第23卷 第4期
2005年12月

干 旱 气 象
ARID METEOROLOGY

Vol. 23 NO. 4
Dec,2005

文章编号:1006-7639(2005)-04-0041-05

支持向量机方法在冰雹预报中的应用

吴爱敏,郭江勇,张洪芬,路亚奇

(甘肃省庆阳市气象局,甘肃 庆阳 745000)

摘 要:支持向量机(Support Vector Machines 简称 SVM)方法是近年发展起来的一种新的统计学习理论方法,本文通过对这一方法的学习,总结陇东主要降雹的环流形势特点,利用这一新方法对冰雹分类预报进行了探讨,经检验,效果较好。并与传统的天气分型后制作预报模式进行了比较,验证了SVM方法不需要进行天气分型,这样总样本数多,建立的预报模型效果好。这为基层台站制作天气预报模式提供了一种新方法和新思路。

关键词:支持向量机;冰雹;环流特征;预报模式

中图分类号:P456.8 文献标识码:A

引 言

随着计算机技术的飞速发展,处理各种气象资料信息并将其及时分析应用于天气预报业务成为可能。如何从多种大气探测资料、数值预报模式等这些海量信息中获取可用于预报的关键信息,通过机器学习就可能有效地解决这一问题。当我们面对数据而又缺乏理论模型时,统计分析方法是最先采用的方法。然而传统的统计方法只有在样本数量趋于无穷大时才能有理论上的保证。而在实际应用中样本数目通常都是有限的,甚至是小样本,基于大数定律的传统统计方法对此难以取得理想的效果。多元回归方法、卡尔曼滤波方法[1~2]等是建立在线性相关基础上的统计方法,把它们用于具有非线性特征的气象要素或天气现象(比如降水)的预报上有局限性,人工神经元模型方法通过给定数目的观测所得的解往往是局部最优,在气象上的应用也存在不足,近年来发展起来的支持向量机(Support Vector Machines 简称 SVM)方法[3]为解决这一类问题提供了比较有效的手段。

V. N. Vapnik 等人提出的这种统计学习理论是一种专门的小样本理论,数学推导严密,理论基础坚实,它避免了人工神经网络等方法的网络结构难于确定的过学习和欠学习以及局部极小等问题,被认为是目前针对小样本的分类、回归等问题的最佳理论,是在统计学习理论的基础上发展起来的新一代学习算法,陈永义[4]、冯汉中[5]等对 SVM 方法的原理和在气象预报领域中的应用进行了一些尝试探讨,结果表明,SVM 方法能用于具有显著非线性特征的气象预测预报。冰雹作为一种小天气概率事件,具有离散性,本文应用 SVM 分类方法建立冰雹预报模型,探讨其在强对流天气预报中的应用。

1 支持向量机(SVM)分类方法简介

SVM 方法的基本思路是:定义最优线性超平面,并把寻找最优线性超平面的算法归结为求解一个凸规划问题。进而基于 Mercer 核展开定理,通过非线性映射 φ,把样本空间映射到一个高维乃至于无穷维特征空间(Hilbert 空间),使在特征空间中的可以应用线性学习机的方法,解决样本空间中的高度非线性分类和回归等问题,简单地说就是升维和线性化。降维(即把样本空间向低维空间做投影)是人们处理复杂问题常用的简化方法之一,这样做可以降低计算的复杂性。而升维,是把样本空间向高维空间做映射,一般只会增加计算的复杂性,甚至会引起"维数灾",因而人们很少问津。但是作为分类、回归等问题来说,很可能在低维样本空间无法线性处理的样本集,在高维特征空间却可以通过一个线性超平面实现线性划分(或回归)。

