K 近邻非参数回归概率预报技术及其应用

翟宇梅　赵瑞星　肖仁春　王力维
(北京应用气象研究所 ,北京 100029)

摘　要

针对参数回归技术制作概率预报存在拟合好、但预报结果不稳定的现象 ,提出了用 *K* 近邻非参数回归技术制作概率预报的新途径。*K* 近邻非参数回归技术包括历史样本数据库、近邻子集生成和优化以及预报量估计 4 个主要部分。利用该技术进行了单要素概率预报(主要包括云量和降水) 和多维联合概率预报(降水、总云量、风速和气温) 试验 ,并对试验结果进行了检验。实例研究结果表明 :该文所给出的计算方案预报稳定性好 ,准确率较高 ,具有良好的业务应用价值。

关键词 : 相似预报　近邻　非参数回归估计　概率预报

引　言

传统的概率预报方法多是建立在概率密度分布参数估计的基础上。在预报前 ,需要对概率密度分布模型的分布参数进行估计 ,并建立明确的数学模型。为了得到这一参数估计模型 ,在建模过程中 ,常对建模数据作一些限制性假设 ,如假设用于建模的数据可以用有限个参数的数学表达式来拟合 ,分布密度函数是已知的或是可以估计的 ,预报误差和输入变量统计独立等等。但是 ,从一个还未被完全认知的、复杂多变的天气系统中获得的数据可能不完全满足 ,也可能完全不满足这些假设 ,这些模型的结构也并不总能保证与所描述的系统相匹配。如果这些假设肯定不满足 ,则由这种参数回归分析得到的结论就难以令人信服。因此 ,如果能够直接基于样本数据进行预报 ,则可避免概率密度估计误差的影响 ,有望提高预报的准确性和稳定性。

基于范例进行推理分析是人工智能发展较为成熟的一个分支 ,它通过研究历史档案中的基本范例实现对未来相似情况的推测 ,已被广泛应用于那些不存在或还没有比较严密的理论、但积累了大量范例样本的领域。非参数估计技术是一种类似范例推理的启发式建模技术 ,它仅依靠已经积累的、包含系统潜在关系的大量数据对新问题做出估计 ,而不需具有关于模拟过程的先验知识 ,不需建立具体的参数模型 ,它的模型隐含在系统输入和输出状态数据中 ,因此 ,有人将其称为“不对建模数据加任何限制性假设的自由分布函数的非参数估计”[1]。天气预报是一个既有相对成熟的理论 ,又有很多不确定因素的领域 ,大气变化的高度复杂性 ,使其很难用一个或一组简单的线性或非线性参数模型精确描述 ,但客观世界所具有的规整性和重现性 ,为相似预测提供了合理性依据。基于上述考虑 ,

2004-07-26 收到 ,2005-05-23 收到再改稿。

并针对用参数回归技术制作概率预报存在拟合效果好,但预报结果不稳定的现象,本文提出了用 K 近邻非参数回归技术制作概率预报的新途径。预报试验表明,利用该技术进行预报,结果稳定,准确率较高,因而具有良好的业务推广前景。

1 基本原理

通常情况下,给定一组输入和输出数据(X, Y),参数估计方法通过拟合寻找函数的具体关系式

$$Y = f(X, \beta) \tag{1}$$

其中 β 是模型参数,拟合的过程就是在全部训练数据对$\{(x_i, y_i)\}$样本集上最小化准则函数 $J(\beta)$。如果 $J(\beta)$ 是一个最小二乘准则,则当 f 是 β 的线性函数时,式(1)对应的就是线性回归。模型一旦建立,就可用这个模型计算新输入数据矢量 x_q 的输出估计值 $\hat{y}_q$。与参数估计方法相比,非参数估计方法并不关心式(1)中 f 的具体形式,而是在全部训练数据对$\{(x_i, y_i)\}$样本集上寻找关于新输入数据矢量 x_q 的近邻子集$\{(x_i, y_i)^q\}$,并优化这个子集,用这个最优近邻子集生成预报量矢量估计 $\hat{y}_q$。

如果历史资料库已经建好,且较好地搜索近邻的方法已经确定,那么在预报量矢量估计之前,最关键的一步就是最优近邻子集的生成。不同的最优近邻子集生成方法派生不同的非参数回归技术。目前常用的非参数回归技术包括 K 近邻非参数回归技术和窗非参数回归技术[2]。前者的最优近邻子集由子集中近邻的个数定义,而后者的最优近邻子集由包括近邻的窗口(或邻域)尺寸来定义。即 K 近邻非参数回归技术用固定个数的近邻生成预报量估计,而窗非参数回归技术用固定宽度窗口中的近邻生成预报量估计。本文主要讨论 K 近邻非参数估计预报方法。

