文章编号:1000-582X(2010)12-119-08

新的集成预报及其在短期气候预测中的应用

李学明[1,2],郭尚坤[1],王剑柯[1],高阳华[2]
(1. 重庆大学 计算机学院,重庆 400044;2. 重庆市气象科学研究所,重庆 401147)

摘 要:分析了传统的基于加权的集成预报等方法及其在气象预测应用中存在的问题,在此基础上提出了一种新的基于数据挖掘的集成预报方法,并选用 BP 人工神经网络建立集成预报分类器来对各种子预报方法的预报结果进行集成和综合;该方法可以根据不同预报对象的特性,对集成预报权值进行动态改变,克服了传统的集成预报方法中权值一旦确定就不能改变的不足,也克服了现有的集成预报不能得到最优结果的不足。通过对 2001~2007 年重庆市城口县 1 月的降水和平均气温以及重庆市的春旱指数进行预报,实验结果显示,集成预报结果的可靠性和准确性不但高于集成之前的各种子预报方法,而且高于传统的其它集成预报方法,验证了方法的有效性。

关键词:BP 人工神经网络;数据挖掘;集成预报;气象预测;环流特征
中图分类号: P456 **文献标志码**:A

A new integrating forecast and its application in short-term climate prediction

LI Xue-ming[1,2], *GUO Shang-kun*[1], *WANG Jian-ke*[1], *GAO Yang-hua*[2]
(1. College of Computer Science, Chongqing University, Chongqing 400044. P. R. China;
2. Chongqing Institute of Meteorological Sciences, Chongqing 401147, P. R. China)

Abstract: The existing problems of the traditional weight integrating forecast methods and the application in climate prediction are analyzed. A new method based on data mining is presented, which uses BP artificial neural network to build the integrating forecast classifier to integrate the forecast results of sub-methods. According to the features of different forecast objects, this method can change weight dynamically, which overcomes the shortage of the traditional weight integrating forecasts that cannot change weight after been decided and overcomes the shortage that cannot get the optimal results. By predicting the precipitation and average temperature of Chengkou County in January, and spring drought index of Chongqing from 2001 to 2007, the experiment results show that the reliability and accuracy of the proposed model are better than those of the sub-methods and other integrating forecast methods, which proves the effectiveness of this method.

Key words: BP artificial neural network; data mining; integrating forecast; meteorological prediction; circulation features

---

收稿日期:2010-06-12
基金项目:国家科技支撑计划重大资助项目(2007BAC03A06);科技部农业成果转化资助项目(2007GB24160446);中国气象局新技术资助项目(CMATG2009MS21);重庆市重大科技攻关资助项目(CSTC2009AB2221)。
作者简介:李学明(1967-),男,重庆大学副教授,主要从事数据挖掘、信息安全、网络计算等方向研究,(E-mail)lixuemin@cqu.edu.cn。

目前，在对天气现象或气象要素进行预报时，通常使用动力学、统计学和天气学等几种预报方法，常用的有 BP 人工神经网络、多元回归、均生函数、最优气候均态模型等方法。在实际应用中，对每个具体的问题，各种预报方法得出的结果通常是不一致的，因而不知道如何将它们统一起来[1]。一般地，每个具体预报方法的预报思想不同，其适应的具体环境也就不同，得到的预报结果的准确程度也不相同，对某类数据有较好预报结果的方法，对其他数据不一定有较好结果，例如，在气象的旱涝预测中，有时用神经网络能得到较好的结果，有时用最优气象模型能得到较好的结果。因此需要采用一种较好的处理方法，将各种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 1 个优于单一预报方法的预报结论，这就是预报方法的集成问题。

随着统计预报方法的不断发展，集成预报可以比集成的各子预报方法获得更好的预报效果，这已越来越被人们所认识和承认。面对众多的预报方法，为了更好地得出最终的预报结果，研究、探讨各种子预报方法的集成预报新方法对实际业务工作具有重要意义。近年来，对于集成预报方法的研究，国内外的气象学者作了大量的工作[1-2]。目前较为常用的集成预报方法除了回归、平均、多数表决和加权集成方法、采用概率回归和典型相关等集成预报方法以外，也有用人工神经网络等其他集成方法[3-4]。然而这些集成预报方法本质上都属于加权集成方法，这些权值一旦确定，在进行预报时就无法调整；在进行实际预报时，这类方法没有考虑到被预报对象的差异性，也没有充分考虑到不同预报方法的适应性。

