴于 E 网格特点，其输出产品的分辨率0.5°×0.5°。在该模式中，云和降水过程采用 BIAM 显式冷云微物理过程方案和改进的 Betts 对流调整方案，行星边界层过程采用非局地边界层参数化方案，辐射方案采用基于 Benjamin 理论的 MM5 辐射方案，地表通量处理方案采用多层结通量廓线方案。

本文使用的资料为 2002—2007 年汛期 5—8 月的 AREM 模式产品和同期的常规观测报文资料，预报时效为 12～36 h，其中建立模型时间为 2002—2006 年，预报试验的时间为 2007 年。

1.2 雷暴指标的介绍与计算

论文中采用的指标分为 4 类，合计 72 个。

(1) 模式输出量：地面温度和气压；高空温度、露点、风场、散度、涡度、垂直速度和位势高度，其中高空量有 850、700、500 和 300 hPa 4 层。这 4 层代表着低(850 hPa)、中低(700 hPa)、中(500 hPa)和高空(300 hPa)。

(2) 能量指标[9]：对流有效位能(CAPE)、最优对流有效位能(BCAPE)、下沉对流有效位能(DCAPE)、对流抑制能(CIN)、925 hPa 总能量(TEI_{925})、850 hPa 总能量(TEI_{850})。

(3) 不稳定度指标[9]：Adedokun 1 指数、Adedokun 2 指数、Rackliff 指数、Boyden 指数、Bradbury 指数、Cross Totals 指数、Vertical Total 指数、Total Total 指数、Lifted 指数、K 指数、Showalter 指数、Potential Instability 指数、S 指数、Jefferson 指数、Humidity 指数、Deep Convective 指数、KO 指数、Thompson 指数、自由对流高度、平衡高度、0℃层高度和 $\Delta\theta_{se}$。

(4) 综合指标[9]：Severe Weather Threat 指数、SWISS00 指数、SWISS12 指数、Yonetani 指数、Modified Yonetani 指数、Storm Relative Helicity 指数、Energy Helicity 指数、Storms Severity 指数、Bulk Richardson Number 指数和 Wind 指数。

由于后续的建模是针对站点，因此，因子的计算按照以下规则进行：(1)考虑到雷暴量变到质变的积累过程和常规报文的 6 h 间隔，这里选择预报时刻和预报前一时刻的加权和进行因子场的计算，权重值以预报时刻为主。(2)雷暴往往发生在站点附近地区，对于某一站点，雷暴指标除了该站点外还需要考虑到站点周围的格点。这里模式产品的水平分辨率近似 50 km×50 km，雷暴的水平尺度也就数 10 km，因此仅选择距离站点最近的两个格点。论文中站点的雷暴指标值采用周围 4 个格点的距离加权平均进行插值处理。

1.3 样本和因子的处理

雷暴属于小概率事件，小概率事件的预报建模一般都要求样本集的尽量均衡。在小概率天气现象样本集的均衡中，首先需要考虑样本的消空，消空的实现一般借用消空指标库的建立。通过试验发现：如果采用多个消空因子进行消空预报时，很容易错过有雷暴的天气。因此，这里仅采取 1 个消空因子作为消空的条件。

正常情况下，对于雷暴样本筛选和均衡，单靠因子消空还是不够的。在同类中，从多到少，或选择压缩，或选择提取。鉴于计算的简便实用，这里采用 K-平均聚类[10]的方法进行样本的压缩。由于 K-平均聚类算法中初始聚类中心的选取是随意的，但是初始聚类中心的选取却又直接影响到样本压缩的效果，因此初始聚类中心的选取采用袁方等[11]的初始聚类中心的选择算法。

因子的筛选分为初选和精选，由于消空的过程

中某些重要对流参数的影响容易被覆盖掉,因此不再运用消空因子作为预报因子,而是采用 F-分值[10]进行因子的初选。F-分值思想来源于 Fisher 判别,即实现同类之间差别小,异类之间差别大。具体选择时,给定初选因子的个数 d,把各个对流参数的 F-分值按降序排列,然后找出前 d 个最大值对应的下标,这些下标对应的对流参数就作为初选的因子。