上述为非参数估计技术的基本原理。对于概率预报,输出的预报量或预报量矢量的各分量是概率值,且预报量矢量估计是输入条件特征矢量出现时的条件概率。可以证明,非参数估计概率预报技术的基础是 Bayes 后验概率公式,其结果等价于 Bayes 后验概率估计结果。

2 计算方案

K 近邻非参数估计算法主要由 4 个部分组成,即存放全部历史数据对样本集的数据库 近邻样本子集的搜索、近邻子集的优化以及预报量估计。其数据流程关系参见图 1[3]。

2.1 历史样本数据库

历史样本数据库,由表示系统状态的全部历史观测资料集组成。历史数据越多越有利于非参数回归估计更加真实、完整地表达系统状态的特征,越有利于得到准确的预报量估计。在历史样本数据库中,系统状态用状态矢量存储。数据库是动态的,只要有新的数据对就放到数据库中。

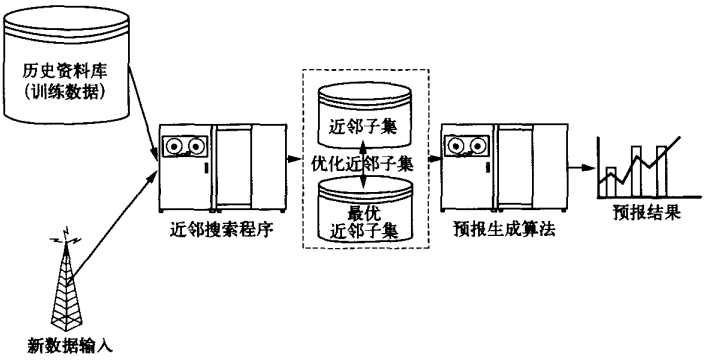

图 1　非参数估计算法示意图

2.2　近邻子集生成及其优化

搜索近邻的过程就是根据事先定义的相似性测度,在历史数据库中寻找和当前状态条件特征相似的历史记录,并把搜索到的有相似特征的历史记录标记为一个近邻,所有搜索到的近邻就组成了近邻子集。

最优近邻子集是指那些在搜索得到的所有近邻中对预报量矢量估计贡献最大的近邻组合而成的子集,一般由控制参数和一个最优化指标确定。K 近邻非参数回归技术的控制参数就是最优近邻子集的样本容量,最优化指标可同参数估计模型中的指标类似,如最小二乘准则等,目前使用较多的是最小预测误差平方和准则[4]。

最优化近邻子集的过程是一个函数寻优过程。所有函数寻优的方法都可以使用,如优选法、梯度法等,但使用较多的是枚举法和遗传算法。最优化计算一般在一组测试样本集上进行,测试样本集的选取要有一定的代表性,通常用随机抽取的方法获得,也可用和最新输入矢量时间上最接近的一段样本组成测试样本集。最优化计算可在搜索过程结束后进行,也可与搜索近邻过程同步进行。

2.3　预报量估计

最优近邻子集确定之后,就可用最优子集生成预报量估计。在数据库中已把状态特征矢量分成了条件特征矢量和预报量特征矢量,这样在最优近邻子集中的每一个近邻都有一个相应的预报量特征矢量存在,这些预报量特征矢量可以看作是与每一个近邻的条件特征矢量相对应的输出。由于最新输入的条件特征矢量 x_q 和近邻子集中的每一个近邻有一定的"距离",所以,它们各自对应的预报量矢量可能不同,但它们之间的差距相比而言是最小的。因此,对所有近邻子集中的输出矢量进行综合最有可能得到与最新输入的条件特征矢量 x_q 相对应的输出矢量。若用 Y_q 表示条件特征矢量的输出,则预报量矢量 Y_q 的估计 $\hat{Y}_q$ 为

$$\hat{Y}_q = \Phi(K, Y_i), \quad i = 1, 2, \cdots, K \tag{2}$$

式(2)中 K 是近邻子集的参数,Y_i 是第 i 个近邻对应的预报量矢量,Φ 是综合算子,用来表示对近邻子集中所有近邻相应的预报量矢量的综合技术。一般假定综合算子 Φ 是一个线性算子或一个随机算子,也可假定为非线性算子。最常用的是算术平均算子,即

$$\hat{Y}_q = \frac{1}{K}\sum_{i=1}^{K} Y_i, \quad i = 1, 2, \cdots, K \tag{3}$$

或加权平均算子

$$\hat{Y}_q = \sum_{i=1}^{K} \theta_i Y_i, \quad i = 1, 2, \cdots, K \tag{4}$$

其中 θ_i 是权重系数,满足:

$$\sum_{i=1}^{K} \theta_i = 1, \quad i = 1, 2, \cdots, K \tag{5}$$

权重系数可取距离倒数的平均值、相似系数的平均值、距离序列号倒数的平均值等。

3 应 用

非参数回归估计技术是一种基于数据驱动的启发式预报技术。它对数据不做任何限制性假设,不用建立具体的参数模型,能较好地运用蕴含于历史数据中的输入输出关系对系统的未来状态做出估计,因此,特别适合对那些内部行为不明确或不很明确的系统的未来状态进行估计。由于它的这些特性,非参数回归估计技术可用于复杂时间序列的分析与预测。

3.1 云量和降水预报

资料选用 1990~1996 年欧洲数值天气预报中心的逐日网格点资料(包括 500 hPa 高度场和地面气压场,水平分辨率为 5°×5°)以及北京站的天气实况资料。1995 年之前的样本作为训练数据,1996 年 1 月和 7 月的样本作为检验数据。预报前,对资料进行了处理,用 500 hPa 高度场和地面气压场诊断出了 500 hPa 上的高度梯度和地面气压梯度,分别用来表示 500 hPa 高度上的风场和地面风场。

近邻搜索用分矢量 VP 树[5]方法,近邻优化与搜索过程同时进行,用枚举法确定近邻个数为 4。为减少计算量,在搜索前,确定空间邻域为以预报站点为中心的区域,范围为 15×15 个经纬度,时间邻域为所要预报日的前 20 d 和后 20 d 共 40 d 的时间。

近邻搜索时所用的相似性测度[6]为

$$D_j(x, y) = \left[\left(1 - \frac{\sum_{i=1}^{m} \left|\left(X_i - \bar{X}\right) - \left(Y_i - \bar{Y}\right)\right|}{\sum_{i=1}^{m}\left(\left|X_i - \bar{X}\right| + \left|Y_i - \bar{Y}\right|\right)}\right)^2\right]^{1/2} \tag{6}$$

其中

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i, \quad \bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$$

m 是条件特征矢量的维数。依据上述方法,利用式(6)计算 1996 年数值预报资料及其诊断产品与 1990~1995 年的数值预报资料的分析场及其诊断产品间的“距离”,并在此基础上生成最优近邻子集。

在估计前,将总云量和低云量按成数分为“0~3,4~5,6~7,8~10”4 档,将降水分为无雨、小雨、中雨、大雨和暴雨 5 档。对预报结果的检验方法包括预报准确率(用于对降水

有无预报结果的评价) 和有序概率得分 R_{PS}(Ranked Probability Scores ,用于对总云量、低云量和分级降水预报结果的评价) 。有序概率得分的计算公式[7]为

$$R_{PS} = \frac{3}{2} - \frac{1}{2(L-1)} \sum_{i=1}^{L-1} [(\sum_{n=1}^{i} P_n)^2 + (\sum_{n=i+1}^{L} P_n)^2] - \frac{1}{L-1} \sum_{i=1}^{L} | i - j | \cdot P_j , \quad i = 1, 2, \cdots, L \tag{7}$$

其中 L 为概率预报类别数 , P_i 为对第 i 类天气的预报概率 , j 表示实际出现的天气类别。R_{PS}越大 ,预报准确率越高。预报完全正确时 , R_{PS}为 1 ;预报完全不正确时 , R_{PS}为 0 。图 2 是北京地区降水有无预报的准确率 ,图 3 是分级降水、总云量和低云量的有序概率得分。