为此，提出了基于数据挖掘思想的集成预报方法，理论分析和实验都表明，该方法克服了现有一些集成预报方法的不足，取得了较好的结果。

1 提出问题

在实际的预报工作中，随着预报的气象要素的地理位置变化、预报样本自身的差异，并没有哪一种子预报方法得出的预报结果永远是最准确的；即便是对于同一地理位置同一气象要素的预报，2 次不同时间的预报，其预报最准确的子预报方法也极可能是不同的。所以没有任何一种预报方法在任何情况下的预报结果都是最准确的，最理想的集成预报方法应该能够识别出最优子预报方法并将其预报结果作为自身的预报结果。

在传统的基于加权的集成预报模型中，对于各种子预报方法都只是给予一定的权值，这些权值一旦确定，在进行预报时就无法调整。而预报方法的本质是利用过去的知识来对未来的发展进行预测，因此基于权值的集成预报方法在预报时由于权值无法根据预报对象进行动态调整，无法考虑预报对象的特性。例如预报不同年份的降水，其采用的集成预报权值都是不变的，没有考虑到不同年份的差异性。下面分析这样的基于加权的集成预报模型存在的问题。

设 P 是集成预报结果，E 是集成预报的绝对误差；P_i 是第 i 个子预报方法的预报结果，E_i 是第 i 个子预报方法的绝对误差，w_i 是第 i 个子预报方法的权值；n 是子预报方法的数量。加权集成预报的模型为

$$P = \sum_{i=1}^{n} w_i P_i ; \tag{1}$$

$$E = \sum_{i=1}^{n} w_i E_i ; \tag{2}$$

$$\sum_{i=1}^{n} w_i = 1 ; \tag{3}$$

设 P_j 是最优的子预报方法的预报结果，E_j 是最优的子预报方法的绝对误差，容易得到子预报方法中最小的绝对误差 $E_{\min}=E_j$。

$$\begin{aligned} E - E_{\min} &= \sum_{i=1}^{n} w_i E_i - E_{\min} \sum_{i=1}^{n} w_i = \\ &\sum_{i=1}^{n} w_i E_i - \sum_{i=1}^{n} w_i E_{\min} = \\ &\sum_{i=1}^{n} w_i (E_i - E_{\min}) = \\ &w_1 (E_1 - E_{\min}) + \cdots + w_n (E_n - E_{\min}) \geqslant 0 \end{aligned}$$

只有在 $E_1=E_2=E_3=\cdots=E_n=E_{\min}$ 时，等号成立，即所有子预报方法的预报结果都一致。而在实际的预报工作中，所有子预报方法几乎是不可能全部一致，即 $E>E_{\min}$。

在基于加权的集成预报模型中，当 $E_i>E_{\min}$，且 $w_i\neq 0$，集成预报的绝对误差 E 必大于 $E_{\min}$，得到的集成预报结果 P 也就不是最优的，而最希望得到的是最优子预报方法的预报结果 P_j。

经过以上分析，可以看出在所有子预报方法的预报结果出现不一致的时候，基于加权的集成预报方法得到的预报结果并不是最优和最准确的。产生上述结果的原因在于，每次进行预报时，由于不知道哪个子预报方法最好，因此只能把这些子预报方法的预报结果按照以前的经验所确定的权重进行加权求和。这个问题就变为：在对 1 个天气现象或气象要素每次进行预报时，能否根据以前的经验找到一个最合适或最佳的子预报方法来进行本次的预报呢？如果能，就不需要对各个子预报结果进行加权求和。该问题的答案显然是可以的，这就是提出的基于数据挖掘方法的出发点。其核心思想在于该方法考虑了预报的差异性，能够根据预报对象的不同，而采用权值可变的集成方法，即不同预报对象的集

成预报权值是不同的；例如预报不同年份的降水，方法能够根据不同预报年份的差异性以及预报年份与以往数据的关联度，利用关联度最高的以往年份的预报方法来对预报年份的降水进行预报。

