SVM方法在降水预报中的应用及改进

熊秋芬[1] 曾晓青[2]

(1. 中国气象局培训中心,北京 100081; 2. 兰州大学大气科学学院)

提 要:以T213数值模式输出产品为基础,结合常规观测的降水资料,利用SVM方法,进行了大量多因子的随机交叉验证,从而选出最优参数,建立了全国72个站点的降水预报模型,并用独立的样本对预报模型进行了检验。再通过计算正、负样本的贴近度来分析预报因子,实现了预报因子的筛选和降水预报模型的改进;检验结果表明:改进后的降水模型的预报结果优于改进前的。实时业务试运行的结果也显示SVM模型的降水预报效果好于T213模式直接输出的降水预报。

关键词:SVM方法 降水预报 贴近度 因子

Application and Improvement of SVM Method in Precipitation Forecast

Xiong Qiufen[1] Zeng Xiaoqing[2]

(1. CMA Training Centre, Beijing 100081; 2. Atmospheric Science Academy of Lanzhou University)

Abstract: Based on T213 NWP(Numerical Weather Prediction)model outputs and precipitation observations, cross-validation is performed with random samples to find the samples with best predictors and optimal parameters. The forecast models of precipitation are established at 72 meteorological stations in China by the SVM (Support Vector Machine) statistical method. The models are verified with independent samples. The predictors are selected and the precipitation forecast models are improved by pressing close degree. Forecast experiments show that the improved models are better. The precipitation forecasted by SVM models is superior to the precipitation of T213 DMO (direct model output) in real-time experiments.

Key Words: support vector machine(SVM) method precipitation forecast pressing close degree predictor

资助项目:"中国气象局数值模式创新基地"开放课题(2007)

收稿日期:2008年1月30日; 修定稿日期:2008年10月30日

引　言

降水是各种尺度的天气系统共同作用的结果,其形成机制非常复杂,具有非线性的特点,MOS、卡尔曼滤波、神经元网络等方法被广泛应用到降水预报中,以提高降水预报准确率。近年来,能处理非线性问题的 SVM (Support Vector Machine)方法[1-2]被引入非线性特征十分明显的大气科学领域,已取得了初步的成果[3-8]。但已有的关于 SVM 方法在降水预报的应用中,主要是该方法在某一地区的局部应用[3,7-8],能否推广应用到全国、如何推广应用到全国,需要进一步的探讨。

本文基于 T213 数值模式预报产品,在全国范围内选取了 72 个站点的降水作为预报对象,利用 SVM 方法,进行大量交叉验证,选出最优的参数,建立降水预报模型,并用独立的样本对模型进行了检验;再通过分析样本的贴近度来分析预报因子,改进预报模型,以提高模型的预报效果。

1　资料选取及处理

在全国选取 72 个气象站的日降水量作为预报对象,并读取这 72 个站 2003—2005 年 4—9 月以及 2006 年 4 月 1 日至 7 月 31 日逐日 08—08 时观测的降水量。选取了对应段内 T213 数值预报产品作为主要的预报因子,将 T213 的基本要素及其通过动力诊断得出的反映降水的扩充物理量,用双线形插值的方法插值到对应的 72 个站点上,建立起所需要的站点因子库;再通过计算相关系数,在不同层次、不同时次的因子中选出一批与实况降水量相关系数较大的因子,然后按相关系数由大到小的顺序排列,选取 100 个左右的预报因子,这些因子中包含有从 00、12、24 到 48 小时预报时效的因子。

这样就形成了 72 站 2003—2005 年 4—9 月、2006 年 4 月 1 日至 7 月 31 日共 600 个左右的历史学习训练样本集。

另外以 2006 年 8 月 1 日至 9 月 30 日共 60 个样本作为独立检验样本集。

2　降水预报模型的建立及模型的检验

SVM 方法的基本思想[1-2]是升维和线性化,通过非线性映射(核函数),把样本空间映射到一个高维乃至于无穷维的特征空间,在特征空间中,应用线性学习机的方法解决样本空间中的高度非线性问题。在整个求解过程中不需要知道非线性映射的显式表达式,而是通过支持向量(关键样本)来表达预报因子与预报对象的关系。

这里用 SVM 两类分类方法[2],在建模之前,先对预报对象进行分类:

由于西北地区降水较少,因此将西北 9 站(乌鲁木齐、克拉玛依、酒泉、民勤、兰州、呼和浩特、盐池、太原、银川)日降水量≥1mm 标定为正样本(+1 类)、日降水量<1mm 标定为负样本(−1 类);其余站 63 站日降水量≥10mm 标定为正样本(+1 类)、日降水量<10mm 标定为负样本(−1 类)。再对每个站全部样本的每一个因子按公式(1)分别做归一化处理,使每个因子的数据在[0,1]之间,这样避免了预报因子之间量级的差异。

$$a = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

式(1)中 $x_{\max}$ 和 $x_{\min}$ 分别为因子的最大和最小值。

选取最常用的径向基核函数,即 $K(x, y) = \mathrm{e}^{-g * \sum_{i=1}^{n} (x_i - y_i)^2}$,通过调整核参数 g 和惩罚系数 C 的值,进行大量随机交叉验证,分析比较所建模型 T_s 评分的高低,从而选择出最优模型对应的参数 g 和 C。

从全部 600 个样本中随机抽取的 10% 样本作为检验样本，其余样本作为建模样本，C、g 给定的初值分别为 100 和 0.005，再按一定的步长递增（分别为 5 和 0.005），对每个站进行 1000 次随机交叉验证，从中选出最高 T_s 评分值对应的参数 C、g。然后以 2003—2005 年 4—9 月、2006 年 4 月 1 日至 7 月 31 日共 600 个左右的训练样本集分别建立各站的降水预报模型，再用 60 个独立的样本来检验预报模型的预报能力，即预报 2006 年 8 月 1 日至 9 月 30 日期间逐日的降水，部分站预报评分结果见表 1；T213 模式的直接输出的降水预报（简记 DMO）评分检验也在表 1 中。

表 1 部分站 SVM 预报和 T213 DMO 的 T_S 评分对比

站名	SVM 预报	T213 DMO
民勤	0.308	0.207
兰州	0.462	0.286
太原	0.444	0.394
哈尔滨	0.500	0.333
漠河	0.571	0.333
呼和浩特	0.375	0.320
平凉	0.367	0.250
本溪	0.400	0.364
北京	0.500	0.067
济南	0.500	0.286
泰山	0.440	0.308
昌都	0.333	0
略阳	0.333	0.077
汉中	0.545	0.444
安康	0.600	0.364
遵义	0.333	0
安庆	0.500	0
河池	0.429	0.400
漳州	0.556	0.217
厦门	0.375	0.176
广州	0.316	0.286
海口	0.500	0.294
三亚	0.333	0.231

独立样本检验的结果（表 1）表明，SVM 建立的模型对降水有较好的预报能力，且比 T213 模式直接输出的降水预报的 T_s 评分高，这些站的降水预报模型可以投入业务试运行。

3 贴近度分析及模型的改进

贴近度是在训练样本数据集中引入的一种相似性度量函数，它刻画两个训练样本之间的相似或贴近的程度。它是分析训练数据集的一种方法。

设样本向量为 $\boldsymbol{x}_i=(x_{i1},\cdots,x_{in})$ 和 $\boldsymbol{x}_j=(x_{j1},\cdots,x_{jn})$，且经过了归一化处理，则它们在样本空间的贴近度为：

$$N(x_i,x_j)=\frac{2(x_i\cdot x_j)}{(x_i\cdot x_i)+(x_j\cdot x_j)} \quad (2)$$

这里通过经向基核函数 $K(x,y)=e^{-g*\sum_{i=1}^{n}(x_i-y_i)^2}$ 的非性线映射，在特征空间中 x_i 和 x_j 分别变为 $\psi(x_i)$ 和 $\psi(x_j)$，则它们在特征空间的贴近度为：

$$N(x_i,x_j)=\frac{2(\psi(x_i)\cdot\psi(x_j))}{(\psi(x_i)\cdot\psi(x_i))+(\psi(x_j)\cdot\psi(x_j))} \quad (3)$$

式(3)表示在特征空间中两个训练样本的相似或贴近的程度，完全相同的两个样本的贴近度为 1，区别最大的两个样本的贴近度为 0，一般两个训练样本的贴近度介于 0 和 1 之间。

由于 72 站中正、负样本数不同，组成样本的因子也各不相同，所以经计算所得到的样本间的贴近度也不同，表 2 给出了各站贴近度的变化范围（在最小值和最大值之间）。表 1 与表 2 对比分析发现，表 1 中 SVM 模型预报 T_s 评分较高的站对应着表 2 中该站同类样本（正样本与正样本、负样本与负样本）之间的样本的贴近度较大、正负样本间的贴近度较小。而 SVM 模型预报 T_s 评分低（表 3 中改进前）的站同类样本之间的样本的贴近度小、正负样本间的贴近度大。