SVM 方法的核心是支持向量。SVM 建模方法,

收稿日期:2005-03-07;改回日期:2005-07-28

基金项目:科技部2002年社会公益类项目"西北地区人工消雹防雹技术"(2002DIB10046)资助

作者简介:吴爱敏(1967-),女,河南温县人,高级工程师,主要从事天气预报应用研究. E-mail:qxjwam@sohu.com

其本质就是通过对各种典型空间(支持向量)的充分表述来描述因子群与预报对象之间的关系,是一种基于事实的转导式推理。

SVM 分类即模式识别,就是依据有限的观测数据(训练样本)来寻求蕴涵着的分类关系(建立分类模型),进而用求得的分类模型对未来数据(预报数据)进行预报。即根据给定的样本数据集:

$(x_{i1},x_{i2},\cdots,x_{in};y_i)$,$x_{ij} \in \mathbf{R}$, $y_i \in \mathbf{N}$

其中 $x_i = (x_{i1},x_{i2},\cdots,x_{in})$ 为待识别对象的特征因子数值,为 N 维向量, y_i 为决策结果值。

$y_i \hat{I}$ {-1,1} (二类划分) 或

$y_i \hat{I}$ {1,2,……,k} (多类划分)

求函数关系 $y_i = f(x_i,\alpha)$ (α 为参数向量),使其对于样本数据集符合率最大,且具有好的推广能力。

对于冰雹天气来说,只是有和无的问题,是二类划分,就是在规定的函数类中寻找一个合适的实函数 $y = f(x)$,使得对任何 x_i,当 $f(x_i) \geqslant 0$ 时把 x_i 归入有冰雹,当 $f(x_i) < 0$ 时把 x_i 归入无,总决策函数可以表示为:

$$y = M(x) = \mathrm{Sgn}(f(x))$$

其中的 Sgn(·)为符号函数,它当自变量非负时取值 +1,当自变量为负时取值 -1。

2 冰雹天气的环流背景分析

制作冰雹预报模型,首先对环流背景进行分析。冰雹天气由中小尺度局地强对流系统产生,它的发生、发展仍然受到大气环流和天气尺度影响系统的制约,与高空和中空的干湿、热冷情况和大气运动有密切关系,同时又受地形条件影响。甘肃陇东区域性降雹多数出现在长波槽的高空西北气流冷平流中的大尺度下沉运动区内[6],这种天气过程较强,往往连续几天发生雹灾,有时在盛夏偏北气流中也出现冰雹,过程较弱;其次是偏西和西南气流为主的天气型里,多为一扫而过,且伴有大风、大暴雨,可发生综合性灾害,其中西北气流型概括了 80% 以上的冰雹天气,下面的分析和建模讨论主要以该天气型为主。

西北气流型是夏半年(4~10 月)降雹的主要天气形势。依陇东短期冰雹预报经验,表现为西高东低,春季(4~5 月),500 hPa 中纬度(30°~65°N)为两槽一脊,我国东北黑龙江及其以南为冷性深槽,新西伯利亚为浅槽区,新疆—贝加尔湖西部地区为高压脊区(图略),沿新疆脊前不断有冷空气南下,在河西及河套上游有明显的冷温度槽形成并加强东移,陇东与河西及河套北部高空温度之差达到 10~15 ℃,斜压性较强,沿低槽底部甩下的冷平流影响陇东,产生冰雹。

夏季(6~8 月)(图 1),高纬及极地南下冷空气明显减弱,蒙古—我国东北的低槽向西向南扩充,高原成为热源,为一相对低值区,西北地区处于脊前西北气流中。初夏(6 月)降雹的环流形势说明西伯利亚仍有冷空气南下,中纬度槽脊经向度较小,以纬向环流为主;而在盛夏(7~8 月),降雹天气形势主要表现在东北—华北的低槽加强,贝加尔湖高脊向北扩展,沿脊前有冷空气南下,由于低层气温高,500 hPa 降温强,上冷下热,易使大气层结不稳定,形成强对流雷电冰雹天气。