方法的大致思想如下：首先利用数据挖掘中的分类方法来构建 1 个选择最优子预报方法的分类器，其次，在对某一天气现象或气象要素进行预报时，利用该分类器选出 1 个最优的子预报方法；最后，把该子预报方法的预报结果作为集成预报方法的预报结果。

2　集成预报模型

2.1　资料

实验中各种算法所使用的输入数据是 74 项环流特征量资料（北半球副高面积指数、北非副高面积指数、北半球副高强度指数等），是由国家气候中心气候系统诊断预测室再处理资料，起始时间是 1951-2007 年，资料数据全都为整型。

实验中各种算法的预测对象有：降水、平均气温、5 种干旱指数（春旱、夏旱、伏旱、秋旱、冬旱）和洪涝指数。降水、平均气温等要素的数据来自于重庆 34 个地面气象观测站的逐日观测资料（1971-2007 年），各种干旱指数和洪涝指数来自于重庆市旱涝灾害监测预警决策服务系统的计算结果[5-7]。

2.2　预报因子和预报量

预测工作包括 2 种类型：年度预测和季度预测。

年度预测将从在每年 11 月的基础上向前推算 3 个月再提前 6 个月的环流特征组成 6×74 个环流特征序列中选取，季度预测将在预测月份之前 3 个月的基础上再提前 6 个月的环流特征组成 6×74 个环流特征序列中选择；然后将这 6×74 个环流特征序列分别与对应的目标输出序列求相关系数，最后取相关系数绝对值最大的 N 个环流特征序列作为各种算法的输入数据[6]。

2.3　子预报方法

第一种子预报方法是 BP 人工神经网络。BP 是目前应用最多的神经网络，这主要归结于 BP 算法的多层感知器具有非线性映射能力，只要能提供足够多的样本模式对供 BP 网络进行学习训练，它便能完成由 n 维输入空间到 m 维输出空间的非线性映射；BP 算法还有很强的泛化能力和容错能力[8-10]。

第二种子预报方法是多元回归。回归分析是一种处理变量的统计相关关系的一种数理统计方法，它的基本思想是：虽然自变量和因变量之间没有严格的、确定的函数关系，但可以设法找出最能代表它们之间关系的数学表达形式。多元回归分析是研究多个变量之间关系的回归分析方法[11]。

第三种子预报方法是均生函数。当一要素未来状况难以与现有的其它因素建立联系时，用历史资料的自身变化规律来预测未来一定时段这一要素的状况是一种有效方法。均生函数是原系列生成的、体现各种长度周期的基函数，在基函数的基础上，相继给出了几种适于不同类型系列的建模方案。均生函数预测模型既可以作多步预测，又可以较好地预测极值。均生函数模型是借助多元分析的手段，解决时间系列预测问题的一种尝试[12]。

第四种子预报方法是最优气候均态模型。最优气候均态法是美国气候中心用于制作温度预报的一种方法，它是用最近 K 年的要素平均值作为下一年的预报值，制作简便。最优气候均态法被定义为最近 K 年的要素平均。选取 K 年的标准是，用 K 年平均值作为下一年的预报值能得到 1 个最好的预报。世界气象组织推荐最近 30 年的平均作为气候均值。如何确定最佳的 K，是此方法的关键[13]。

2.4　其他集成预报

回归、平均、多数表决集成预报方法在本质上是属于加权集成方法，只是确定权值的方式不同而已。加权集成预报的模型在公式(1)～(3)已经阐述。

确定权值的基本原则是预报效果好的方法给予较高的权值。

文献 3 中基于人工神经网络的集成预报方法进行训练时的输入数据是子预报方法的预报值，期望输出预报对象的实际值；在进行预测时的输入数据是子预报方法的预报值，得到的输出是集成预报的结果。而提出的基于数据挖掘的集成预报方法进行训练时的输入数据是各项环流因子，期望输出是最优的子预报方法；在进行预测时的输入数据是各项环流因子，得到的输出是集成预报模型认可的最优子预报方法。

经过分析，可以得出文献 3 中基于人工神经网络的集成预报方法其本质仍然是加权集成预报。只是确定权值的方法是不断地进行学习训练，权值存在于复杂的人工神经网络结构中，并不是线性和容易理解的。

