文章编号:1671 - 6833(2007)01 - 0114 - 04

AdaBoost - NN 模型在浊漳河水质评价中的应用

何争光,孙晓峰,马勇光

(郑州大学 环境与水利工程学院,河南 郑州 450001)

摘 要:为了克服传统 BP 网络的不足,将 AdaBoost 与神经网络结合,提出了基于 AdaBoost - NN 的水质评价模型.利用浊漳河水质监测数据比较 AdaBoost - NN 模型与传统 ANN 法和内梅罗综合指数法评价模型的差异,结果表明:AdaBoost - NN 水质评价模型有效弥补了 BP 模型自身的缺陷,评价准确度更高,结果更加客观、合理.

关键词:水质评价;BP 网络;AdaBoost;泛化

中图分类号:X 824 **文献标识码**:A

0 引言

早期的水质综合评价主要有模糊数学法、灰色系统理论法、数理统计法、物元可拓法等,在这些方法中,多数需要设计评价指标对各级标准的隶属函数及各指标的权重,评价结果受主观因素影响较大[1,2].

20 世纪 80 年代,人工神经网络(Artificial Neural Network,ANN)迅速崛起,在模式识别领域有着广泛的应用,近年来也被大量应用于水环境质量评价中.这种基于"数据驱动"的建模方法省去了建模过程中人为因素带来的影响,评价结果客观,精度也高,与传统水质评价方法相比,ANN 方法体现出了巨大的优越性[3].目前运用较多的是基于误差反向传播算法的前馈型 BP(Back Propagation)网络模型,这类方法在河流水质、湖泊水体富营养化程度及地下水水质评价方面都有所应用[4,5].但是,BP 网络存在收敛速度较慢、稳定性差、易陷入局部极小等局限性[6,7].

AdaBoost 算法能够提高任意给定弱分类器的分类精度,在许多机器学习问题中都取得了成功应用,尤其是应用于决策树[8,9].因此,针对 BP 网络自身的局限性和训练样本选择的主观因素,为进一步提高分类精度,使评价结果更加可靠,将 AdaBoost 与神经网络相结合,建立 AdaBoost 结合神经网络(AdaBoost - NN)水质评价模型.该模型采用 BP 神经网络作为弱分类器,通过对其进行训练,产生一个弱分类器序列,利用 AdaBoost 的提升作用,对序列中较精确的分类器给予较高的影响.笔者以某水质资料为例,将模型评价结果与传统的 ANN 方法和内梅罗综合指数法进行比较.

1 AdaBoost - NN 算法介绍

AdaBoost - NN 算法即利用 AdaBoost 算法提升神经网络分类器的性能,作者采用 BP 网络作为待提升的神经网络分类器.

1.1 BP 神经网络

BP 神经网络是前馈型神经网络中研究最为成熟且应用最广的一种网络[10],由一个输入层、一个或多个隐层和一个输出层组成.每一层都包含不同数量的神经元,同一层节点间互不相连,相邻层节点间单向全互连.隐层为一层的 BP 网络结构如图 1 所示.网络共分为 3 层:i 为输入层节点,j 为隐层节点,k 为输出层节点.

BP 学习算法根据给定的学习样本对进行学习,通过调整网络连接权来体现学习的效果.在学习阶段,先将学习样本对的输入加在网络的输入端,沿着前向在各层神经元按输入和激励函数的方式产生输出.定义网络的学习误差函数为

$$E = \frac{1}{2}\sum_k (d_k - y_k)^2 \qquad (1)$$

式中:d_k 为网络的期望输出,y_k 为网络的实际输出.然后将学习误差逆向传播到各层神经元,根据误差的大小和符号相应地调整各连接权值.此过程一直进行到神经网络权连接方式能在给定输入

收稿日期:2006 - 10 - 29;**修订日期**:2006 - 12 - 11

作者简介:何争光(1963 -),男,河南孟州人,郑州大学教授,博士,主要从事水污染控制理论与技术等方面的研究工作.

样本条件下以一定精度产生输出为止. 由于网络中 Sigmoid 函数的存在,BP 算法易陷入局部极小,泛化能力受到限制,制约了模型的性能,因此需要寻找合适的方法弥补 BP 算法的这些不足.

