彭昱忠,王 谦,元昌安,等. 数据挖掘技术在气象预报研究中的应用[J]. 干旱气象,2015,33(1):19-27, [PENG Yuzhong, WANG Qian, YUAN Changan, et al. Review of Research on Data Mining in Application of Meteorological Forecasting[J]. Journal of Arid Meteorology, 2015, 33(1):19-27], doi:10.11755/j.issn.1006-7639(2015)-01-0019

数据挖掘技术在气象预报研究中的应用

彭昱忠[1,2],王 谦[2],元昌安[1],林开平[3]

(1. 广西师范学院,科学计算与智能信息处理广西高校重点实验室,广西 南宁 530001;2. 广西师范学院,北部湾可持续发展监测与优化教育部重点实验室,广西 南宁 530001;3. 广西壮族自治区气象台,广西 南宁 530022)

摘 要: 气象预测是现代世界最重要和最有挑战的问题,准确的气象预测常需要使用比较先进的方法和计算机模型。本文分析了数据挖掘方法在气象预报中应用的国内外研究现状,简要介绍了目前在大气科学领域应用的一些数据挖掘方法的相关概念、原理和特点,综述了数据挖掘方法在气象预报中的最新应用研究进展,讨论了这些数据挖掘方法在气象预报中的优缺点。最后指出了当前基于数据挖掘方法的气象预报技术存在的一些困难,并对未来的研究重点和和发展趋势进行展望。

关键词: 数据挖掘; 气象预报; 气象数据挖掘; 气象建模; 计算智能

文章编号: 1006-7639(2015)-01-0019-09 doi:10.11755/j.issn.1006-7639(2015)-01-0019

中图分类号: TP18 **文献标识码**: A

引 言

天气预报是根据气象观测资料,应用天气学、动力气象学、统计学等原理和方法,参考某一区域的气候背景和天气演变规律的基础上,对该区域未来一定时段的天气状况做出定性或定量的预测。作为一门预测学科,由于受到各种尺度的天气系统之间的制约和相互作用错综复杂,影响天气变化的因素众多,且关系极为复杂,气象预报存在一定程度的不确定性。目前气象预报主要使用4种预报处理方法:第一种是经验预报方法,在天气图形势预报的基础上,根据天气系统的未来位置和强度,对未来的天气分布做出预测;第二种是统计预报方法,通过统计某一现象在历史上特定的环境条件下出现的概率,来推测在未来存在类似环境时出现的可能性;第三种是数值预报方法,利用大气运动方程组,在一定的初值和边值条件下对方程组进行积分,预报未来的天气;还有一种是集成预报方法,即把不同预报方法对同一要素的多种预报结果综合在一起,从而得出一个优于单一预报方法的预报结果[1]。

近几十年来,天气预报研究取得了巨大的进步,预报质量也逐步提高,但与社会需求尚有一定的差距。近年,国际上借助先进的智能计算和数据挖掘方法研究和改进气象预报方法和模型,以期提高对未知气象规律的认识和提高气象预测预报能力,已逐渐成为气象、数学和计算机领域的专家和学者们关注的热点。2010年美国科学基金会以1 000万美元经费资助橡树岭国家实验室等7个研究单位合作的关于运用先进的人工智能、数据挖掘和数学理论等方法研究气候变化问题的重大研究项目,该项目更激励了国际上数据挖掘研究在气象领域的拓展,近年多个相关国际会议上设置了相关的专题和workshop。数据挖掘就是利用统计、机器学习、软计算方法、数据库等多个学科领域的先进技术,对大量历史数据进行分析处理,从中挖掘出隐含的、事先未知的和有价值的知识,为人们的决策分析提供更高层次的技术支持[2-3]。气象观测数据和探测资料是海量数据,有文本格式和图像格式等,具有空间和时间等属性。随着新的气象观测设备的普及与应用,气象观测的空间密度进一步提高,数以万计的气象观测站点的地面和高空实况资料,以及卫星、雷达探测资料成倍增长[4]。然而大量的气象资料未能有

收稿日期: 2014-04-25; **改回日期**: 2014-05-25

基金项目: 国家自然科学基金(#61363037)、广西自然科学基金青年项目(2012GXNSFBA053161)、广西高校科学技术研究重点项目(ZD2014083)共同资助

作者简介: 彭昱忠(1980-),男,硕士,副教授,主要研究方向为气象数据挖掘、智能信息处理. E-mail: jedison@163.com

通讯作者: 元昌安(1964-),男,教授,主要研究方向为数据库与知识工程、智能计算. E-mail: yca@gxtc.edu.cn

效处理和充分利用，以致未能在天气预报中发挥更多积极的作用[5]。如何去挖掘和发现隐藏在这些海量数据中的天气规律，寻找对天气演变有指示意义的信号，数据挖掘方法可以发挥重要的作用。

目前国内外利用数据挖掘方法进行气象预报的研究主要有 2 个方向：一是基于统计学的数理统计方法，主要采用的数学分析方法有谐波成分分析、方差统计分析、多元回归分析、EOF 展开式、属性相关度分析、关联分析、基于小波分析等；另一个是基于机器学习、软计算方法的研究。主要应用于气象数据建模、空间数据分析、气象时态序列数据分析、气象模式识别等。本文将对当前应用于气象预报中的一些数据挖掘方法的基本原理及优缺点分别进行论述，以期能够对气象工作者及相关领域学者继续深入研究提供有用的参考。

1 数据挖掘方法在气象预报中的应用研究

气象数据同时具有时空属性、多维、多尺度、非平稳、不确定、周期性强、属性相关度高等特点，仅用传统方法对气象数据进行分析和处理会遇到不少困难。将数据挖掘方法应用于气象领域数据的分析和处理，探索各种气象要素间的、及其与天气现象间的内在联系，寻找各种潜在规律去揭示未知的气象理论，不但对气象科学研究很重要，而且能够在丰富天气预报方法、提高天气预报水平等方面产生积极重要的影响。目前，较为常用的基于数据挖掘方法的预报技术有人工神经网络、遗传算法、支持向量机、贝叶斯、决策树和关联规则挖掘等，下面将对这几种方法在气象预报中的研究应用分别进行论述。