2.5　一种新的基于数据挖掘的集成预报

系统框架如图 1。用数据挖掘中数据分类的方法构建 1 个选择最优子预报方法的集成预报分类器；每次预报某一天气现象或气象要素，这个分类器能够选择出 1 个最优的子预报方法，然后将该子预报方法的预报结果作为集成预报方法的预报结果。构建设想的集成预报分类器的数据挖掘方法有决策树方法、统计方法、人工神经网络方法等，选择具有非线性准动力系统特征的 BP 人工神经网络方法，其多层感知器具有非线性映射能力，只要能提供足够多的样本模式对供 BP 网络进行学习训练，它便能完成由 n 维输入空间到 m 维输出空间的非线性映射。

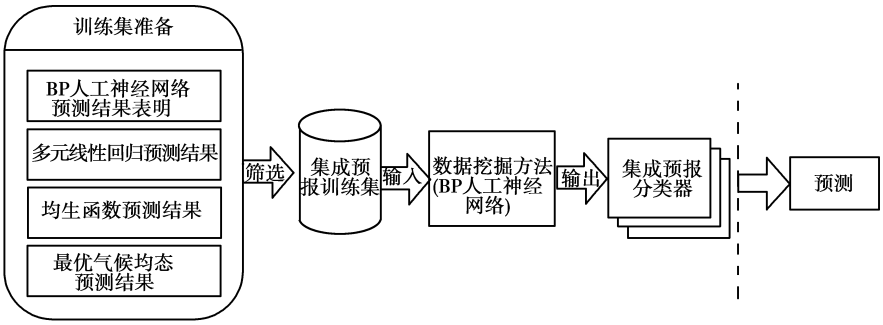

图 1 系统框架

基于数据挖掘的集成预报方法过程如下

1) 准备用于构建集成预报分类器的训练集；

用 1971-1986 年的数据作为 4 种子预报方法的训练集，然后用 4 种子预报方法去预测 1986-2000 年的预测对象值；从 4 种子预报方法的预报结果中可以筛选出相应的最优子预报方法，并将 1986-2000 年的输入数据以相应的最优子预报方法标记，筛选出的输入数据（相关最大的 N 个环流特征）和期望输出数据（最优子预报方法）作为集成预报分类器的训练集；表 1 是重庆市城口县 1 月降水训练集的准备过程。

表 1 集成预报分类器的训练集的准备过程

输入（相关最大的 N 个环流特征）	4 种子预报方法的预报结果	筛选	期望输出
X_1(1986)	$E_{bp}=0.189, E_{mr}=0.131, E_{mgf}=0.26, E_{ocn}=0.191$	$E_{min}=E_{mr}$	2
X_2(1987)	$E_{bp}=0.125, E_{mr}=0.32, E_{mgf}=0.036, E_{ocn}=0.155$	$E_{min}=E_{mgf}$	3
……	…	…	…
X_{15}(2000)	$E_{bp}=0.203, E_{mr}=0.104, E_{mgf}=0.043, E_{ocn}=0.264$	$E_{min}=E_{mgf}$	3

E_{bp}是 BP 人工神经网络预报的绝对误差，E_{mr}是多元回归预报的绝对误差，E_{mgf}是均生函数预报的绝对误差，E_{ocn}是最优气候均态模型预报的绝对误差；期望输出的数值表示最优子预报方法（1 是 BP 人工神经网络，2 是多元回归，3 是均生函数，4 是最优气候均态模型）。

筛选出的输入数据（相关最大的 N 个环流特征）和期望输出数据（最优子预报方法）作为集成预报分类器的训练集。

2) 用数据挖掘的方法训练得到集成预报分类器；

集成预报分类器利用筛选的训练集进行训练，得到集成预报分类器，该集成预报分类器可以根据环流特征的输入，直接得到一种最优子预报方法，然后利用得到的最优子预报方法去预测，将最优子预报方法的预测结果作为集成预报的预测结果；

3) 利用训练好的分类器进行独立预测。

用 2001-2007 年的数据去评估集成预报分类器和四种子预报方法和其它集成预报方法的预报准确性。

2.6 基本原理

BP 人工神经网络（back propagation，误差反传网络）是一种多层前馈网络，即信息处理的方向是从输入层到各隐层再到输出层逐层进行的，采用最小均方差学习方式，是应用最广泛的神经网络，是一种有导师监督的学习网络[8-10]。