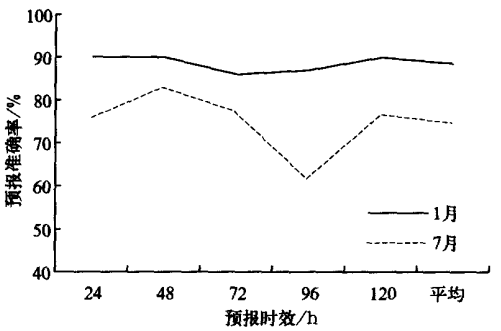

图 2　1996 年北京地区降水有无预报准确率

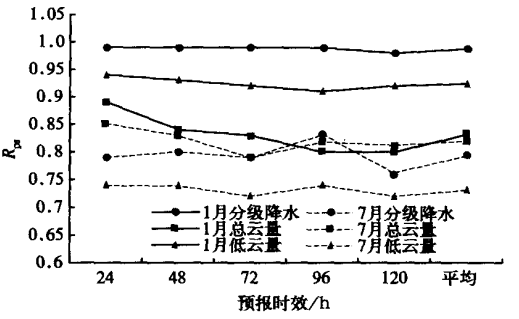

图 3　1996 年北京地区分级降水、总云量和低云量的有序概率得分

从图 2 和图 3 可以看出 ,用 K 近邻非参数回归技术制作的降水有无预报的准确率各时效平均值 1 月份为 88 .6 % ,7 月份为 75 % ;分级降水、总云量和低云量的有序概率得分平均值 1 月份分别为 0 .99 ,0 .83 和 0 .92 ,7 月份分别为 0 .79 ,0 .82 和 0 .73 ,且预报结果比较稳定。为进一步验证该方法的预报效果 ,与传统的相似预报结果进行了比较 ,表 1 是两种方法各时效(24 ,48 ,72 ,96 ,120 h) 预报结果的平均值 ,显然 ,前者的预报效果明显优于后者。这主要是由于 K 近邻方法用了前 K 个相似样本 ,所含预报信息更多 ,从统计意义上讲 ,更能代表要预报的样本属性。

表 1　K 近邻非参数回归预报技术与相似预报技术的预报结果比较

方法	1 月份预报评分				7 月份预报评分			
	降水有无	总云量	低云量	降水分级	降水有无	总云量	低云量	降水分级
K 近邻	89	83	92	99	75	82	73	79
相似	82	74	88	98	67	71	61	73

从 K 近邻非参数回归技术的物理基础看 ,它对条件状态矢量和预报量矢量均不加任何限制 ,不需具有关于模拟过程的先验知识 ,仅用足够多的历史数据来建立输入和输出之间的内在关系。因此利用该技术可实现多种要素或天气现象的同时预报。值得指出的是 ,非参数回归技术的预报效果与历史数据库容量有关 ,如果拥有能反映系统状态可能变

化范围的较大历史数据库,预报准确率可稳定提高。

3.2 2003 年 7 月淮河流域降水预报

2003 年汛期,淮河流域出现持续降雨,降雨量一般为 80 ~ 150 mm,其中安徽西南部和河南部分地区降雨总量达 200 ~ 400 mm,与常年同期相比,淮河流域降雨量偏多 9 成到 1.7 倍。为进一步检验 *K* 近邻非参数回归技术的预报能力,从 2003 年 7 月 1 日起,用该技术进行了降水分级的实时预报试验,直至淮河流域 7 月 22 日降水区北抬,淮河水情减弱结束,并对预报结果进行了评价。

近邻搜索算法和时空邻域选择方案同 3.1 节,训练数据为 1990 年以后至预报日前 1 天的逐日地面气象观测资料和欧洲数值天气预报中心的分析场及用 T106 模式每天制作的 24 ~ 240 h 预报产品,预报站点包括信阳、蚌埠、安康、阜阳、南阳 5 个气象站。图 4 是预报结果的有序概率得分。

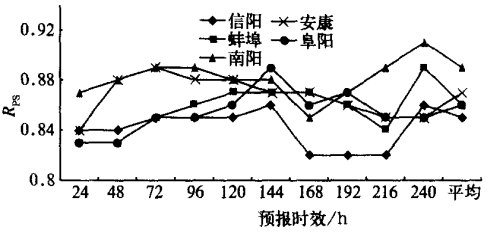

图 4 日降水量的有序概率得分

从图 4 看出,有序概率得分可达 0.82 以上,最高达 0.91,各站各时效的平均值接近 0.85 以上,且预报稳定性较好。可见,用非参数回归方法对 2003 年 7 月淮河流域降水的预报是成功的。

3.3 多变量联合概率预报

天气对民用设备、军用装备以及农事活动等的影响往往表现为多种气象要素的共同作用,如飞机起降的最低气象条件就与能见度、风、降水、雷暴等气象要素或天气现象有关。因此,多种气象要素的综合预报(以概率形式表示时,为多变量联合概率预报)具有特别重要的实际应用价值。为了计算多变量事件的联合概率,Abramowitz 从连续变量的联合分布出发,设计了一套近似计算联合概率的半经验方法[8]。由于所给方案涉及到多重积分的计算,所以一般只计算到 3 个变量事件的联合概率,对于 3 个以上变量事件的联合概率因其计算的高度复杂性成为统计预报领域的一个难题。