1.1 基于人工神经网络方法的气象预报方法研究

神经网络是一种计算模型，由大量的节点(神经元)和之间相互的联接构成。每个节点代表一种特定的输出函数，称为激励函数(activation function)。每 2 个节点间的连接都代表一个对于通过该连接信号的加权值，称之为权重。在结构上，可以把一个神经网络划分为输入层、输出层和隐含层。神经网络技术具有很强的非线性映射能力，已被学者用数学证明了可对任意连续函数进行逼近[6-7]，并且具有较强的并行性、自适应性、容错性及自学习能力，被广泛用于各种复杂系统的建模和解决因果关系复杂的非确定性推理、判断、预测与分类和非线性问题。

由于神经网络是一个大规模的非线性自适应系统，通过对样本的学习建立起记忆，然后将未知模式判定为其最接近的记忆，这与具有耗散的、多个不稳定源的高阶非线性特性的气候系统有着极其相似的特点，两者的相似性决定了用神经网络进行气象预报的可能性。

近年来，国内外大气科学领域开展了大量的人工神经网络研究，并取得了显著成果。金龙等[8-9]用神经网络方法开展了不同区域范围的暴雨预报研究，认为神经网络方法确实可以通过对网络的学习训练，从原始数据中提取足够的分类信息，从而达到较高的预报准确率。另外，还针对神经网络方法在预报建模中存在的“过拟合”致低泛化性能等问题，提出了采用主成分分析构造神经网络低维学习矩阵的预报建模方法，比传统的神经网络预报建模方法及逐步回归预报模型，泛化能力有显著提高；段文广等[10]基于时序分析技术，建立适合于 BP 神经网络的输入样本模型，通过反复学习从温度时序中建立预测模型，将其用于未来 24 h 的精细化温度预报；Tang 等[11]利用神经网络方法和不完全动力系统相耦合方法建立了一个混合神经网络——动力模型，用混合模型来模拟一些数据变量，结果表明该方法具有良好的预报性能。Mihalakakou 等[12]利用神经网络方法对总太阳辐射时间序列进行短期预报，将其结果与自回归模型预报结果进行比较，发现神经网络预报模型的预报效果远远好于自回归模型；Hung 等[13]利用曼谷的 75 个观测站 4 a 的每小时降水序列数据，通过神经网络进行建模，对未来 1～6 h 的降水量进行预报和洪水管理；Vamsidhar 等[14]构建了一个反向传播神经网络模型，利用印度某地的湿度、露点温度和气压数据预测降水量，在训练集中获得 99.79% 的准确率和测试集中获得 94.28% 的准确率；Chadwick 等[15]利用神经网络建立全球气候模式降尺度分析预报欧洲的区域温度和雨量；Gordon Reikard[16]利用频域算法和神经网络构建了一个空间天气预报模型，该模型先用频域算法捕捉，后用神经网络捕捉短期独立信号和趋势信息，获得了较好的预报效果。

在集成预报方面，人工神经网络也显示了它强大的处理非线性问题的能力。陈云浩等[17]提出了基于人工神经网络的集成降雨时序分析与预报模型，并根据上海市降雨和实际监测点分布的特点，建立了城市降雨空间差异预报模型，经验证，该模型具有较高精度，可为城市规划、排水市政工程建设服务；吴建生等[18-19]以 BP 算法为基本框架，引入改进后的神经网络算法对个体进行训练，并通过 PSO

等优化和回归方法建立气象集成预测模型,以广西的月降水量进行实例分析,计算结果表明该方法学习能力强、泛化性能高,预报精度高、而且稳定; Langella [20] 利用神经网络建立高时空分辨率的降雨预测集成推断系统框架; Wang L 等 [21] 设计了一个复合神经网络集成预报模型进行降水预报,该预报模型先采用 Bagging 和 Boosting 方法进行数据采样,训练各个神经网络子模型,然后利用偏最小二乘法选择合适的子模型个数,用小波支持向量机对这些选择出来的神经网络子模型集成构成预报模型,其试验预报效果比多种神经网络预报模型及集成模型的效果要好。Karthik Nadig [22] 提出了用神经网络集成模型提高已有的单个神经网络模型空气温度和露点温度的预测准确率和减少预测异常值的次数。

但是,神经网络在实际应用中也暴露出自身的一些弱点。如,收敛慢、学习时间相对长,算法不完备,容易陷入局部极小点; 鲁棒性不好,网络性能对网络的初始设置比较敏感,机理分析比较困难。这些缺点都限制了它在实际天气预报中的进一步广泛应用。

1.2　基于遗传计算及其融合算法的气象预报方法研究

遗传计算(包括遗传算法和遗传编程等遗传进化算法) ,是一种借鉴自然界生物种类遗传和进化过程而形成的自适应全局优化搜索算法。其主要特点是群体搜索策略和群体中个体之间的信息交换,搜索不依赖于梯度信息,具有较强移植性和通用性,它尤其适用于处理传统搜索方法难以解决的复杂非线性问题,对于处理需要进行全局优化的问题也有着较强的优势 [7] 。而这些特点与具有明显非线性演变特征的气象预报问题有着极其相似的地方,两者之间的共同点决定了将遗传算法应用于气象预报中的可能性。