学习过程由信号的正向传播和误差的反向传播 2 个过程组成。正向传播时，输入样本从输入层传入，经各隐层逐层处理后，传向输出层。若输出层的实际输出与期望的输出（教师信号）不符，则转向误差的反向传播阶段。误差反传是将输出误差以某种形式通过隐层向输入层逐层反传，并将误差分摊给各层的所有单元，从而获得各层单元的误差信号，此误差信号即作为修正各单元权值的依据。这种信号的正向传播与误差反向传播的各层权值调整过程是周而复始进行的。权值不断调整的过程，也就是网络的学习训练过程。此过程一直进行到网络输出的误差减少到可接受的程度，或者进行到预先设定的学习次数为止[14-16]。

尽管 BP 神经网络的研究与应用已取得巨大的成功，但是在网络的开发设计方面至今还没有 1 套完整的理论作为指导。应用中采用的主要设计方法是，在充分了解待解决问题的基础上将经验与试探相结合，通过多次改进性实验，最终选出 1 个比较好的设计方案[17-18]。

BP 神经网络的模型参数有：输入层节点数、隐藏层节点数、输出层节点数、最大训练次数、收敛误差、学习因子和动量系数。

在实验中，输入层节点取值为 10，隐藏层节点取值范围为[1，10]，最大训练次数取值范围为(0，

2000],收敛误差取值范围为(0,1],学习因子取值范围为(0,1],动量系数取值范围为[0,1]。

3　集成预报方法的对比分析

3.1　新的集成预报与各子预报方法的比较

为了进行对比分析,选用以下 2 种统计评价指标:

1)平均绝对误差 $MAE=\frac{1}{n}\sum_{i=1}^{n}|P_i-\hat{P}_i|$,其中 P_i 是预报值,$\hat{P}_i$ 是实际值。

2) 预报结果最优百分比

3.1.1　预报降水

表 2 给出了基于数据挖掘的集成预报模型和各个子预报方法在 2001-2007 年对重庆市城口县 1 月的降水的预报结果(年度预测)。

表 2　城口县 1 月降水预报结果(与子预报方法的对比)

年份	实际值	BP 人工神经网络		多元回归		均生函数		最优气候均态		数据挖掘集成预报	
		预报值	绝对误差	预报值	绝对误差	预报值	绝对误差	预报值	绝对误差	预报值	绝对误差
2001	0.5	0.491	0.009	0.526	0.026	0.451	0.049	0.500	0	0.500	0
2002	0.3	0.471	0.171	0.781	0.481	0.334	0.034	0.473	0.173	0.334	0.034
2003	0.1	0.449	0.349	0.600	0.005	0.701	0.601	0.427	0.327	0.427	0.327
2004	0.1	0.434	0.334	0.861	0.761	0.147	0.047	0.373	0.273	0.147	0.047
2005	0.3	0.419	0.119	0.527	0.227	0.317	0.017	0.318	0.018	0.317	0.017
2006	0.3	0.414	0.114	0.842	0.542	0.601	0.301	0.309	0.009	0.309	0.009
2007	0.6	0.41	0.19	0.743	0.143	0.396	0.204	0.282	0.318	0.396	0.204
平均绝对误差			0.184		0.383		0.179		0.16		0.091
最优百分比			0/7		1/7		3/7		3/7		6/7

由表 2 的比较可以看到,基于数据挖掘的集成预报 2 种统计指标明显优于各个子预报方法。基于数据挖掘的集成预报与 4 种子预报方法相比,平均绝对误差分别减少了 0.093,0.292,0.088 和0.069,减少的百分比分别为 51%,76%,49%,43%;4 种子预报方法预报结果最优百分比分别为 0,14.3%,42.9%,42.9%,而基于数据挖掘的集成预报为 85.7%,有明显的提高。

如在进行 2001 年预报时,方法得出权值是(0,0,0,1),而在进行 2002 年预报时,方法得出的权值是(0,0,1,0),说明法在进行集成预报时能够动态改变权值,适应不同预报年份的差异性,克服了传统的基于加权的集成预报方法权值一旦确定就无法改变的不足。

3.1.2　预报平均气温

表 3 给出了基于数据挖掘的集成预报模型和各个子预报方法在 2001-2007 年对重庆市城口县 1 月的平均气温的预报结果(年度预测)。

表 3　城口县 1 月平均气温预报结果(与子预报方法的对比)