所谓多变量联合概率实际上是从多个子事件角度说的,如果从一个合成事件的角度出发,也可以认为是一维事件概率。因此,可从事件的定义入手,将多变量事件转换为"单"变量事件,进而借助单变量事件的概率预报方法进行多变量联合概率预报。作为 1 个例子,给出 1 个"好天"事件的概率预报试验。"好天"事件的具体定义是,如果某个时次满足条件:无降水、总云量 ≤3 成、风速 ≤6 m/s、温度 ≤30 ℃,则定义这个时次为"好天"。对于"好天"这一"单"变量事件,就可以利用 *K* 近邻非参数回归概率天气预报技术进行预报。

试验所用资料为欧洲数值天气预报中心的网格点资料以及北京站的天气实况资料。1990 年 1 月至 2002 年 6 月的样本作为训练数据,2002 年 7 月的样本作为检验数据。近邻搜索算法和时空邻域选择方案同 3.1 节,为简单起见,预报生成方法用算术平均。用 *K* 近邻非参数回归方法制作了 2002 年 7 月"好天"的概率预报,并用预报准确率指标对预报结果进行了评价,表 2 是预报评价结果。从中可以看出预报准确率比较高,72 h 之内预报准确率均在 80 %以上,最低也达到 72.5 %。

表 2　2002 年 7 月北京站“好天”概率预报评价结果

预报时效/ h	24	48	72	96	120	144	168
预报准确率/ %	80.0	83.8	81.3	76.3	73.8	72.5	77.8

4　结　语

非参数回归估计概率预报技术是一种类似于范例推理的启发式天气预报技术,是天气预报员根据天气学原理和天气预报经验进行天气分析和预报的仿真算法。它不需具有关于模拟过程的先验知识,不需建立具体的参数模型,从而避免了概率密度估计误差的影响。基于蕴含于数据中的关联知识进行预报是它区别于参数估计模式的最显著特点。

用 *K* 近邻非参数估计技术进行预报时,对待预报的天气样本,需要计算它与已储存的所有天气样本间的“距离”,并用前 *K* 个最相似样本计算待预报样本的天气事件属性,从这一意义上讲,它是相似预报方法的推广。

该方法的缺点是需要储存大量的历史天气样本资料,而且每次预报都要计算待预报样本和全部历史样本间的“距离”。因此,计算量较大,制作预报的时间相对较长,但通过对近邻搜索算法和程序的优化,可以满足业务预报对时效性的要求。

本文所给出的 *K* 近邻非参数回归估计计算方案思路清晰、实现简单、运行稳定可靠且预报准确率较高,在复杂时间序列的分析和预报方面具有广阔的应用前景。

参 考 文 献

1　Schaal Stefan. Nonparametric Regression for Learning. Proceeding of the Conference on Prerational Intelligence. Germany, 1994.

2　Altman N S. An introduction to kernel and nearest neighbor nonparametric regression. *The American Statistician*, 1992, **46**(3):175 ~ 185.

3　Oswald R K, William T S, Brian L S. Traffic Flow Forecasting Using Approximate Nearest Neighbor Nonparametric Regression. Research Report, No. UVACTS-15-13-7, 2001.

4　施能. 气象科研与预报中的多元分析方法. 北京:气象出版社,1995. 85 ~ 88.

5　Yianilos P N. Data Structure and Algorithms for Nearest Neighbor Search in General Metric Space. Proceeding of the ACM-SIAM Symposium on Discrete Algorithms. 1993. 311 ~ 321.

6　罗阳.一种新的相似性度量 ——高分辨相似系数.空军气象学院学报,1996,**17**(1):23 ~ 32.

7　朱盛明,曲学实. 数值预报产品统计解释技术的进展. 北京:气象出版社,1988. 163 ~ 165.

8　张军,葛军,田俊杰,等. 概率天气预报及其应用. 北京:气象出版社,1998. 29 ~ 43.

K-NEAREST NEIGHBOR NONPARAMETRIC REGRESSION FOR PROBABILITY FORECASTING WITH ITS APPLICATIONS

Zhai Yumei　Zhao Ruixing　Xiao Renchun　Wang Liwei

(*Beijing Institute of Applied Meteorology*, *Beijing* 100029)

Abstract

Although probability forecasts based on a parametric regression scheme have good fitting rates the results are not so stable. For this reason, a new approach is proposed to such forecasts by means of a *K*-nearest neighbor nonparametric regression technique, and the technique includes 4 main components such as a database of historical samples, production of nearest neighbor subsets, their optimization and estimate of predictands. Case experiments are conducted on univariate (cloudiness or precipitation) and multivariate joint (e.g., rainfall, total cloudiness, wind speed and temperature) probability forecasting, with the results tested. Results show that forecasts from the nonparametric regression scheme are high-stability, with good prospects in operational weather forecast.

Key words: Analogue forecasting Nearest neighbor Nonparametric regression estimation Probability forecasting