熊聪聪等 [23] 根据天津市气象科学研究所提供的历史气象统计资料,选取相邻的 20 个站点 3 个月的历史数据作为一次试验样本,利用遗传算法得出各预报模式预报结果的权重系数,进而得到该气象要素的集成预报模型,对影响集成预报准确性的因素进行了分析和验证,结果表明利用遗传算法可以实现较好的集成性天气预报; 吴清佳等 [24] 提出了一种神经网络和遗传算法相结合的天气预报方法,对卫星雷达观测到的气象数据加以处理分析,并以上海地区的天气预报作为试验,试验结果得到了上海市气象局有关专家的肯定; 吴建生等 [25] 采用由遗传算法改进后的神经网络模型,并以广西全区的月降水量作为实例进行分析和计算,计算结果表明,该方法预报精度高、而且稳定; Li Chen 等 [26] 提出了一种改进的遗传编程模型利用多变量气象卫星数据进行 SSM/I 估计海上的台风降水; Hisham Ihshaish 等 [27] 提出了一种基于遗传算法和数值预报模式的天气预报策略(G - Ensemble) 在墨西哥湾飓风等历史气象事件中试验,该策略利用遗传算法优化数值预报模式的输入参数,从而提升数值模式的预报效果; Siva Venkadesh 等 [28] 在每个预测水平利用遗传算法确定神经网络模型输入的每个环境变量的最佳时段和分辨率,改进预报效果; Ka Yan Wong [29] 基于遗传算法框架设计了一个通用的天气系统识别模型,该系统模型可从多维的数值预报数据中识别出天气系统,达到 80% ~100% 的定位准确率,且还可能发现预报者易忽略的可揭示成因或耗散预报特征要素。

由于遗传计算方法具有良好的全局搜索能力,可以快速地将解空间中的全体解搜索出来,而不会陷入局部最优解的快速下降陷阱; 并且利用它的内在并行性,可以方便地进行分布式计算,加快求解速度。但是遗传算法的局部搜索能力较差,导致单纯的遗传算法比较费时,在进化后期搜索效率较低; 而遗传编程算法过程相对复杂,处理效率较低,并可能引起代码膨胀影响搜索效率等。在实际应用中,现有遗传计算容易产生早熟收敛的问题。采用何种方法既能够使优良个体得以保留,又能够维持群体的多样性一直是遗传计算中较难解决的问题。

1.3　基于支持向量机方法的气象预报方法研究

支持向量机(SVM,Support Vector Machine) 是统计学家 Vapnik 提出的一种建立在统计学习理论的 VC 维理论和结构风险最小原理基础上的有监督机器学习的新方法 [30] ,近年来受到了学术界的重视,并得到了广泛的应用。SVM 方法的基本思想是: 通过非线性映射把样本空间映射到一个高维乃至于无穷维的特征空间(Hilbert 空间) ,在特征空间中寻求最优划分或回归线性超平面,从而解决样本空间中的高度非线性分类和回归等问题。

SVM 方法与传统的气象预测方法(如多元回归方法,卡尔曼滤波方法等) 相比有明显的优势。首先,它不依赖于模型的选择,且 SVM 本身对不同方法具有一定的不敏感性,能够一定程度地避免维数过高和过拟合等问题,具有预测精度高,求解速度快等优点,更适合解决实际中的小样本问题。此外,SVM 具有良好的泛化能力和抗过敏能力,在处理具

有非线性特征的气象要素或天气现象(如,降水)的预报时有着明显的优势[31]。陈水义和冯汉中等人对 SVM 方法在气象预报领域进行了一些探讨性的试验,表明 SVM 方法能用于具有显著非线性特征的气象预报中,所得出的 SVM 推理模型具有良好的预报能力,在短期天气预报、数值预报释用、实时短期预报业务等方面有良好的预报前景[32-33]。王小萍等[34]用 2002 ~ 2004 年 5 个月的国家气象中心 T213L31 模式的初始资料和预报资料,对 500 hPa 高度场的预报技巧——距平相关系数进行了统计分析,针对各季节分别设计了基于支持向量机的预报模型,研究结果表明:支持向量机模型预报值可以较好地反映预报技巧变化的趋势,将其应用于气象预报中是可行的,有意义的;王在文[35]收集整理了北京奥运场馆 2006 ~ 2007 年 6 ~ 9 月的历史气象资料并形成样本集,利用 SVM 方法对所得到的样本集进行参数调优建模,对 2008 年 6 ~ 9 月该站点要素提供的业务释用预报进行检验,并和原中尺度模式 MM5V3 的预报进行对比。结果显示,在原数值模式预报的基础上释用预报能提供更加精确的场馆业务预报;樊高峰等[36]选择 SVM 方法,利用 8 月南方涛动指数等 15 项因子,基于径向基核函数建立浙江省秋季的干旱预测模型并进行了预测,结果表明,建立的干旱预测模型能直接对秋季干旱进行预测,并且有较高的准确率,可为气候预测中从气候要素预测到气象灾害预测提供一种有效途径。汪春秀[37]以浙北地区 5 ~ 9 月的气象资料(通过均生函数处理的一维气象数据——降雨量)作为实验样本建立 SVM 模型,并对汛期降水量进行预测。仿真实验表明,基于均生函数的 SVM 模型用于降水量预测中具有较高的精度,并在处理蕴含丰富信息的一维时间序列问题上具有较好的优势;滕卫平等[38]利用影响热带气旋的气候学和持续性提取高相关因子,然后分别采用支持向量机和逐步回归方法建立西太平洋地区热带气旋路径中期预报模型。试报结果表明:支持向量机回归模型无论在历史样本拟合的精度上还是在实际预报的能力上都优于逐步回归模型,而且这种优势随着预报时效的延长越来越明显;Hong[39]利用混沌粒子群优化的 SVM 进行降水预报,实验效果比递归神经网络等模型好;Radhika 等[40]利用非线性回归方法训练的 SVM 模型预测某区域的 2 ~ 10 d 的气温,结果比利用反向传播的多层感知网络模型效果好;Ozgur Kisi 等[41]提出一个 wavelet - SVM 联合模型进行日降水预报,该模型先用离散小波对降水时间序列进行分解重构,然后用支持向量回归机进行建模预报,实验结果比神经网络预报模型好;R. Usha Rani 等[42]提出了一个多核的增强支持向量回归(ESVR)天气预报模型,该模型采用多个可解释的核函数分别处理输入的气温、气压、湿度和露点温度等气象要素,并通过自组织映射树模型集成输出,可获得比多层感知网络模型更好的预测性能。