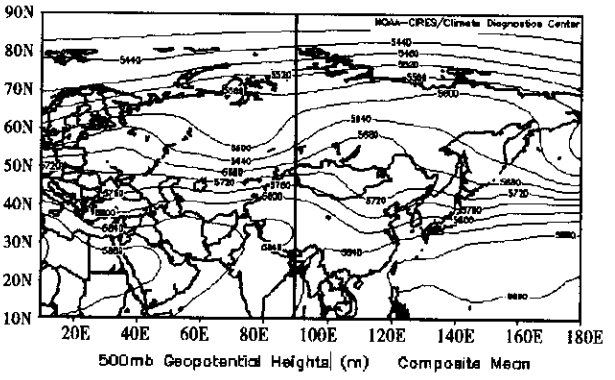

图 1 夏季西北气流型冰雹 500 hPa 环流形势

Fig. 1 500 hPa circulation situation of hail under the northwest airflow pattern in summer

秋季(9~10 月)(图略),与盛夏冰雹环流差异较大,中纬度槽脊经向度又开始变小,趋向纬向环流,从新西伯利亚群岛有冷空气开始向南扩充,但远没有春季冷空气势力强,巴尔喀什湖由槽区变为脊区,新疆脊开始变宽广,乌拉尔山北部为脊,南部为槽,说明大气环流正处在夏季向冬季的过渡阶段,强冷空气还没有大规模南下,此时地面温度还比较高,高空弱冷空气活动也能造成大气层结的不稳定。

3 用支持向量机(SVM)分类方法建立冰雹预报模型

3.1 构建预报因子

利用 1980~2002 年 4~9 月共 23 a 的高空探测资料构建预报因子库。由于 SVM 方法对因子的数

量没有明显的限制,是通过支持向量构建推理模型,可选与预报对象有明确意义的各种因子来表述预报对象和预报因子之间变化的关系。依据冯汉中的工作,样本越多,建立的 SVM 模型效果越好。

根据对陇东冰雹天气气候特征分析,春末夏初、盛夏不同季节降雹天气存在差异,分月构建冰雹预报的因子库,建立分月推理模型,就可消除季节因素。针对产生冰雹的天气系统,以 500 hPa、700 hPa 高度场、温度场为主来构造因子,主要有所选关键区的差值、24 h 变化量等 60 个因子,以反映高空环流形势场、冷平流、锋区作用等。

3.2　确立预报对象

陇东 15 个气象观测站任一站出现冰雹作为一雹日个例,由于 10 月份冰雹个例很少,年降雹日数 <3 d,概率 <0.1%,因此,主要制作 4 ~ 9 月冰雹的预报模式。1980 ~ 2002 年 4 ~ 9 月共有 295 个降雹日,出现概率为 7%,是小概率事件,其中 5 月和 6 月出现较多,在 8% 以上,9 月和 4 月较少,低于 6%(表 1)。

表 1　陇东降雹日数及概率

Tab. 1　The hail days and probability in Longdong region

月　份	4 月	5 月	6 月	7 月	8 月	9 月	合计
降雹日(d)	40	61	57	50	50	37	295
概率(%)	5.8	8.6	8.3	7.0	7.0	5.4	7.0

过去制作小概率天气预报方法或模式时,为了提高样本概率,首先进行天气分型,然后建立不同天气型下的概念模型。本文分别建立了各月不分型和天气分型后冰雹个例库,将结果进行比较。