鉴于主成分分析较为常见,因此这里不再介绍。

2　预报模型的建立

2.1　支持向量机算法简介

本文采用支持向量机的两种变形形式——线性规划支持向量机(LP_SVM)和最小二乘支持向量机(LS_SVM)[10]。考虑到便于拓展的需要,算法均由 FORTRAN 语言实现。

支持向量机算法实现的过程中,核函数较为重要,核函数具体实现了空间变换,使数据空间的维数得到改变,解决了非线性问题。可以这样说,核函数为线性时,相应的支持向量机也为线性的,反之,为非线性的。常见的核函数主要有 4 种。

(1) 线性核函数 $K(x,x')=(x\cdot x')$

(2) 多项式核函数 $K(x,x')=(x\cdot x')^d$

(3) 径向基核函数 $K(x,x')=\exp(-\gamma\|x-x'\|^2)$

(4) Sigmoid 核函数 $K(x,x')=\tanh[s(x\cdot x')+c]$

以上核函数中的 d、γ、s 和 c 均为参数,在算法的具体实现中需要借用 K-折交叉确认(K-fold cross-validation)的思想逐个循环进行核函数的确定和最优核参数的选择。

(1) LP_SVM 是基于 C 支持向量机的原问题和对偶问题的变形而得到的,其形式为:

$$\max_{\alpha,b,\beta} -lC-\sum_{i=1}^{l}\alpha_i+C\sum_{i=1}^{l}y_i\Big[\sum_{j=1}^{l}\alpha_j y_j K(x_j,x_i)\Big]+Cb\sum_{i=1}^{l}y_i-C\sum_{i=1}^{l}\beta_i$$

$$st\quad \xi_i=1-y_i\Big(\sum_{j=1}^{l}\alpha_j y_j K(x_j,x_i)+b\Big)+\beta_i \qquad i=1,\cdots,l$$

$$\alpha_i,\beta_i,\xi_i\geqslant 0 \quad i=1,\cdots,l$$

对于线性规划问题这里采用能够克服退化和无限循环的基于 Bland 规则的单纯形算法(Simplex Method)方法[12]解决。

(2) LS_SVM 由 Suykens 在 1999 年提出,其形式:

$$\begin{pmatrix}\boldsymbol{\Omega} & Y\\ Y^T & 0\end{pmatrix}\begin{pmatrix}\boldsymbol{\alpha}\\ b\end{pmatrix}=\begin{pmatrix}\boldsymbol{I}\\ 0\end{pmatrix}$$

其中 $\boldsymbol{\Omega}_{ij}=y_i y_j K(x_i,x_j)+\dfrac{\delta_{ij}}{C}$

由于 $\boldsymbol{\Omega}$ 是对称的半正定矩阵,这里可以利用该矩阵的性质,采用追赶法或平方根方法[13]进行解决。

最小二乘支持向量机将不等式约束转化为等式约束,失去了支持向量的稀疏性,而最小二乘支持向量机的难点也在于支持向量的稀疏性处理。Suyken 等[14]提出了一种简单易行的剪枝方法进行支持向量的稀疏性处理,即按照一定的步长逐步去掉绝对值较小的 Lagrange 乘子对应的支持向量。通过试验发现,具体的剪枝阈值根据最小二乘支持向量机的分类效果而定。

两类支持向量机的决策函数均为:

$$f(x)=\mathrm{sgn}[g(x)]=\mathrm{sgn}\Big\{\sum_{i=1}^{l}\alpha_i^* y_i[x_i\cdot\Phi(\boldsymbol{x})]+b^*\Big\}$$

其中 $x\xrightarrow{\Phi}\boldsymbol{X}=\Phi(\boldsymbol{x})$ 属于引进核函数时做的空间变换,以应对部分线性不可分情况。

2.2　预报模型建立

释用预报中,一般依据天气型建模或者逐月(季)建模。天气分型主要有两个大的方面,一是主观经验分型;二是客观的环流指标分型以及典型场相似分型[15]。在预报对象复杂、天气型较难判断或者判断结果不符合实际的情况下,可以建立逐季(月)的释用预报模式[2]。在本文,采用逐月建模。