然而,支持向量机在实际应用中也存在 2 点不足:(1)传统 SVM 对于实际大规模问题的训练速度较慢;(2)用 SVM 解决多分类问题存在困难。这 2 个问题也在不同程度上制约了 SVM 在气象预报中的发展。

1.4 基于贝叶斯方法的气象预报方法研究

贝叶斯方法是将关于未知参数的先验信息与样本信息综合,再根据贝叶斯公式,得出后验信息,然后根据后验信息去推断未知参数的一种主观推测方法。其估计取决于先验知识的正确性和新证据的丰富积累,新证据的积累有利于推测分类的准确性,新证据越多,分类的准确性越高。这种推测方式适用于对不能重复的实验进行推测,在模式识别、人工智能、数据挖掘等领域都得到了广泛的研究[7]。

在气象预报领域中,之所以采用贝叶斯网络进行气象预报建模主要是基于以下几点的考虑:(1)贝叶斯网络的拓扑结构——节点和边的集合,用一种精确而简洁的方式描述了在域中成立的条件独立关系,便于气象预报人员建立其拓扑关系;(2)贝叶斯网络的推理能力比较强,能够充分描述人类的推理模式,其图形化的表示方式贴切的蕴含了网络节点变量之间的因果关系及条件相关关系,能够直观地展现气象属性与气象预报之间的因果联系,便于帮助预报人员的理解和开发;(3)贝叶斯网络将不确定性的考虑融入了条件概率表(Conditional Probability Table)中,采用条件概率表达各个信息要素之间的相关关系,能在有限的、不完整的、不确定的信息条件下进行学习和推理,符合气象学中处理不确定性的要求,对于处理气象信息的不确定性提供了一个简洁而又准确的分析方法[43]。以上 3 点决定了使用贝叶斯方法进行气象预报的可能性及其优势所在。

何伟等[44]用朴素贝叶斯分类器和预测降雨量的朴素贝叶斯算法对降雨量预测问题进行了分类研究,分别以郑州市 1971 ~ 2000 年 6 ~ 8 月的气象数据为训练集,对郑州市 2001 ~ 2004 年相应月份的月降雨总量进行了预测。实验表明,基于朴素贝叶斯分类器和朴素贝叶斯算法的预测精度明显高于目前

短期气候预测中采用的回归分析、聚类分析等其它预测方法; 陈朝平等[45]在贝叶斯概率决策理论的基础上,利用四川境内 1951 ~2004 年 147 个预报站点暴雨的气候概率对西南区域中尺度集合预报模式提供的预报产品进行了修正。从预报试验结果来看: 基于贝叶斯方法修正后的集合概率预报产品在一定程度上消除了空报; 顾锦荣等[46]利用贝叶斯网络建立了天气条件对于飞行威胁度的计算模型,并进一步利用贝叶斯网络产生的计算数据和支持向量机方法建立了天气威胁度的预报模型,根据对小样本的学习、预报检验,证明该模型效果良好; 梁莉等[47]利用淮河流域加密站点逐日降水和日最高、最低温度资料,以及对应的 T213 数值预报产品的集合预报,采用贝叶斯模型平均(Bayesian Model Averaging, BMA) 方法进行水文概率预报。结果表明: 采用 BMA 方法以概率分布的形式描述预报不确定性,这对减少降水预报误差、提高预报准确率、做好洪水预报及防灾减灾工作有重要意义; Rafael Cano 等[48]利用贝叶斯网络对 1959 ~ 2000 年伊比利亚半岛 100 个气象站观测数据及部分数值预报模式的输出数据进行预报建模和数据挖掘,成功用于本地日降水和最大风速预报,以及观测数据缺损值修补和插值等; Yu Jiang 等[49]根据贝叶斯理论和结构突变模型原理,提出一个贝叶斯结构突变模型用于短时风速预报,在实验中获得了比 ARMA 模型和神经网络模型更高的预报质量; Cyril Voyant 等[50]为了获得更准确的每小时太阳辐射总量预测值,构建了由贝叶斯模型、多层感知网络模型、ARMA 模型和持续模型(persistence models) 组成的混合模型,该模型中主要利用贝叶斯模型进行模式的选择和各子模型预测结果的概率集成输出, 实验结果的 nRMSE 较单个子模型更好; Valmik B Nikam 等[51]利用贝叶斯理论建立一个有监督学习分类器,并以此为基础建立数据密集型的降水预报模型,并利用印度多个城市的历史降水数据进行试验,效果良好,且相比数值预报模式等计算密集型的预报模型计算更简单,需要资源更少; McLean 等[52]通过利用二元分布密度函数扩展了贝叶斯模型平均(Bayesian model averaging) 理论对集合预报的输出进行后验处理构建了提前 48 h 的风矢量概率预报模型,并使用华盛顿大学的中尺度集合对 2003 年北美太平洋西北地区的风矢量校准概率预报比原始中尺度集合预报获得更好的效果,且比概率气候预报的输出结果更清晰。

但是,贝叶斯分析法并非是一种完美无缺的方法。首先,如何确定先验概率确实是一个非常麻烦且需进一步研究的问题。贝叶斯分析法在求后验概率的过程中,同时使用相对频数和主观概率的作法也不一定是非常合理的。其次,如何确定损失函数、给出公式化的损失函数,也是一个值得研究的问题。如果没有具体的公式化的损失函数,各具体损失值是难以计算的。另外,在贝叶斯决策理论中也有许多未能解决的问题。如对所有可能的事件是否应同等对待,怎样更好地对不确定事件进行数量化评价等,都是需要进一步明确的。当然,我们指出贝叶斯分析法的不足并非意味着否定这一方法的科学性及实用性。