年份	实际值	BP 人工神经网络		多元回归		均生函数		最优气候均态		数据挖掘集成预报	
		预报值	绝对误差	预报值	绝对误差	预报值	绝对误差	预报值	绝对误差	预报值	绝对误差
2001	3.8	3.061	0.739	2.295	1.505	2.105	1.695	3.257	0.543	3.257	0.543
2002	4.9	3.094	1.806	2.009	2.891	4.026	0.874	3.200	1.700	4.026	0.874
2003	4.0	3.170	0.830	2.080	1.920	2.552	1.448	3.279	0.721	3.279	0.721
2004	3.4	3.221	0.179	−0.487	3.887	1.516	1.884	3.393	0.007	3.393	0.007
2005	3.3	3.226	0.074	2.465	0.835	3.026	0.274	3.400	0.100	3.026	0.274
2006	3.8	3.231	0.569	2.498	1.302	3.921	0.121	3.393	0.407	3.921	0.121
2007	3.3	3.247	0.053	3.155	0.145	2.923	0.377	3.407	0.107	2.923	0.377
平均绝对误差			0.607		1.784		0.953		0.512		0.417
最优百分比			2/7		0/7		2/7		3/7		5/7

由表3的比较可以看到,基于数据挖掘的集成预报2种统计指标明显优于各个子预报方法。基于数据挖掘的集成预报与4种预报方法相比,平均绝对误差分别减少了0.19,1.367,0.536和0.095,减少的百分比分别为31%,77%,56%,19%;4种子预报方法预报结果最优百分比分别为28.6%,0%,28.6%,42.9%,而基于数据挖掘的集成预报为71.4%,有明显的提高。

3.1.3 预报春旱指数

表4给出了基于数据挖掘的集成预报模型和各个子预报方法在2001-2007年对重庆市春旱指数的预报结果(年度预测)。

表4 重庆市春旱指数预测结果(与子预报方法的对比)

年份	实际值	BP人工神经网络		多元回归		均生函数		最优气候均态		数据挖掘集成预报	
		预报值	绝对误差	预报值	绝对误差	预报值	绝对误差	预报值	绝对误差	预报值	绝对误差
2001	0.559	0.483	0.076	−0.046	0.605	0.084	0.475	0.415	0.144	0.415	0.144
2002	0	0.425	0.425	0.550	0.550	0.581	0.581	0.518	0.518	0.518	0.518
2003	0.603	0.421	0.182	0.542	0.061	0.513	0.090	0.346	0.257	0.542	0.061
2004	0.118	0.318	0.200	0.810	0.692	0.486	0.368	0.382	0.264	0.318	0.200
2005	0	0.412	0.412	0.791	0.791	0.222	0.222	0.320	0.320	0.222	0.222
2006	0.029	0.086	0.057	0.633	0.604	0.343	0.314	0.180	0.151	0.086	0.057
2007	0.368	0.153	0.215	1.121	0.753	0.862	0.494	0.188	0.180	0.188	0.180
平均绝对误差			0.224		0.579		0.363		0.262		0.197
最优百分比			4/7		1/7		1/7		1/7		5/7

由表4的比较可以看到,基于数据挖掘的集成预报2种统计指标均优于各个子预报方法。基于数据挖掘的集成预报与4种子预报方法相比,平均绝对误差分别减少了0.027,0.382,0.166和0.065,减少的百分比分别为12%,66%,46%,33%;4种子预报方法预报结果最优百分比分别为57.1%,14.3%,14.3%,14.3%,而基于数据挖掘的集成预报为71.4%,有明显的提高。

3.2 新的集成预报与其它集成预报方法的比较

3.2.1 预报降水

表5给出了基于数据挖掘的集成预报模型和其它集成预报方法在2001-2007年对城口1月的降水的预报结果(年度预测)。

表5 城口1月降水预报结果(与其它集成预报方法的对比)