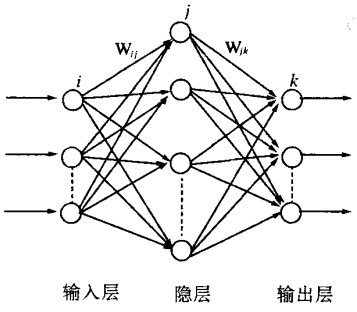

图 1　隐层为一层的 BP 网络模型结构

Fig. 1　BP structure with one hidden

1.2　AdaBoost 算法

1995 年 Freund 和 Schapire 提出了实用性更强,通过调整权重而运作的 AdaBoost 算法[9],解决了早期 Boosting 算法很多实践上的问题. 目前,它已成为 Boosting 算法家族的代表算法.

使用 Boosting 算法是为了根据给定的训练样本(x,y),找到假设 $H(x)$ 算法的输入为训练集 $S=[(x_1,y_1),\cdots,(x_N,y_N)]$. 每个成员$(x_i,y_i)$都是带期望类别号的训练样本,$x_i\in X, y_i\in Y$,$X$ 为领域或实例空间,Y 为类别标号集. 学习器接受的例子是从分布 P 的 $X\times Y$ 上随机选择出来的. 假定要学习的是两类问题,$Y=\{-1,+1\}$,它是多类问题扩展的基础. AdaBoost 反复调用给定的弱学习算法,其主要思想是:在训练集中维护一套权重分布,在第 $t(t=1,\cdots,T,T$ 为迭代次数)次迭代时样本$\{X_i,Y_i\}$上的分布权值记为 $D_t(i)$. 初始时,所有例子的权重都设为相等 $1/N$,但是每一次错分的实例其权重将增加,以使弱学习器被迫集中在训练集的难点上. 弱学习器的任务就是根据分布 D_t 找到合适的弱假设 $h_t: X\to R$. 最简单的情况下每个 h_t 的范围是二值的:$\{-1,+1\}$. 于是,该学习器的任务就是最小化错误 $\varepsilon_t=\Pr_{D_t}[h_t(x_i)\neq y_i]$,一旦得到 h_t,AdaBoost 选择一个参数 $\alpha_t\in R$,该参数直观的测量 h_t 的重要程度. 最终假设 H 是 T 次循环后用加权多数投票把 T 个弱假设的输出联合起来得到的. 对二值 h_t,设 $\alpha_t=1/2\ln(1-\varepsilon_t)/\varepsilon_t)$. 可以证明 AdaBoost 调用给定的弱学习算法 WeakLearning 时,将产生错误率为 $\varepsilon_1,\cdots,\varepsilon_T$ 的假设. 假设每个 $\varepsilon_t\leqslant 1/2$,可以证明最终假设 h_f 的错误率 $\varepsilon=\Pr_{i\sim D}[h_t(x_i)\neq y_i]$ 的上边界为:$\varepsilon\leqslant 2^T\prod_{t=1}^{T}\sqrt{\varepsilon_t(1-\varepsilon_t)}$,即随着弱分类器数目 T 的增加,最终假设 h_f 的分类错误率指数级下降. 因此,只要弱分类器的分类正确率比随机猜测好,AdaBoost 就能提升其分类正确率. 基于 AdaBoost 的这种优越性,作者将 AdaBoost 和神经网络结合起来用于水质评价.

1.3　AdaBoost－NN 算法的实现过程

AdaBoost 和神经网络相结合的 AdaBoost－NN 水质评价方法的算法框图如图 2 所示.

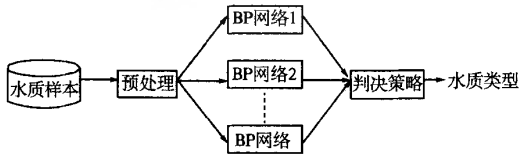

图 2　AdaBoost－NN 结构图

Fig. 2　AdaBoost－NN structure

对上述的 AdaBoost 算法,要成功提高弱分类器的分类精度,必须满足每次生成的弱分类器的分类正确率不小于 0.5. 那么,直接将上述的 AdaBoost 扩展到多分类中,如果弱学习器不能获得至少 50% 的分类正确率则该方法失效. 为此,作者选择 AdaBoost. M2 算法来解决多分类问题. AdaBoost. M2 算法的核心思想在于采用简化多类问题为多个二值问题加以解决,即每个有正确类别号 y_i 的样本 x_i 和每个非正确类别号 y(y 表示除 y_i 外的其它 $k-1$ 个类)这样的二值问题. 弱分类器用于预测样本的可能类别集,并通过置信度表示样本被判为某一类别的可能性大小,即产生 $h_t: X\to[0,1]$,这样就不能再用以前的分类错误率来度量弱分类器的性能好坏,而是采用"伪误差"来度量. "伪误差"的定义式为

$$\text{ploss}_q(h,i)\approx 0.5\times\Big(1-h(x_i,y_i)+\sum_{y\neq y_i}q(i,y)h(h_i,y)\Big)\qquad(2)$$