这里根据是否采用主成分分析和运用支持向量机模型,共设计 6 个方案。即仅采用 Fisher 线性判别的 RAW_Fisher,主成分-Fisher 判别的 PCA_Fisher,仅采用最小二乘支持向量机的 RAW_LSSVM,主成分-最小二乘支持向量机的 PCA_LSSVM,仅采用线性规划支持向量机的 RAW_LPSVM,主成分-线性规划支持向量机的 PCA_LPSVM。其中这 6 个方案涉及到 3 种模型,即 Fisher、LS_SVM、LP_SVM。

2.3 预报技巧的评价

目前，天气预报的检验和评估方法较多，雷暴属于小概率事件，本文采用适合小概率事件的临界成功指数 *CSI*（风险评分 TS）、Heidke 技巧评分 *HSS* 和 Gilbert 技巧评分 *GSS*，这三种评分方法进行预报效果的评估[16]。

（1）临界成功指数 *CSI*（critical success index）：临界成功指数又称之为风险评分 TS（threat score），反映了事件“出现且报对”时的预报水平。

（2）Heidke 技巧评分 *HSS*（Heidke skill score），*HSS* 表示实际预报准确率比随机预报准确率到底提高了多少，其值在[−1，+1]之间，*HSS*>0 表示预报水平高于随机预报水平，预报完全正确是 *HSS*=1。

（3）Gilbert 技巧评分 *GSS*（Gilbert skill score），*GSS* 实际上是 *CSI* 扣除随机预报正确次数后得到的，因此又称之为技巧临界成功指数，其值在[−1/3，+1]之间，*GSS*>0 表示预报水平高于随机预报水平，预报完全正确是 *GSS*=1。

以上 3 种预报评分方法可以综合使用，当 *HSS* 和 *GSS* 均大于零时，*CSI* 越大则预报方法越好；或者 *CSI* 无太大差异时，*HSS* 和 *GSS* 越大时预报方法越好。

3 个例试验与结果分析

试验站点选为海口站。海口站是雷暴发生的甚高密度区，其一年四季均有雷暴发生，5—8 月是雷暴发生的集中期。海口位于琼州海峡南侧附近，受海陆热力影响更多，其雷暴的日分布也有明显的变化。

通过统计：海口站 2002—2007 年 5—8 月的各月雷暴日天数的分布总体较为平缓；但是雷暴的日分布却有很明显的变化，5—8月各月情况类似，即每天下午到前半夜是雷暴的高发时间段，其他时间段雷暴发生的概率相比显得极小，如图 1。

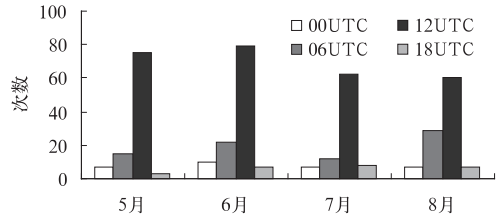

图 1 海口站 5—8 月雷暴的日分布

Fig. 1 The daily distribution of thunderstorm from May to August at Haikou Station

在采用预报模型预报之前，采用了 1 个消空因子进行了消空预报。因此对于“消空成功”或者“过消空”的结果都放在预报方程的预报结果中进行统计和最后预报方法的评估。

首先分析不同方案的总体预报效果（表 1）。由结果可知：

（1）当仅进行因子的初选时，RAW_LSSVM、RAW_LPSVM、RAW_Fisher 这 3 种方案的 *HSS* 和 *GSS* 评分均大于零，雷暴的预报结果好于随机预报，这说明预报流程的设计是有效果的，整个预报流程是合理的；3 种方案中，最小二乘支持向量机的预报效果最好，*CSI* 评分约为 0.23，*HSS* 和 *GSS* 评分均大于其他两种方案，两种形式支持向量机的预报效果均要好于线性的 Fisher 判别，因而支持向量机的预报效果要好于线性的 Fisher 判别预报模型。