1.5　基于决策树方法的气象预报方法研究

决策树是空间数据挖掘进行自动分类的方法之一,它是以规则形式对数据进行自动分类[16]。由于该方法以图形化的方式表示数据挖掘结果,浅显易懂,易于做出判断,目前已在遥感影像处理、环境演变、灾害天气预测等方面得到了广泛应用。

张海玲等[53]根据香港天文台提供的 2001 年 8 月和 2002 年 6 ~7 月业务区域谱模式(ORSM) 物理量场预报值,运用决策树方法探讨空间环境物理量场与暴雨中心暴雨量之间关系的研究思路。结果表明,研究区暴雨中心降雨量的多少与周边相对湿度、海平面高度、温度和经向风速以及这些物理量场所处的纬度位置关系密切,而与这些物理量场所处的经度位置和纬向风速关系较小; 徐会明等[54]利用 8 个月的闪电定位仪、常规探空以及 T213 数值预报产品等资料作为样本,应用决策树方法生成四川省未来 12 h 雷电潜势预报决策树,并通过雷电潜势预报决策树作四川省 2007 年 8 ~9 月雷电潜势预报,效果较好; Prasad N 等[55]研究用 GINI 系数作为测试属性的选择标准的 SLIQ 决策树对与降水密切相关的湿度、温度、大气压、风速和露点温度等气象要素的历史数据进行降水预报建模,获得了 72.3% 的准确率; Royston S 等[56]利用语义决策树(linguistic decision tree) 对现有的水位和气象数据进行风暴潮的规则挖掘和预报建模,提供了在泰晤士河口精度为 0.1 m 提前 8 h 的风暴潮预报。

决策树也存在着一些缺点。首先,决策树算法不易处理连续数据。数据的属性域必须被划分为不同的类别才能处理,但是并非所有的分类问题都可以明确划分区域类型; 其次,决策树算法对缺失数据难以处理,这是由于不能对缺失数据产生正确的分支进而影响了整个决策树的生成; 最后,决策树的过程忽略了数据库属性之间的相关性。上述这些问题都或多或少的限制了决策树方法在气象预报中的进

一步应用。

1.6 基于关联规则挖掘的气象预报方法研究

关联规则挖掘就是通过对历史数据进行查询和遍历，从大量数据中提取或者挖掘出有用的知识，找出存在于数据之间的频繁模式、潜在联系或因果结构[16]。

在气象预报业务的实际应用中，万谦等[57]在正态云关联规则算法的基础上进行了进一步的研究，并利用求正态云关联规则的支持率和信任度来进行预测。将这种方法应用于气象预报，结果表明：该预测方法简单易懂，更容易被人理解和使用；Guo Z等[58]提取了MCS周边的环境物理量场特征值，进而运用空间数据挖掘中的关联规则建立了高原上MCS东移传播与其环境物理量场之间的数学模型，确立了长江流域暴雨中心周边的环境场特征与暴雨中心暴雨量之间的关系，以此提高长江中下游的灾害天气预报的准确性；朱冬梅等[59]针对贵州1995～2010年全省各个气象站的气象资料进行分析，并将改进的关联规则生成算法应用到成都信息工程学院飓风研究所开发的DSS系统进行分析与预测，挖掘出地形与干旱之间的潜在联系。结果表明在改进Apriori算法的帮助下，预报员可以根据当地的地形更加准确地预测降水量的大小，从而提高预报准确率；林万涛等[60]提出一种应用关联规则方法挖掘模式风速模拟数据与气压、温度、湿度及较大风速预测误差之间的关联关系，进而用统计及智能优化方法修正模式误差，极大地提高了风电场风速预报精度；迟德中[61]提供了一种新的基于WRF模式和关联规则优化的风电场短期风速预报方法。利用WRF模式预测了2010年4～6月承德红松风电场的风速数据，之后采用关联规则优化方法对这种模式的预测结果进行校正。研究表明这种新提出的短期风速预报方法，能有效提高短期风速的预测精度，保证电力系统的稳定运行；王红霞等[62]提出一种基于月粒度统计数据进行挖掘属性构造的方法，通过对离散化处理后的数据做关联规则挖掘，进行干旱预测模型的建立。对实验数据处理结果表明：挖掘规则合理、预测结果可信、预测效率提高；郑忠平[63]使用围绕极端值的三分聚类算法和Apriori算法对重庆地区的气象资料进行了深入挖掘，得出了一系列有积极意义的结果。这些结果对于发现该地区灾害性天气的分布特征和导致灾害性天气的相关因素有着积极的现实意义；Sherri Harms等[64]提出了用关联分析方法做中、长期预报，表明该方法在客观化、定量化地确定气候周期，寻找相似的前期特征等方面能获得令人满意的结果；Tsegaye Tadesse等[65]通过对区域气象、海洋数据及区域干旱历史的数据进行关联规则挖掘，挖掘Nebraska地区的一系列干旱预测指标，提高了该区域的干旱气候预报准确度；Yang等[66]用关联规则挖掘方法对北大西洋飓风历史追踪数据进行分析挖掘，并据此预测北大西洋飓风强度的变化。

当前，已有的气象数据关联规则挖掘研究极大部分都是使用Apriori算法进行。需要指出的是，在实际应用中，Apriori算法暴露了2个缺点：(1)可能产生大量的候选集；(2)可能需要重复扫描数据库。这2个缺点也在一定程度上制约了关联规则挖掘在气象预报中的应用。