3.3　资料处理

3.3.1　天气分型

分型原则,主要根据 500 hPa 高度差,各月略有不同,以 6 月份为例,编程序如下:

```
H1 = X(14) - X(22)
H2 = X(22) - X(34)
H3 = X(16) - X(27)
H4 = X(27) - X(37)
H5 = X(29) - X(31)
H6 = X(31) - X(41)
H7 = X(20) + X(29)
H8 = X(34) + X(81)
H9 = X(32) - X(41)
H10 = X(65) - X(81)
hh1 = H1 + H3 + H5
hh2 = H2 + H4 + H6
hh3 = H7 - H8
hh4 = H9 + H10
ffx $ = "2"
If hh1 > 2 And hh2 > 2 And hh3 > 2 And hh4 > 2 Then hh5 = 1 Else hh5 = 0
If hh1 > 11 And hh2 > 11 And hh5 = 1 Then ffx $ = "1"
If hh1 > 11 And hh3 > 11 And hh5 = 1 Then ffx $ = "1"
If hh1 > 11 And hh4 > 11 And hh5 = 1 Then ffx $ = "1"
If hh2 > 11 And hh3 > 11 And hh5 = 1 Then ffx $ = "1"
If hh2 > 11 And hh4 > 11 And hh5 = 1 Then ffx $ = "1"
If hh3 > 11 And hh4 > 11 And hh5 = 1 Then ffx $ = "1"
If (hh1 < 5 And hh2 < 2) Or (hh3 < -5 And hh4 <= -2) Then ffx $ = "3"
```

其中 X(14)、X(16)、X(20)、X(22)、X(27)、X(29)、X(31)、X(32)、X(34)、X(37)、X(41) X(65)、X(81) 分别为 500 hPa 天气图上 51656、51777、52203、52652、52818、52866、52889、53336、53543、53845、56096、57036 站的高度。

ffx $ 代表天气型,"1"为西北气流,"2"为偏西气流,"3"为西南气流。

1980 ~ 2002 年 23 a 中,6 月份共有 690 个样本,天气分型后(表 2),西北气流型样本 226 个,占总样本的 33%,包含了 73% 的冰雹个例,偏西气流型样本 277 个,占总样本的 40%,包含了 18% 的冰雹个例,西南气流型样本 187 个,占总样本的 27%,包含了 8% 的冰雹个例。由分析可见,分型后大大提高了西北气流型下的气候概率。

表 2　6 月不同天气型下冰雹概率

Tab. 2　The hail probability in different weather types in June

天气型	历史样本	占总样本比例(%)	冰雹概率(%)
西北气流	226	33	73
平直气流	277	40	18
西南气流	187	27	8

3.3.2　归一化处理

归一化处理有利于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做归一化处理,使每

一因子的数据落入区间[0,1]。具体算法是：

$$\bar{x}_i = \frac{x_i - \min(x_k)}{\max(x_k) - \min(x_k)}$$

其中 max（x_k）和 min（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、最小值不只是对于训练集，要考虑到未来可能的数据。

3.3.3 整理预报因子及预报对象

将预报因子和预报对象整理成下列格式：

```
-1  1:2263  2:114   3:5458  4:3222  5:4351  6:4088  7:7612 …
 1  1:8364  2:1574  3:7218  4:5430  5:6734  6:3211  7:2333 …
-1  1:5     2:2538  3:3676  4:3988  5:4215  6:6731  7:4567…
-1  1:3737  2:318   3:5452  4:677   5:7845  6:261   7:6312 …
-1  1:6875  2:4517  3:366   4:2392  5:6712  6:4537  7:6415 …
-1  1:2778  2:2156  3:6728  4:6775  5:5012  6:4657  7:4798 …
-1  1:5362  2:1173  3:6821  4:7659  5:7101  6:3222  7:5899…
-1  1:4782  2:3617  3:5897  4:2309  5:5334  6:4356  7:6487 …
……
```

其中第 1 列为预报对象，第 2 列以后为预报因子及其序号。

3.4 确定核函数

以径向基函数（满足 Mercer 定理条件，又称高斯核，简记为 RBF）做为核函数建立推理试验模型。径向基函数形为：

$$K(x,x_i) = \exp(-r \| x - x_i \|^2)$$

在分类预报中，基于 RBF 核求得的最终决策函数形为：

$$M(x) = \mathrm{Sgn}\Big(\sum_{\text{支持向量}} \alpha_i y_i K(x,x_i) + b\Big) =$$

$$\mathrm{Sgn}\Big(\sum_{\text{支持向量}} \alpha_i y_i \exp(-r \| x - x_i \|^2) + b\Big)$$