式中:$h(x_i,y_i)$ 为弱分类器 h 将 x 分类为 y 的置信度. 函数 $q:\{1,\cdots,N\}\times Y\to[0,1]$ 称为类别权重函数,它与区分非正确类别和正确类别相联系,说明不同的区分问题很可能在不同情况下有不同的重要性,对 i 有 $\sum_{y\neq y_i}q(i,y)=1$. 弱学习器的目标是对给定分布 D 和权重函数 q 最小化预期的伪损失为 $\text{ploss}_{D,q}(h):=E_{i\sim D}[\text{ploss}_q(h,i)]$. 通过控制 D 和 q,本算法有效地使弱学习器不仅集中于困难

的实例,也集中于最难于排除的非正确类别上.相反地,这种伪损失衡量可能使弱学习器易于得到弱优势.

在AdaBoost.M2算法中,要求弱学习器的输出是一个在[0,1]范围内的向量,因此我们采用BP神经网络作为弱分类器,神经网络的输入节点数目由水质影响因子(如COD、BOD_5、氨氮等)数目决定;神经网络的输出由待确定的水质类别数目来决定,对于期望输出采用最简单的编码方式,即正确输出类别响应为1,错误类别响应为0;我们采用只含有一个隐层的BP网络,隐层节点数通过实验来确定.

AdaBoost - NN算法具体实现过程如下:

首先对原始水质样本特征向量进行归一化处理,根据在弱学习器中制定的编码规则,对训练样本的期望类别进行编码,从而生成训练样本集

$$S = \langle (x_1, y_1), \cdots, (x_N, y_N) \rangle$$

初始化权值,对 $i = 1,2,\cdots,N$,置权重 $W_{i,y}^1 = D(i)/(k-1)$,$y \in Y - \{y_i\}$,$D(i) = 1/N$.这样样本的初始分布服从均匀分布,在第一次学习时,所有样本的学习难易度都是一样的.

其次,对所有的 $t = 1,2,\cdots,T$:

(1) 置 $W_i^t = \sum_{y \neq y_i} W_{i,y}^t$,$q_t(i,y) = W_{i,y}/W_i^t$,对 $y \neq y_i$,设 $D_t(i) = W_i^t / \sum_{i=1}^{N} W_i^t$;

(2) 调用Weaklearning,传递分布 D_t 和类别权重函数 q_t 给它;返回假设 $h_t: X \times Y \to [0,1]$;

(3) 计算 h_t 的伪损失:

$$\varepsilon_t = 0.5 \sum_{i=1}^{N} D_t(i)(1 - h_t(x_i, y_i) + \sum_{y \neq y_i} q_t(i,y) h_t(x_i, y)).$$

如果 $\varepsilon_t \geq 0.5$,$T = t - 1$,直接输出;

(4) 置 $\beta_t = \varepsilon_t/(1 - \varepsilon_t)$;

(5) 计算新的权重

$$W_{i,y}^{t+1} = W_{i,y}^t \beta^{0.5(1 + h_t(x_i, y_i) - h_t(x_i, y))}.$$

在(3)中 $\varepsilon_t \geq 0.5$ 的情况不会出现,因为我们在BP网络的训练收敛条件中加入了限制因素,使得所有的BP网络分类正确率至少要大于0.5.

最后,利用最大加权投票原则融合各基本分类器 $BP_t(x)$ $(t = 1,2,\cdots,T)$,得到最后的判决结果

$$h_f(x) = \arg\max_{y \in Y} \sum_t (\log 1/\beta_t) h_t(x, y).$$

可见,ε_t 越小,$\log 1/\beta_t$ 越大,即分类器的置信度越高,在组合最后分类器时,其影响因子也越大.

2 实例应用

为检验AdaBoost - NN模型在水质评价应用中的正确性,选取文献[5]中的断面监测资料作为实例,并将几种方法的评价结果进行比较,评价指标参照水质评价标准(GB3838-2002),如表1所示.

神经网络的输入层神经元数目由评价因子决定,根据该地区实际情况选取pH值、COD_{cr}、BOD_5、NH_3-N、石油类、硫化物、挥发酚、氰化物等8项指标作为参评因子(如表2所示),考虑到pH值均在正常范围内,故未将pH值作为参评因子,因此输入层神经元数目为7.网络输出应为水质分类结果,设定Ⅰ类、Ⅱ类、Ⅲ类、Ⅳ类、Ⅴ类水的期望输出结果分别为:(10000)、(01000)、(00100)、(00010)和(00001),因此输出层神经元数目为5.采用包含一个隐层的神经网络结构,经比较计算,隐层神经元数目定为10.因此弱分类器的网络结构为7—10—5.学习参数取:最大学习次数1 000,学习速度0.5,学习精度0.001.