（2）因子的精选涉及到主成分分析时，PCA_LSSVM、PCA_LPSVM、PCA_Fisher 3 种方案的 *HSS*、*GSS* 和 *CSI* 评分相较于未进行因子精选的预报效果有明显的提高，*CSI* 评分最低的 Fisher 也达到了 0.28，说明在释用预报中减少因子数目、进一步综合因子和进行因子的精选是很有必要的；3 种方案中，仍然是最小二乘支持向量机的预报效果最好，其 *CSI* 评分达到了近 0.40，*HSS* 和 *GSS* 评分均大于其他两种方案，两种形式支持向量机预报

表 1 海口站不同方案的总体预报效果

Table 1 The forecasting results of different project at Haikou Station

方案	出且报对/次	漏报/次	空报/次	不出且报对/次	总次数/次	*CSI*	*HSS*	*GSS*
RAW_LSSVM	54	49	133	379	615	0.2288	0.1995	0.1108
RAW_LPSVM	39	64	120	392	615	0.1749	0.1185	0.0630
RAW_Fisher	42	61	156	356	615	0.1622	0.0753	0.0391
PCA_LSSVM	57	27	63	99	246	0.3878	0.2626	0.1511
PCA_LPSVM	40	44	41	121	246	0.3200	0.2250	0.1268
PCA_Fisher	38	45	55	108	246	0.2754	0.1170	0.0621

CSI 评分均超过了 0.30,进一步说明了支持向量机方法较线性判别方法具有一定的优越性。虽然预报评分有了一定提高,但是这种提高是建立在较大的样本规模基础之上的,即预报时次是那些样本达到一定规模的时次(自行设计的程序中规定样本个数大于 50 时才进行因子的精选),所以预报评分在这里仍不能完全说明预报方法的绝对优越性,但是应当注意的是对于样本较少时次的预报应该不要仅仅局限于定量的方法,可以考虑对流参数区域叠套等方法的运用。

由以上分析发现,对于海口站,最小二乘支持向量机是最好的预报形式,进行主成分分析后预报效果有明显提高。因此下面按照样本多时采用 PCA_LSSVM,样本少时 RAW_LSSVM 这种组合方式统计各个时次和各月的预报结果。

先分析各个时次的预报结果,由表 2 可知:12 和 36 h 预报效果最好,这两个时次都对应海口当天下午,是海口站雷暴高发时间段;30 h 的 *CSI* 评分有 0.15,这个时间对应海口当天的中午前后,雷暴也经常发生;18 和 24 h 的预报效果虽然 *HSS* 和 *GSS* 评分均大于零,但是总体都较差,这也与海口站夜间到第二天鲜有雷暴的时间段相对应;此外 12 和 36 h 两个预报时刻其实均对应于 12 UTC,但是发现 36 h 的预报 *CSI* 评分要高于 12 h 的,这说明模式的输出存在形势场的过快或者过慢,某些物理场也随着预报时效的增加变得过大或者过小。

表 2　海口站各时次预报结果(LS_SVM)

Table 2　The 6-h interval forecasting results at Haikou Station (LS_SVM)

时次/h	出且报对/次	漏报/次	空报/次	不出且报对/次	总次数/次	*CSI*	*HSS*	*GSS*
12	27	15	28	53	123	0.3857	0.2766	0.1605
18	1	4	20	98	123	0.0400	0.0120	0.0061
24	2	1	18	102	123	0.0952	0.1373	0.0737
30	6	5	29	83	123	0.1500	0.1444	0.0778
36	30	12	35	46	123	0.3896	0.2493	0.1424

由这些分析可以发现,对于雷暴的模式产品释用预报,如果一直运用一种方法和均匀 6 h 的间隔进行预报,不考虑当地的雷暴的日分布规律,预报效果并不好,这就要求某些时间段可以进行加密预报,某些时间段可以放稀疏些;此外,为了适应样本,预报方法也可以采取灵活的方式,比如一些定性的、非方程形式的方法。

接着分析各月的预报结果,如表 3。由该表可知:6 月的预报效果最好,7 月预报效果较差;由于逐月建模是分别将各月作为建立不同气候背景的依据,考虑到海口所处的地理位置,其发生雷暴时的环流形势在 5—6 月都是分别有所侧重的,但是 7 和 8 月的环流形势经常被副热带高压、台风和低槽等天气型交替影响,因此预报效果也受到了一定的影响。