综上所述，数据挖掘经过多年的发展，现已有大量经典的数据挖掘算法拓展到气象预报中实践和应用，它们各有优缺点，在不同的问题背景上具有不同的性能和优越性。其中从国内外的相关学术文献分析可知，以感知学习为基础的神经网络和支持向量机是近年国内外气象预报领域研究和应用的最广泛的数据挖掘方法。而关联规则挖掘方法在各种灾害天气的规律分析及其预测研究中应用非常多且具有较强的效能。

2 未来研究重点及展望

传统的统计预报方法在短期内已经很难取得质的飞跃，随着计算机计算速度的大幅提升，数值预报方法应运而生，而且很快从研究领域过渡到了实际应用领域，并成为了气象预报工作中极为重要的方式，然而现有的一些数值预报方法都或多或少存在着缺陷和问题，运行效率和准确率都需要不断提高。另一方面，基于数据挖掘方法的气象预报技术目前是一个非常炙手可热的研究领域，尽管使用数据挖掘方法挖掘气象资料进行气象预报的文献不是太多，但国内外都已经展开了这方面的研究工作，且已经积累了不少优秀成果，在气象预报的诸多方面已取得了突破性的进展，但是很多理论和方法还不够成熟，还有待探究出更多、更有效的气象数据挖掘方法来提高预报能力。随着全球变暖，灾害性天气频发，未来天气变得越来越难以捉摸，更突显这项工作的重要意义和迫切性。纵观当前基于数据挖掘方法的气象预报技术的研究进展情况，认为需要进一步研究的重点和难点问题主要包括：(1)如何加快学习速度，从而减少计算时间代价；(2)如何减少出现过度训练以便有效地提高泛化性能；(3)如何准确、全面的选择预报因子，从而提高预报的准确率；(4)

如何将近年新发展起来的优秀的智能计算方法(如,基因表达式编程算法、蜂群算法等)应用于气象数据挖掘和气象预报中,提高预报的质量;(5)由于气象数据的特殊复杂性,使得传统的数据挖掘中的数据预处理算法难以适应。因此,探究有效的适应复杂气象数据特点的数据预处理算法是可望提高预报质量的重要研究方向;(6)当前在气象预报技术中应用最广泛的数据挖掘方法是以感知学习为核心的神经网络和支持向量机建模预报方法,这些方法都属于隐模型建模方法,可解释性较差,不利于理解和发现新的气象规律知识。因此,探究模型可解释性强的数据挖掘方法也是一个较重要的研究方向。

因此,后续的研究工作可以探究更多优秀的数据挖掘算法来继续开展气象预报方法的研究,也可以探索数据挖掘方法在其他气象领域的研究和应用。此外,随着计算机的计算能力不断得到加强,以及数值预报模式的逐步完善,中长期的气象预报也已逐渐引起人们的关注,如何将基于数据挖掘方法的预报技术应用于中长期的气象预报中也是广大气象学者和相关专家亟待解决的问题。

3　结束语

气候系统是一个耗散的、具有多个不稳定源的高阶非线性系统,其复杂的内部相互作用和自由变化导致了气候的可变性和复杂性。进入 21 世纪以来,全球气候变化异常,如何有效地提高气象预报的精度,提高积极应对各种气象灾害和由其引发的次生灾害的能力,减少人员的伤亡和财产损失,已成为社会关注的焦点,也是当今气象预报领域中的重点和难点。本文以基于数据挖掘方法的气象预报技术的发展现状研究为主,简要介绍了文中所提及的一些数据挖掘方法的相关概念和特点,并对这些数据挖掘方法应用于气象预报工作中的基本原理及优缺点分别进行了论述,最后探讨了当前基于数据挖掘方法的气象预报技术存在的困难和发展趋势。与传统的预报方法相比,基于数据挖掘方法的气象预报技术对解决诸如暴雨、台风等非线性特征明显的天气预报问题有更强的能力,是一门极具前景的实用技术。目前国内外有关基于数据挖掘方法的气象预报技术的研究性论文正日趋增多,但仍有大量的拓展空间,尤其是将关联规则挖掘、聚类挖掘、和异常点挖掘等应用于气象预报中的文献相对较少,希望本文的工作能够对气象工作者及相关领域学者继续深入研究提供有用的参考。

参考文献:

[1] 彭九慧,丁力,杨庆红. 几种降水集成预报方法的对比分析[J]. 气象科技,2008,36(5):520 - 523.

[2] Han J, Kamber M. Data Mining: Concepts &Techniques[M]. San Francisco, CA: Morgan Kaufmann Publishers, 2001. 30 - 33.

[3] 陈宝学,俞经善,关宏伟. 数据挖掘技术应用于天气预报的可行性研究[J]. 应用科技,2004,31(3):48 - 50.

[4] Michael Steinbach, PangNing Tan, Vipin Kumar, et al. Discovery of climate indices using clustering[A]//Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining[C]. 2003.

[5] 马廷淮,穆强. 气象数据挖掘研究[J]. 武汉理工大学学报,2010,32(16):110 - 114.

[6] 董聪,郦正能,夏人伟. 多层前向网络研究进展及若干问题[J]. 力学进展,1995,5(2):186 - 195.

[7] 元昌安,邓松,李文敬,等. 数据挖掘原理与 SPSS Clementine 应用宝典[M]. 北京: 电子工业出版社, 2009. 234 - 267; 86 - 87; 147 - 150; 176.

[8] 金龙,吴建生. 基于遗传算法的神经网络短期预报预测模型[J]. 灾害学,2004,8(1):15 - 16.

[9] 金龙,况雪源,黄海洪. 人工神经网络预报模型的过拟合研究[J]. 气象学报,2004,62(1):62 - 70.

[10] 段文广,周晓军,石永炜. 数据挖掘技术在精细化温度预报中的应用[J]. 干旱气象, 2012,30(1):130 - 135.

[11] Tang Y M, William W Hsieh. Coupling Neural Network to Incomplete Dynamical System via Variational Data Assimilation[J]. Monthly Weather Review, 2000,129(4):818 - 834.