表1 地表水环境质量标准(GB3838-2002)

Tab.1 Environment quality standard for surface water(GB3838-2002) $mg \cdot L^{-1}$

水质级别	pH	COD_{cr}	BOD5	NH_3-N	石油类	硫化物	挥发酚	氰化物
Ⅰ	6-9	15	3	0.15	0.05	0.05	0.002	0.005
Ⅱ	6-9	15	3	0.5	0.05	0.1	0.002	0.05
Ⅲ	6-9	20	4	1.0	0.05	0.2	0.005	0.2
Ⅳ	6-9	30	6	1.5	0.5	0.5	0.01	0.2
Ⅴ	6-9	40	10	2.0	1.0	1.0	0.1	0.2

表 2　浊漳河某河段水质状况(2002)

Tab. 2　Water quality situation of the certain sections in Zhuozhang river　　mg·L^{-1}

断面	pH	COD_{cr}	BOD_5	NH3-N	石油类	硫化物	挥发酚	氰化物
1	8.18	31.9	3.9	7.47	0.19	0.05	0.004	0.009
2	8.25	31.9	6.4	8.79	0.03	0.05	0.013	0.011
3	8.54	19.9	5.2	1.43	0.21	0.05	0.001	0.002
4	8.33	23.7	6.0	2.01	0.04	0.05	0.006	0.002
5	8.28	29.9	4.6	7.85	0.04	0.05	0.014	0.009

AdaBoost 的训练最大轮数 $T=6$，采用 1.3 中描述的训练算法进行训练，当 AdaBoost 训练结束，最后的判决函数由已经训练过的 BP 函数组成，为 $h_f(x)=\arg\max_{y\in Y}\sum_t(\log 1/\beta_t)h_t(x,y)$，这里 T 为实际训练的轮数.

考虑到以地表水环境质量标准作为训练样本，训练样本过少，影响网络效果，因此在各级水质评价标准内按随机均匀分布方式内插生成训练样本，各级水质标准生成 300 个样本，其中 200 个样本作为训练样本，剩下 100 个作为检验样本，则总的训练样本为 1 000 个，检验样本为 500 个. 采用归一化数据，经过反复的训练学习后，网络模型收敛，可以发现，随着神经网络个数的增加，AdaBoost－NN 模型的训练误差指数级单调下降，更重要的是，当其训练误差为 0 时，检验样本的识别误差仍在下降，可见该模型有着良好的泛化性能. 通过训练好的 AdaBoost－NN 模型对浊漳河 5 个断面的水质进行评价，并将评价结果与文献[6]中传统的 ANN 和内梅罗综合指数法对各断面的评价结果进行比较，详见表 3.

表 3　AdaBoost－NN、ANN 方法与内梅罗综合指数法评价结果

Tab. 3　The assessment results of the AdaBoost－NN, ANN and Nernrow

断面	AdaBoost－NN 输出值					AdaBoost－NN	ANN	内梅罗综合污染指数法
1	0.231	0.554	0.326	0.973	0.421	Ⅳ	Ⅳ	Ⅳ
2	0.263	0.328	0.452	0.924	0.506	Ⅳ	Ⅲ	Ⅴ
3	0.318	0.362	0.875	0.431	0.252	Ⅲ	Ⅲ	Ⅰ
4	0.183	0.459	0.902	0.207	0.372	Ⅲ	Ⅲ	Ⅰ
5	0.363	0.496	0.183	0.953	0.207	Ⅳ	Ⅳ	Ⅳ

从表 3 可以看出，3 种方法对断面 1、断面 5 评价结果一致，但对断面 2、断面 3 和断面 4 水质类别判定有差异.

对于断面 2，内梅罗综合指数法片面强调了水中污染物的相对污染值的最大值，使得大量有用信息被丢弃，污染评判结果偏重；传统 ANN 方法评定为 III 类，这反映出了该方法泛化能力不足，对于各评价参数参差不齐的非训练样本，无法正确判断.

对于断面 3、4，AdaBoost－NN 与 ANN 方法的评价结果一致，而与内梅罗综合指数法评价结果相差 2 级，由断面 3、4 评价参数的实测值与其评价结果比较可知，其 COD、BOD_5、NH_3－N、挥发酚的实测值都已超过了评价标准 II 类，内梅罗综合指数法仍将断面 3、4 水质级别评价为Ⅰ级，恰恰反应了该方法在确定权重时受主观因素影响的缺点.