表 3　海口站各月预报结果(LS_SVM)

Table 3　The monthly forecasting results at Haikou Station (LS_SVM)

月份	出且报对/次	漏报/次	空报/次	不出且报对/次	总次数/次	*CSI*	*HSS*	*GSS*
5	17	6	33	99	155	0.3036	0.3295	0.1972
6	21	7	32	90	150	0.3500	0.3629	0.2217
7	12	13	27	103	155	0.2308	0.2221	0.1249
8	16	11	38	90	155	0.2462	0.2121	0.1186

4　结　论

论文利用 2002—2006 年 AREM 模式产品和常规观测报文资料,逐月建立了支持向量机的单站雷暴释用预报模型,并针对海口站 2007 年 5—8 月进行了具体的预报,结果表明:

(1) 在不同的预报方案中,支持向量机较线性 Fisher 判别的预报效果好,其中最小二乘支持向量机的预报效果最好。由此可见,支持向量机结合 AREM 模式产品进行释用预报是合适的,用支持向量机方法建立预报方程较线性 Fisher 判别效果优。

(2) 在主成分分析前后,支持向量机和 Fisher 判别的预报效果均有明显改变,这些说明:对于雷暴这种较为复杂的天气现象的预报,通过一定方法综合和精简其预报因子,会对最终的预报结果起到积

极的作用。

（3）通过对各个时次和各月的预报结果分析发现：雷暴的预报效果在雷暴活跃期较好，而在雷暴的沉寂期却相对较差，这些结果一方面说明了预报方程的预报效果与样本的规模有关，另一方面也说明了对雷暴沉寂期缺乏细致的研究，规律难觅，预报仍较难。

参考文献

[1] 孔玉寿，章东华. 现代天气预报技术(第二版)[M]. 北京：气象出版社，2005.

[2] 谌志刚，王婷，汪瑛，等. 广东省后汛期强对流天气潜势预报方法研究[J]. 气象，2011，37(8)：936-942.

[3] 王新敏，张霞，徐文明，等. T213/T639 数值产品在河南省雷电潜势预报中的释用[J]. 气象，2011，37(5)：576-582.

[4] 陈翔，彭丽霞，高文亮，等. 洪泽湖地区强雷暴天气气候特征与雷达回波分析[J]. 气象，2011，37(9)：1118-1125.

[5] 冯汉中，陈永义. 处理非线性分类和回归问题的一种新方法——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)：355-365.

[6] 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[J]. 气象科技，2009，37(2)：145-148.

[7] 贺皓，罗慧. 基于支持向量机模式识别的大雾预报方法[J]. 气象科技，2009，37(2)：149-151.

[8] 宇如聪，薛纪善，徐幼平，等. AREMS 中尺度暴雨数值预报模式系统[M]. 北京：气象出版社，2004.

[9] Haklander A J, Delden A V. Thunderstorm predictors and their forecast skill for the Netherlands[J]. Atmospheric Research, 2003, 67-68, 273-299.

[10] 邓乃杨，田英杰. 支持向量机——理论、算法与拓展[M]. 北京：科学出版社，2009.

[11] 袁方，孟增辉，于戈. 对 K-means 聚类算法的改进[J]. 计算机工程与应用，2004，(36)：177-178.

[12] Kuenzi H P, Tzschach H G, Zehnder C A. Numerical Methods of Mathematical Optimization[M]. New York: Academic Press, 1971.

[13] 李庆扬，王能超，易大义. 数值分析(第四版)[M]. 北京：清华大学出版社，2001.

[14] Suykens J A K, Lukas L, Vandwealle J. Sparse least squares support vecter machines for adaptive communication channel equalization[J]. International Journal of Applied Science and Engineering, 2005, 11(3): 51-59.

[15] 贾丽伟，李维京，陈德亮. 东北地区大气环流型与哈尔滨气候关系的初步研究[J]. 气象学报，2006，64(2)：236-245.

[16] 罗阳，赵伟，翟景秋. 两类天气预报评分问题研究及一种新评分方法[J]. 应用气象学报，2009，20(2)：129-135.