[12] Mihalakakou G, Santamouris M, Asimakopoulos D N. The total solar radiation time series simulation in Athens Using Neural Networks[J]. Theoretical and Applied Climatology,2000,66(3):185 - 197.

[13] Hung N Q, Babel M S, Weesakul S, et al. An artificial neural network model for rainfall forecasting in Bangkok, Thailand[J]. Hydrology and Earth System Sciences, 2008(5):183 - 218.

[14] Enireddy Vamsidhar, Varma K V S R P, Sankara Rao P, et al. Prediction of Rainfall Using Back propagation Neural Network Mode[J]. International Journal on Computer Science and Engineering, 2010,2(4):1119 - 1121.

[15] Chadwick R, Coppola E, Giorgi F. An artificial neural network technique for downscaling GCM outputs to RCM spatial scale[J]. Nonlinear Processes Geophys, 2011,18(6):1013 - 1028.

[16] Gordon Reikard. Combining frequency and time domain models to forecast space weather[J]. Advances in Space Research, 2013(52):622 - 632.

[17] 陈云浩,史培军,李晓兵. 不同热力背景对城市降雨(暴雨)的影响(Ⅲ)——基于人工神经网络的集成预报模型[J]. 自然灾害学报,2001,10(3):26 - 31.

[18] 吴建生,刘丽萍,金龙. 粒子群 - 神经网络集成学习算法气象预报建模研究[J]. 热带气象学报,2008,24(6):679 - 686.

[19] Wu J A. Semi - parametric Regression Ensemble Model for Rainfall Forecasting Based on RBF Neural Network[J]. Lecture Notes in Artificial Intelligence and Computational Intelligence, 2010, 6320(2):284 - 292.

[20] Langella G, Basile A, Bonfante A, et al. High - resolution space - time rainfall analysis using integrated ANN inference systems [J]. Journal of Hydrology, 2010(387): 328 - 342.

[21] Wang L, Wu J. Application of Hybrid RBF Neural Network Ensemble Model Based on Wavelet Support Vector Machine Regression in Rainfall Time Series Forecasting [A]. Fifth International Joint Conference on Computational Sciences and Optimization (CSO) [C]. IEEE, 2012. 867 - 871.

[22] Karthik Nadig, Walter Potter, Gerrit Hoogenboom, et al. Comparison of individual and combined ANN models for prediction of air and dew point temperature [J]. Appl Intell, 2013(39): 354 - 366.

[23] 熊聪聪, 王静, 宋鹏, 等. 遗传算法在多模式集成天气预报中的应用 [J]. 天津科技大学学报, 2008, 23(4): 80 - 84.

[24] 吴清佳, 张庆平, 万健. 遗传神经网络的智能天气预报系统 [J]. 计算机工程, 2005, 31(14): 176 - 177.

[25] 吴建生, 金龙, 汪灵枝. 遗传算法进化设计 BP 神经网络气象预报建模研究 [J]. 热带气象学报, 2006, 22(4): 411 - 416.

[26] Chen L, Yeh K C, Wei H P, et al. An improved genetic programming to SSM/I estimation typhoon precipitation over ocean [J]. Hydrological Processes, 2011, 25(16): 2573 - 2583.

[27] Ihshaish H, Cortés A, Senar M A. Genetic ensemble (G - Ensemble) for meteorological prediction enhancement [A]. Proceedings of The 2011 Internacional Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA2011) [C]. 2011. 404 - 410.

[28] Venkadesh S, Hoogenboom G, Potter W, et al. A genetic algorithm to refine input data selection for air temperature prediction using artificial neural networks [J]. Applied Soft Computing, 2013, 13(5): 2253 - 2260.

[29] Wong K Y, Yip C L, Li P W. Automatic identification of weather systems from numerical weather prediction data using genetic algorithm [J]. Expert Systems with Applications, 2008, 35(1): 542 - 555.

[30] Vapnik V. 统计学习理论的本质 [M]. 张学工译. 北京: 清华大学出版社, 2004.

[31] 韦慧红, 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.

[32] 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.

[33] 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.

[34] 王小萍, 谭季青, 吴书成. 基于支持向量机的预报技巧的预报模型研究 [J]. 科技通报, 2006, 22(6): 747 - 752.

[35] 王在文. 基于非线性支持向量机方法的奥运场馆气象预报 [A]. 第 27 届中国气象学会年会城市气象, 让生活更美好分会场论文集 [C]. 2010.

[36] 樊高峰, 张勇, 柳苗, 等. 基于支持向量机的干旱预测研究 [J]. 中国农业气象, 2011, 32(3): 475 - 478.

[37] 汪春秀. 基于支持向量机的气象预报方法研究 [D]. 南京: 南京信息工程大学, 2011.

[38] 滕卫平, 胡波, 滕舟, 等. SVM 回归法在西太平洋热带气旋路径预报中的应用研究 [J]. 科技通报, 2012, 28(11): 49 - 53.

[39] Hong W C. Rainfall forecasting by technological machine learning models [J]. Applied Mathematics and Computation, 2008, 200(1): 41 - 57.

[40] Radhika Y, Shashi M. Atmospheric temperature prediction using support vector machines [J]. International Journal of Computer Theory and Engineering, 2009, 1(1): 1793 - 8201.

[41] Kisi O, Cimen M. Precipitation forecasting by using wavelet - support vector machine conjunction model [J]. Engineering Applications of Artificial Intelligence, 2012, 25(4): 783 - 792.

[42] Usha Rani R, Dr. T. K Rama Krishna Rao. An Enhanced Support Vector Regression Model for Weather Forecasting [J]. IOSR Journal of Computer Engineering, 2013, 12(2): 21 - 24.