表 2　72 站样本贴近度的变化范围

站名	正样本与正样本	负样本与负样本	正样本与负样本	站名	正样本与正样本	负样本与负样本	正样本与负样本
乌鲁木齐	0.371－0.970	0.372－0.993	0.275－0.976	西宁	0.675－0.962	0.435－0.988	0.426－0.979
克拉玛依	0.654－0.972	0.514－0.987	0.481－0.983	延俺	0.398－0.967	0.321－0.978	0.253－0.961
民勤	0.494－0.951	0.414－0.987	0.415－0.979	沈阳	0.541－0.963	0.433－0.982	0.366－0.973
酒泉	0.337－0.962	0.375－0.988	0.406－0.977	天津	0.542－0.949	0.357－0.979	0.306－0.962
兰州	0.597－0.971	0.522－0.988	0.383－0.975	大连	0.547－0.960	0.483－0.985	0.348－0.971
呼和浩特	0.519－0.967	0.571－0.991	0.326－0.976	青岛	0.662－0.974	0.545－0.990	0.521－0.981
盐池	0.526－0.962	0.384－0.984	0.356－0.976	丽江	0.691－0.985	0.540－0.990	0.518－0.990
太原	0.527－0.956	0.491－0.982	0.346－0.971	郑州	0.324－0.981	0.393－0.987	0.426－0.978
银川	0.357－0.950	0.325－0.980	0.237－0.961	宜昌	0.559－0.988	0.406－0.992	0.345－0.975
漠河	0.546－0.962	0.558－0.981	0.389－0.974	武汉	0.267－0.942	0.243－0.981	0.119－0.969
哈尔滨	0.711－0.981	0.501－0.989	0.475－0.989	重庆	0.601－0.978	0.502－0.990	0.436－0.988
平凉	0.508－0.959	0.577－0.987	0.431－0.978	岳阳	0.590－0.981	0.486－0.985	0.432－0.963
本溪	0.571－0.978	0.518－0.987	0.425－0.977	长沙	0.137－0.942	0.204－0.987	0.103－0.964
北京	0.734－0.981	0.633－0.989	0.562－0.989	郴州	0.147－0.933	0.179－0.970	0.076－0.968
济南	0.502－0.960	0.549－0.981	0.331－0.971	赣州	0.812－0.995	0.839－0.998	0.775－0.996
泰山	0.514－0.977	0.519－0.984	0.465－0.976	徐州	0.309－0.978	0.298－0.985	0.321－0.979
昌都	0.439－0.945	0.520－0.977	0.233－0.974	蚌埠	0.471－0.966	0.434－0.986	0.329－0.982
略阳	0.552－0.975	0.514－0.989	0.309－0.979	六安	0.664－0.989	0.650－0.994	0.564－0.991
汉中	0.697－0.981	0.602－0.989	0.421－0.987	合肥	0.308－0.967	0.396－0.968	0.082－0.967
安康	0.521－0.975	0.554－0.989	0.283－0.986	常州	0.267－0.939	0.096－0.967	0.082－0.967
遵义	0.613－0.979	0.465－0.990	0.363－0.980	上海	0.107－0.912	0.123－0.976	0.038－0.957
安庆	0.563－0.976	0.571－0.992	0.349－0.985	黄山	0.378－0.981	0.426－0.993	0.263－0.980
河池	0.592－0.985	0.575－0.988	0.399－0.985	庐山	0.439－0.968	0.408－0.986	0.319－0.974
漳州	0.495－0.974	0.519－0.989	0.202－0.979	杭州	0.345－0.966	0.353－0.990	0.230－0.977
厦门	0.518－0.974	0.556－0.983	0.079－0.973	南昌	0.292－0.959	0.218－0.988	0.177－0.986
海口	0.562－0.977	0.520－0.987	0.221－0.984	韶关	0.448－0.975	0.483－0.989	0.399－0.986
三亚	0.557－0.976	0.591－0.988	0.315－0.984	梧州	0.147－0.933	0.179－0.970	0.076－0.967
石家庄	0.359－0.963	0.342－0.985	0.499－0.975	广州	0.689－0.993	0.740－0.994	0.430－0.992
长春	0.401－0.977	0.322－0.991	0.571－0.989	深圳	0.455－0.977	0.458－0.989	0.380－0.985
拉萨	0.361－0.960	0.241－0.982	0.413－0.981	湛江	0.342－0.984	0.367－0.987	0.411－0.986
贵阳	0.331－0.984	0.454－0.986	0.459－0.979	西沙	0.318－0.974	0.405－0.988	0.223－0.984
井冈山	0.413－0.979	0.373－0.987	0.317－0.986	珊瑚岛	0.029－0.958	0.215－0.968	0.100－0.948
南京	0.248－0.944	0.173－0.974	0.140－0.959	桂林	0.368－0.984	0.470－0.988	0.436－0.985
福州	0.355－0.969	0.446－0.991	0.302－0.973	汕头	0.029－0.958	0.247－0.982	0.018－0.963
九仙山	0.218－0.972	0.266－0.981	0.089－0.978	北海	0.366－0.983	0.406－0.992	0.444－0.858
南宁	0.192－0.968	0.357－0.983	0.124－0.979	武夷山	0.380－0.987	0.337－0.988	0.251－0.982