其中 x_i 是作为支持向量的样本因子向量；x 为待预报因子向量；α_i，b 为建立 SVM 模型待确定的系数，r 为核参数，求和运算只对支持向量进行。

3.5 建立预报模型

把资料分为 3 部分：训练集、测试集、检验集。其中训练集占 75% 的样本，测试集占 20% 的样本，检验集占 5% 的样本（以 2001～2002 年样本作为检验集）。采用中国气象局培训中心 SVM 应用开发小组开发的 CMSVM 应用软件，建立 SVM 分类预报模型。因在建立 SVM 模型中要对参数进行选取，用不同的参数训练得到的 SVM 模型中的支持向量不可能完全一样，因推理模型变化，相应的推理结果也会发生改变，什么样的参数建立的推理模型效果最好，就要对其在测试集中进行测试，我们在这里是依据推理模型对测试集的推理结果所得的 TS 评分值进行参数确定的，把 TS 评分最好的参数对应的支持向量构造的推理模型作为最终确定的推理模型，把该模型用于检验集，以检验其预报效果（推广能力）。

以 6 月份为例，建立的预报模型如下：

```
svmC Version V1.00
2 # 核函数类型 -t  最优模型中核函数参数 -C 100
-1 # 最优模型中核函数参数 -d  分类计算，
产生最优模型时的 TS 评分 11.111111
0.001 # 最优模型中核函数参数 -g
1 # 最优模型中核函数参数 -s
1 # 最优模型中核函数参数 -r
-1 # 最优模型中核函数参数 -u
60 # 训练样本的特征空间的最高维数
500 # 训练样本的个数
179 # 支持向量的个数
1.795551 # threshold b，以下每行代表一个支持向量（每行第一个实数代表 alpha * sign(y)）
-3.9401565990815048 1:-4 2:0 3:2 4:4 5:5 6:0
7:1 8:2 9:0 10:-1 11:2 12:-1 13:1 14:1 15:-2 16:0
17:-1 18:0 19:-2 20:-1 21:0 22:2 23:0 24:-1
25:3 26:1 27:0 28:-1 29:1 30:-1 31:5 32:4 33:7
34:1 35:1 36:2 37:2 38:3 39:2 40:1 41:2 42:2 43:
0 44:1 45:1 46:-1 47:1 48:1 49:1 50:1 51:2 52:1
53:0 54:4 55:5 56:4 57:3 58:1 59:1 60:-1
……
-0.55736367607210369 1:-2 2:0 3:2 4:0 5:0
6:-1 7:0 8:0 9:-3 10:1 11:1 12:1 13:1 14:-2 1
5:9 16:0 17:-2 18:2 19:-7 20:-4 21:-2
22:5 23:-2 24:-3 25:-1 26:0 27:-3 28:-1
29:1 30:-9 31:0 32:1 33:6 34:0 35:0 36:-1 3
7:0 38:0 39:0 40:-2 41:3 42:2 43:4 44:6 45:0
46:2 47:1 48:-1 49:0 50:-1 51:2 52:0 53:0
54:0 55:3 56:-4 57:-1 58:0 59:1 60:0
……
```

4 模型试验效果检验分析

用每月的全部样本资料，建立各月的 SVM 预报模型，统计分类正确率如下（表 3）：

表 3 未分型时 SVM 模型分类检验的正确率

Tab. 3 The correct rate of classified inspection with SVM without weather type division