年份	实际值	加权集成		人工神经网络集成		数据挖掘集成预报	
		预报值	绝对误差	预报值	绝对误差	预报值	绝对误差
2001	0.5	0.492	0.008	0.485	0.015	0.500	0
2002	0.3	0.512	0.212	0.483	0.183	0.334	0.034
2003	0.1	0.544	0.444	0.475	0.375	0.427	0.327
2004	0.1	0.450	0.350	0.447	0.347	0.147	0.047
2005	0.3	0.394	0.094	0.430	0.130	0.317	0.017
2006	0.3	0.540	0.240	0.415	0.115	0.309	0.009
2007	0.6	0.455	0.145	0.411	0.189	0.396	0.204
平均绝对误差			0.213		0.193		0.091

由表5可以看出基于数据挖掘的集成预报的预报结果优于其它集成预报方法,平均绝对误差比加权集成减少了0.122,减少的百分比为57%;比人工神经网络集成减少了0.102,减少的百分比为53%。

3.2.2　预报平均气温

表 6 给出了基于数据挖掘的集成预报模型和其它集成预报方法在 2001-2007 年对城口 1 月的平均气温的预报结果(年度预测)。

表 6　城口 1 月平均预报结果(与其它集成预报方法的对比)

年份	实际值	加权集成		人工神经网络集成		数据挖掘集成预报	
		预报值	绝对误差	预报值	绝对误差	预报值	绝对误差
2001	3.8	2.716	1.084	3.235	0.565	3.257	0.543
2002	4.9	3.136	1.764	3.269	1.631	4.026	0.874
2003	4.0	2.821	1.179	3.381	0.619	3.279	0.721
2004	3.4	2.080	1.320	3.415	0.015	3.393	0.007
2005	3.3	3.067	0.233	3.448	0.148	3.026	0.274
2006	3.8	3.299	0.501	3.431	0.369	3.921	0.121
2007	3.3	3.190	0.110	3.458	0.158	2.923	0.377
平均绝对误差			0.884		0.501		0.417

由表 6 可以看出基于数据挖掘的集成预报的预报结果优于其它集成预报方法,平均绝对误差比加权集成减少了 0.467,减少的百分比为 52.8%;比人工神经网络集成减少了 0.084,减少的百分比为 20.1%。

3.2.3　预报春旱指数

表 7 给出了基于数据挖掘的集成预报模型和其它集成预报方法在 2001-2007 年对重庆市春旱指数的预报结果(年度预测)。

表 7　重庆市春旱指数预测结果(与其它集成预报方法的对比)

年份	实际值	加权集成		人工神经网络集成		数据挖掘集成预报	
		预报值	绝对误差	预报值	绝对误差	预报值	绝对误差
2001	0.559	0.245	0.314	0.455	0.104	0.415	0.144
2002	0	0.517	0.517	0.198	0.198	0.518	0.518
2003	0.603	0.452	0.151	0.194	0.409	0.542	0.061
2004	0.118	0.488	0.370	0.380	0.262	0.318	0.200
2005	0	0.425	0.425	0.499	0.499	0.222	0.222
2006	0.029	0.299	0.270	0.587	0.558	0.086	0.057
2007	0.368	0.559	0.191	0.248	0.120	0.188	0.18
平均绝对误差			0.320		0.307		0.197

由表 7 可以看出基于数据挖掘的集成预报的预报结果优于其它集成预报方法,平均绝对误差比加权集成减少了 0.123,减少的百分比为 38%;比人工神经网络集成减少了 0.11,减少的百分比为 36%。

4　结　论

分析了传统的基于加权的集成预报方法及其在气象预测应用中存在的问题,在此基础上提出了一种新的基于数据挖掘的集成预报方法,该方法选用 BP 人工神经网络建立集成预报分类器,对 4 种子预报方法的预报结果进行集成和综合。基于数据挖掘的集成预报分类器利用从子预报方法中筛选的训练集进行训练,得到集成预报分类器,该集成预报分类器可以根据环流特征的输入,直接得到一种最优子预报方法,然后利用得到的最优子预报方法去预测,将最优子预报方法的预报结果作为集成预报的预报结果。选择 BP 人工神经网络作为集成预报分类器基于以下 2 个主要理由:其他的诸如决策树等方法需要对数据进行离散化处理;另外,理论表明,BP 等人工神经网络能够很好逼近任意的非线性函数。

结果对比显示，基于数据挖掘的集成预报模型预测的可靠性和准确性不但高于集成之前的各种子预报方法，而且高于其它集成预报方法，解决了传统的基于加权的集成预报方法因考虑无益的预报结果而降低预报的准确性和可靠性的问题。说明本方法是有效的，作者深信所提出的方法对其他类似应用是有参考价值的。

参考文献：

[1] 施能. 几个预报的统一[J]. 气象科技，1983，(6)：26-30.
BARPOB H A. The unification of several predictions [J]. Meteorological Science and Technology，1983 (6)：26-30.