这里重点对比分析 SVM 模型在独立检验时 T_s 评分较低、同类样本之间的样本的贴近度小、正负样本间的贴近度大的站,从而实现预报因子的筛选。

SVM 方法要求同类样本的贴近度尽可能大,不同类样本的贴近度尽可能小。以下是对 20 站中同类样本贴近度小、不同类样本贴近度较大的样本中各因子分析的结果,这些因子在样本中应该剔出。它们大致可以分为下面 6 类:

(1) 一些较复杂的热力、动力因子。如螺旋度、Q 矢量、动力综合因子、锋生函数、位涡及湿位涡有关的量、风的垂直切变、风向、K 指数、SI 指数、Ky 指数、温度的指数、比湿的指数、U、V 的指数等。

(2) 层次高的因子。如 200hPa 相当位

温、假相当位温的垂直切变、50hPa 的 U、V，150hPa 的温度及其梯度。

(3) 近地层因子(特别是热带地区)。如2米的温度、湿度，海平面气压、地面气压及变压，10米的风，1000hPa 的变温、变高等。

(4) 高低层差值和累积量。如200hPa与850hPa 的高度差，400、300、200hPa 垂直速度的累积量，700、600、500hPa 温度的累计量，地面到600hPa 水汽通量及水汽通量散度等。

(5) 同一层次、同一要素、同一性质的因子太多。如700hPa 温度的平方、立方和 e 指数，相对湿度、比湿的平方、立方和 e 指数被同时选作因子等。

(6)海拔高度以下的因子，如高山站海平面气压、地面气压、850hPa 的风等。

用贴近度计算分析后，去掉一些因子，各站筛选后的预报因子个数各不相同(表略)，用筛选后的因子重新组成新的训练建模样本集，再进行1000次随机交叉验证，得到最优的参数 C、g，从而得到改进后的预报模型。改进后的预报模型对2006年8月1日至9月30日逐日降水的预报效果的检验见表3，表3也给出了同一时段内 SVM 模型改进前降水预报、T213模式直接输出的降水预报(DMO)结果的检验。

表3 20站 SVM 模型改进后、改进前预报和 T213 DMO 的 T_s 评分对比

站名	SVM 模型改进后	SVM 模型改进前	T213 DMO
乌鲁木齐	0.222	0.167	0.167
银川	0.625	0.200	0.250
石家庄	0.714	0.250	0.500
长春	0.375	0.200	0.333
拉萨	0.250	0	0
贵阳	0.333	0.222	0.182
井冈山	0.200	0.143	0.154
南京	0.250	0.200	0.111
福州	0.400	0.125	0.227
九仙山	0.357	0.214	0.167
南宁	0.286	0.200	0.273
北海	0.267	0.154	0.250
汕头	0.333	0.214	0.250
桂林	0.286	0.143	0.143
酒泉	0.200	0.200	0.143
郑州	0.143	0.125	0.250
徐州	0.200	0	0.111
杭州	0.167	0	0.167
武夷山	0.143	0.143	0.133
湛江	0.167	0.090	0.111

由表3可知，改进后模型的 T_s 评分基本上比改进前的 T_s 评分和 T213 模式直接输出的降水预报 T_s 评分高，表明改进后的模型对降水有一定的预报能力，但仍有少数站的预报效果较差。

4 实时预报试验效果的检验

2007年8月1日开始，SVM 方法建立的降水预报模型在国家气象中心投入业务试运行，对于 T_s 评分>0.25 的站，用原来的预报因子(100个左右)组成的样本集建模；对于 T_s 评分≤0.25 的站，则用贴近度分析后筛选的因子组成的样本集建模。为了增强所建模型的稳定性，实时业务运行时，增加了建模样本的长度，这里使用了2003—2005年4—9月、2006年4月1日至7月31日共600个左右的原训练样本集和2006年8月1日至9月30日独立检验的60个的样本，即建模训练样本的长度为2003—2006年4—9月660个样本的历史资料。共有72个站的降水预报模型在2007年7月31日至9月30日进行了业务试运行，部分站预报检验结果见表4，表4也给出了同一时段内 T213 模式直接的降水预报(DMO)结果的检验。