月 份	4 月	5 月	6 月	7 月	8 月	9 月
分类检验的正确率（%）	100.0	81.8	78.8	82.9	87.9	100

从表 3 可看出,各月冰雹有无分类的正确率除 6 月稍低外,其余都在 80% 以上,效果较好。同样,按照 3.3.1 原则分型后,建立各月不同天气型下的预报模型,以 6 月份为例(表 4),西北气流型分类检验的正确率为 63.6%,偏西气流型为 66.7%(西南气流冰雹样本太少未建),与未分型的正确率 78.8% 相比,正确率都下降了。其他月份也和 6 月份有同样的结论,不再详列。

表 4　6 月份不同天气型分类检验的正确率

Tab. 4　The correct rate of classified inspection under the different weather types in June

6 月	未分型	西北气流型	偏西气流型
样本数(d)	690	226	277
分类检验的正确率(%)	78.8	63.6	66.7

从 3.3.1 分析中已得出,天气分型后,特别是西北气流型下冰雹的气候概率明显提高了,但是也看到西北气流型仅占总样本数的 33%,建模时的样本数明显少于总样本数,检验的正确率低于未进行天气分型的正确率。这基本验证了用 SVM 方法建模时,样本越大,效果越好的说法。这种说法对制作预报模型时,先进行天气分型的常规传统的天气预报思路提出了新的思考。

5　小　结

支持向量机(SVM)方法是近年发展起来的一种新的统计学习理论方法,通过对这一方法的学习,对陇东冰雹天气的主要环流形势进行了分析,对冰雹分类预报进行了探讨,经检验,主要降雹季节各月有无冰雹分类的正确率在 80% 以上,效果较好。与传统的天气分型后制作预报模式进行了比较,分型后,虽然冰雹天气气候概率提高了,但各型中总样本数减少,检验正确率降低,说明 SVM 方法不同于常规预报方法,样本越多,建立的 SVM 模型效果越好。这为基层台站制作天气预报模式提供了一种新方法,也使人们对传统的天气预报思路有了新的思考,不仅局限于特种天气型下将会出现某种天气,需要综合考虑各种因子相互作用,即寻找特征向量。在数值预报模式可靠性已有了很大提高的今天,如果用数值预报产品代替预报因子,或在预报因子中增加数值预报产品,用 SVM 寻找的特征向量将更能反映天气过程的变化。这与 2004 年全国重大天气过程总结和预报技术经验交流会上,陶诗言院士的看法一致,强对流天气预报,要从天气型的预报方法改变为以模式释用为主的预报。

参考文献:

[1] 黄嘉佑,谢庄. 卡尔曼滤波在天气预报中的运用[J]. 气象,1993,19(4):3-7.

[2] 陆如华,徐传玉,张玲. 卡尔曼滤波在天气预报中的运用技术[J]. 数值预报产品释用公报,1996,(5-6):28-36.

[3] 陈永义. 支持向量机方法及其在气象中的应用[M]. 北京:中国气象局培训中心,2004.2.

[4] 陈永义,余小鼎,高学浩,等. 处理非线性分类和回归问题的一种新方法(I)——支持向量方法简介[J]. 应用气象学报,2004,15(2):169-77.

[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(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(2):178-86.

[6] 白肇烨,徐国昌,孙学筠,等. 中国西北天气[M]. 北京:气象出版社,1991.258-372.

The Application of Support Vector Machine Method in Hail Forecast

WU Ai-min, GUO Jiang-yong, ZHANG Hong-fen, LU Ya-qi

(Qingyang Meteorological Bureau of Gansu Province, Xifeng 745000, Gansu, China)

Abstract: The support vector machine (SVM) is a new statistical study theory method which developed in recent years, by using this method in this paper, we analyzed the classification prediction of hail weather based on summarizing the circulation characteristics of hail in the east region of Gansu province, and compared with traditional forecast under different weather types. It is confirmed that SVM method is not need to classify weather types and the effect of forecast model being established is better on many samples. This has provided new method and idea for basic stations on weather forecast.

Key words: support vector machine; hail; circulation characteristic; forecast model