[43] 罗宇智, 陈璟. 贝叶斯网络的气象威胁建模及评估方法研究 [J]. 计算机仿真, 2008, 25(11): 52 - 55.

[44] 何伟, 孔梦荣, 赵海青. 基于贝叶斯分类器的气象预测研究 [J]. 计算机工程与设计, 2007, 28(15): 3780 - 3782.

[45] 陈朝平, 冯汉中, 陈静. 基于贝叶斯方法的四川暴雨集合概率预报产品释用 [J]. 气象, 2010, 36(5): 32 - 39.

[46] 顾锦荣, 焦海军, 朱国涛. 基于贝叶斯网络和支持向量机的天气威胁度模型 [A]. 第 29 届中国气象学会年会分会场: S1 灾害天气研究与预报 [C]. 2012.

[47] 梁莉, 赵琳娜, 齐丹, 等. 基于贝叶斯原理的降水预报偏差订正及水文试验 [A]. 第十四届中国科协年会第 14 分会场: 极端天气事件与公共气象服务发展论坛论文集 [C]. 2012. 1 - 12.

[48] Cano R, Sordo C, Gutiérrez J M. Applications of Bayesian networks in meteorology [M]. Advances in Bayesian networks. Springer Berlin Heidelberg, 2004. 309 - 328.

[49] Jiang Yu, Song Zhe, Kusiak Andrew. Very short - term wind speed forecasting with Bayesian structural break mode [J]. Renewable Energy, 2013, 50: 637 - 647.

[50] Voyant C, Darras C, Muselli M, et al. Bayesian rules and stochastic models for high accuracy prediction of solar radiation [J]. Applied Energy, 2014, 114: 218 - 226.

[51] Valmik B Nikam, Meshram B B. Modeling Rainfall Prediction Using Data Mining Method: A Bayesian Approach [A]. 2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation [C]. IEEE computer society, 2013. 132 - 136.

[52] McLean Sloughter J, Gneiting T, Raftery A E. Probabilistic Wind Vector Forecasting Using Ensembles and Bayesian Model Averaging [J]. Monthly Weather Review, 2013, 141(6): 2107 - 2119.

[53] 张海玲, 过仲阳, 吴建平, 等. 决策树方法在环境物理量场与暴雨之间关系研究中的应用 [J]. 地球信息科学, 2005, 7(4): 39 - 42.

[54] 徐会明, 靳小兵, 季海, 等. 决策树法在雷电潜势预报中的应用 [J]. 高原山地气象研究, 2008, 28(4): 55 - 58.

[55] Prasad N, Kumar P, Mm N. An Approach to Prediction of Precipitation Using Gini Index in SLIQ Decision Tree [A]. 4th International Conference on Intelligent Systems, Modeling & Simulation, Bangkok [C]. 2013. 56 - 60.

[56] Royston S, Lawry J, Horsburgh K. A linguistic decision tree approach to predicting storm surge [J]. Fuzzy Sets and Systems, 2013, 215: 90 - 111.

[57] 万谦,陆建江,宋自林. 正态云关联规则在气象中的应用[J]. 解放军理工大学学报(自然科学版),2002,3(4):1-4.
[58] Guo Z, Dai X, Lin H. Application of Association Rule in Disaster Weather Forecasting[J]. Geographic Information Sciences, 2004, 10(1): 68-72.
[59] 朱冬梅,朱添福,林逢春. 改进 Apriori 算法在气象预报中的应用[J]. 计算机科学,2011,38(7A):111-112.
[60] 林万涛,王建州,张文煜,等. 基于数值模拟和统计分析及智能优化的风速预报系统[J]. 气候与环境研究,2012,17(5):646-658.
[61] 迟德中. 基于数值气象模式和关联规则优化的风电场短期风速预报方法[D]. 兰州: 兰州大学,2012.
[62] 王红霞,李松. 关联规则挖掘在干旱预测中的研究与应用[J]. 微计算机信息(管控一体化),2010,26(4-3):21-23.
[63] 郑忠平. 基于关联规则和聚类分析的异常天气挖掘[D]. 成都: 电子科技大学,2011.
[64] Harms S K, Deogun J S. Sequential association rule mining with time lags[J]. Journal of Intelligent Information Systems, 2004,22(1):7-22.
[65] Tadesse T, Wilhite D A, Harms S K, et al. Drought monitoring using data mining techniques: A case study for Nebraska, USA[J]. Natural Hazards, 2004,33(1):137-159.
[66] Yang R, Tang J, Sun D. Association Rule Data Mining Applications for Atlantic Tropical Cyclone Intensity Changes[J]. Weather & Forecasting,2011,26(3):337-353.

Review of Research on Data Mining in Application of Meteorological Forecasting

PENG Yuzhong[1,2], WANG Qian[2], YUAN Changan[1], LIN Kaiping[3]

(1. *Key Laboratory of Scientific Computing & Intelligent Information Processing in Universities of Guangxi, Guangxi Teachers Education University, Nanning* 530001, *China*; 2. *Key Laboratory of Sustainable Development Monitoring and Optimizing in Beibu Gulf (Guangxi Teachers Education University), Ministry of Education, Nanning* 530001, *China*; 3. *Guangxi Meteorological Observatory, Nanning* 530022, *China*)

Abstract: Meteorological prediction is one of the most important and challenging task in the modern world. In general, climate is highly non-linear and complicated phenomena, which require advanced method and computer modeling for their accurate prediction. This paper provides a survey of available literature of some methodologies employed by different researchers for the weather forecast based on the methods of data mining at home and abroad. It briefly introduced the concepts and traits of some data mining methods in atmospheric science, separately discussed the basic principles of these data mining methods used in weather forecasting, including the advantages and disadvantages. Finally, the author points out some existing difficulties of data mining methods used in weather forecasting, and puts forward key points of future research and current trends of the research.

Key words: data mining; weather forecast; meteorological data mining; weather modeling; computing intelligence