实时业务试运行的结果(表4)也表明 SVM 建立的模型对降水具有预报能力，其 T_s 评分和预报准确率也基本上高于 T213 模式直接输出的降水预报，相对而言，SVM 方法的漏报较多，而 T213 模式的空报较多。

表 4　预报试验中部分站 SVM/T213 DMO 的检验结果

站名	正样本 T_s 评分	正样本正确次数	正样本空报次数	正样本漏报次数	负样本正确次数	正样本个数	全样本准确率
乌鲁木齐	0.333/0.286	3/6	2/14	4/1	52/40	7/7	0.902/0.754
兰州	0.429/0.293	6/12	2/29	6/0	47/20	12/12	0.869/0.525
盐池	0.412/0.407	7/11	4/14	6/2	44/34	13/13	0.836/0.738
太原	0.412/0.381	14/8	18/5	2/8	27/40	16/16	0.672/0.787
长春	0.600/0.375	3/3	2/5	0/0	56/53	3/3	0.967/0.918
沈阳	0.308/0.308	4/4	7/7	2/2	48/48	6/6	0.852/0.852
北京	0.250/0.125	1/1	0/4	3/3	57/53	4/4	0.951/0.885
大连	0.364/0.353	4/6	2/8	5/3	50/44	9/9	0.885/0.820
泰山	0.421/0.211	8/4	9/9	2/6	42/42	10/10	0.820/0.754
宜昌	0.231/0.190	3/4	3/11	7/6	48/40	10/10	0.836/0.721
重庆	0.556/0.400	5/4	4/5	0/1	52/51	5/5	0.934/0.902
郴州	0.308/0.231	4/3	6/6	3/4	48/48	7/7	0.852/0.836
徐州	0.250/0.190	3/4	3/12	6/12	49/40	9/9	0.852/0.721
庐山	0.429/0.333	6/6	1/5	7/7	47/43	13/13	0.869/0.803
福州	0.200/0.192	3/5	8/19	4/2	46/35	7/7	0.803/0.656
九仙山	0.500/0.515	13/17	5/12	8/4	35/28	21/21	0.787/0.738
河池	0.460/0.300	7/6	7/9	4/5	46/41	11/11	0.869/0.770
漳州	0.273/0.235	6/8	8/20	8/6	39/27	14/14	0.738/0.574
厦门	0.313/0.258	5/8	8/23	3/0	45/30	8/8	0.820/0.623
汕头	0.471/0.385	8/10	4/13	5/3	44/35	13/13	0.852/0.738
北海	0.308/0.273	4/6	1/10	8/6	48/39	12/12	0.852/0.738
三亚	0.286/0.200	4/6	3/19	7/5	47/31	11/11	0.836/0.607
珊瑚岛	0.357/0.304	5/7	4/13	5/3	47/38	10/11	0.852/0.738

由表 1、表 3 及表 4 可知,SVM 对降水有一定的预报能力,特别是东北和华南地区,即对于建模、独立检验、预报试验时 T_s 评分较高的站,预报模型可以投入业务试运行。

5　讨论

对 2006—2007 年 8—9 月预报检验可知,SVM 建立的模型在在东北和华南预报效果较好。但仍存在以下问题:

(1)从建模和预报检验的情况来看,有些地区(如西部地区、沿长江流域等)T_s 评分一直很低,如何筛选因子,提高预报准确率需要进一步探讨。当然也可能与 T213 模式本身在这些地区预报误差较大有关。

(2)模型稳定性问题。2006 年预报检验时 T_s 评分较高,2007 年预报试验时却不高,表明模型稳定性差,需要增加更多的建模样本资料。

参考文献

[1] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York:Springer Verlag, 2000.

[2] 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):345-354.

[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[J]. 应用气象学报,2004,15(3):355-365.

[4] 李智才,马文瑞,李素敏,等. 支持向量机在短期气候预测中的应用[J]. 气象,2006,32(5):57-61.

[5] 熊秋芬,顾永刚,王丽. 支持向量机分类方法在天空云量预报中的应用[J]. 气象,2007, 33(5):20-26.

[6] 冯汉中、陈永义. 支持向量机回归方法在实时业务预报中的应用[J]. 气象,2005,31(1):41-44.

[7] Qiufen XIONG, Jie GAO, Huanzhu LIU, et al. Physical Analysis of Precipitation Factors Based on SVM Method[C]. The 9th Japan-China Symposium on Statistics. Sapporo, Japan. 2007.

[8] 王建生,熊秋芬. 支持向量机方法在单站降水预报中的应用探讨[J]. 暴雨灾害,2007,26 (2):159-162.