总体来看，内梅罗综合指数法的评价结果参差不齐，同一河段水质跳跃极大，不能反映因子的重要性. AdaBoost－NN 与 ANN 方法的评价结果基本相同，浊漳河某河段多为Ⅲ、Ⅳ级水质，这与实际情况相符. 可见，AdaBoost－NN 网络模型继承了传统的 ANN 方法的优点，具有很强的自学习、自组织能力，建立起输入与输出之间的复杂的非线性对应关系. 而且网络中的大量参数均由学习所得，避免了人为因素的影响. 同时，当传统 ANN 方法在训练数据量过大、输入参数过多，模型的泛化性能降低，对于非训练样本可能不能正确识别的情况下，AdaBoost－NN 网络模型可以通过融合多个弱分类器的分类信息，更加充分利用给定的水质数据信息，提高模型的泛化能力，对水质样本的判别具有更高的准确度，得到的评价结果更客观、更合理.

3　结论

实践证明，对参数和等级较多的环境质量

(下转第 121 页)

A Method of Logical Join of Multi - sheet Digital Maps Based on ArcGIS and Auto CAD

ZHANG Cheng-cai[1], JI Guang-hui[1], CHEN Jun-bo[2], MENG De-chen[3]

(1. School of Environment and Water Conservancy Engineering, Zhengzhou University, Zhengzhou 450001, China; 2. College of Employee of Architecture Employee of Henan Province, Zhengzhou 450007, China; 3. Bureau of Navigatonal Affairs, Communication Department of Henan Province, Zhengzhou 450052, China)

Abstract: Map digitization is one of the first job of the GIS work. aim for the usually question of logical join of multi - sheet maps and then the logical wrong in this first job, The authors discuss a method of logical join of a multi - sheet digital maps and can correct the space logical wrong based on the ArcGIS , and realize the multi - sheet digital maps logical join in a real meaning, then applied in the practical of the inland river channel GIS of HeNan province ,and get an ideal effect .

Key words: map digitization ; multi - sheet ; logical join

(上接第 117 页)

评价,AdaBoost-NN 只需在标准样本训练时改变输入节点数和隐层节点数即可.将 AdaBoost 算法与神经网络结合,避免了传统的水质评价方法中人为确定权值的主观因素的影响,同时在很大程度上弥补 BP 算法的缺陷,增强了水质评价模型的泛化能力,使得评价结果更加客观、准确,很适合作为水质综合评价的通用模型.

参考文献:

[1] 李祚泳,丁　晶,彭荔红.环境质量评价原理与方法[M].北京:化学工业出版社,2004.

[2] 罗士心,毛红梅,陶守耀.水质评价方法综述[J].水资源研究,2002,23(3):16 - 20.

[3] 李如忠.水质评价理论模式研究进展及趋势分析[J].合肥工业大学学报(自然科学版),2005,28(4):369 - 373.

[4] 楼文高,王延正.基于 BP 网络的水质综合评价模型及其应用[J].环境污染治理技术与设备,2003,4(8):23 - 26.

[5] 李祚泳.基于 B-P 网络的水质营养状态评价模型及效果检验[J].环境科学学报,1995,15(2):186 - 191.

[6] 杨国栋,王肖娟,尹向辉.人工神经网络在水环境质量评价和预测中的应用[J].干旱区资源与环境,2004,18(6):10 - 14.

[7] 阮仕平,党志良,胡晓寒,等.人工神经网络在综合水质评价中的应用[J].水资源研究,2004,25(2):21 - 23.

[8] SCHAPIRE R E. The Strength of Weak Learnability. Machine Leaning,1990,5(2):197 - 227.

[9] FREUND Y. Boosting a Weak Learning Algorithm by Majority. Information and Comutation,1995,121(2):256 - 285.

[10] 楼顺天,施　阳.基于 MATLAB 的系统分析与设计——神经网络[M].西安:西安电子科技大学出版社,1999.

The Application of AdaBoost - NN to Water Quality Assessment of the Zhuozhang River

HE Zheng - guang, SUN Xiao - feng, MA Yong - guang

(School of Environment and Water Conservancy Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract: In order to overcome the shortage of conventional BP Neural Network, the AdaBoost algorithm is applied to water quality assessment based on neural networks to set up an AdaBoost-NN model for comprehensive assessment of water quality. The accuracy of the model is examined based on the data of water quality compared with conventional models. The results show that the AdaBoost-NN model for water quality assessment is more objective and reasonable compared with the traditional methods.

Key words: water quality assessment; BP network; AdaBoost; generation ability