[2] 施能. 气象科研与预报中的多元分析方法[M]. 北京：气象出版社，2002.

[3] 金龙，陈宁，林振山. 基于人工神经网络的集成预报方法研究和比较[J]. 气象学报，1999，57(2)：198-207.
JIN LONG，CHEN NING，LIN ZHEN-SHAN. Study and comparison of ensemble forecasting based on artificial neural network [J]. Acta Meteorological Sinica，1999，57(2)：198-207.

[4] 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.
PENG JIU-HUI，DING LI，YANG QING-HONG. Comparative analysis of several consensus precipitation forecasting methods [J]. Meteorological Science and Technology，2008，36(5)：520-523.

[5] 刘同明，夏祖勋，解洪成. 数据融合技术及其应用[M]. 北京：国防工业出版社，1998.

[6] 黄士松. 决定大气环流的基本因子[J]. 气象学报，1955，26(1/2)：35-64.
HUANG SHI-SONG. Basis factors determining the main features of the general circulation of the atmosphere [J]. Acta Meteorological Sinica，1955，26(1/2)：35-64.

[7] 高阳华，唐云辉，冉荣生，等. 重庆市干旱的分类与指标[J]. 贵州气象，2001，25(6)：16-18.
GAO YANG-HUA，TANG YUN-HUI，RAN RONG-SHENG，et al. The classification and index of the drought in Chongqing [J]. Journal of Guizhou Meteorology，2001，25(6)：16-18.

[8] 欧钊荣，谭宗琨，何燕，等. BP神经网络模型在广西原料蔗产量预报中的应用[J]. 中国农业气象，2007，29(2)：213-216.
OU ZHAO-RONG，TAN ZONG-KUN，HE YAN，et al. Application of BP neural network in yield predication of sugarcane in Guangxi province [J]. Chinese Journal of Agrometeorology，2007，29(2)：213-216.

[9] 韩力群. 人工神经网络理论、设计及应用[M]. 北京：化学工业出版社，2007.

[10] 张立明. 人工神经网络的模型及其应用[M]. 上海：复旦大学出版社，1993.

[11] 宋莹，邓甦. 多元线性回归分析在固定资产投资中的应用[J]. 沈阳师范大学学报，2008，26(2)：160-162.
SONG YING，DENG SU. Multiple linear regression analysis applied to fixed assets investment [J]. Journal of Shenyang Normal University，2008，26(2)：160-162.

[12] 袁本荷. 利用均生函数预测模型作降水预报[J]. 四川气象，2005，25(3)：8-9.
YUAN BEN-HE. Precipitation prediction by using mean generating function [J]. Journal of Sichuan Meteorology，2005，25(3)：8-9.

[13] 武文辉. 最优气候均态模型在贵州月平均温度预报中的应用[J]. 贵州气象，2002，26(6)：33-34.
WU WEN-HUI. The Application of the optimal climate model in the prediction of month average temperature in Guizhou[J]. Meteorology，Journal of Guizhou，2002，26(6)：33-34.

[14] JIN L，LUO Y，LIN Z S. Study on mixed model of neural network for farmland flood /drought prediction [J]. Acta Meteorological Sinica，1997，11 (3)：364-373.

[15] LESLIE L M，HOLLAND G J. Predicting regional forecast skill using single and ensemble forecast techniques [J]. Monthly Weather Review，1991，119(2)：425-435.

[16] KUNG C，SHARIF A. Ling-rang forecasting of Indian summer monsoon onset and rainfall with upper air parameters and surface temperature[J]. J Meteor Soc Japan，1982，60(2)：672-681.

[17] HAN J，KAMBER M. Data mining：concepts and techniques [M]. USA：Morgan Kaufmann，2001.

[18] FAYYAD U M. Data mining and knowledge discovery：making sense out of data [J]. IEEE Expert，1996，11(5)：20-25.

（编辑 侯 湘）