Statistics and Application 统计学与应用, 2019, 8(5), 725-733
Published Online October 2019 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2019.85082

Hans汉斯

Prediction and Assessment of Air Pollution in Shandong Province Based on CART Decision Tree and Radial Basis Function Neural Network

Yanan Zhao

School of Mathematical Sciences, Ocean University of China, Qingdao Shandong
Email: erinyanan@163.com

Received: Sep. 18th, 2019; accepted: Oct. 2nd, 2019; published: Oct. 9th, 2019

Abstract

In order to better monitor air quality and make corresponding air protection measures, this paper uses CART tree to model the air quality level of Shandong Province in 2018, and the data from the first half of 2019 for classifying and predicting. Compared with RBF network, empirical analysis shows that the CART tree has a better fitting effect with higher model accuracy, and this model can also be applied to the forecasting and control of air pollution in Shandong Province.

Keywords

AQI, CART Tree, RBF Neural Network, Model Pros and Cons

基于CART决策树和RBF神经网络的山东省空气污染状况预测评估

赵亚男

中国海洋大学数学科学学院，山东 青岛
Email: erinyanan@163.com

收稿日期：2019年9月18日；录用日期：2019年10月2日；发布日期：2019年10月9日

摘 要

为了更好地监测空气质量，作出相应的空气保护措施，本文运用CART树对山东省2018年的空气质量级

文章引用: 赵亚男. 基于CART决策树和RBF神经网络的山东省空气污染状况预测评估[J]. 统计学与应用, 2019, 8(5): 725-733. DOI: 10.12677/sa.2019.85082

别进行建模，并用2019年上半年的数据进行分类预测，并将此方法与RBF网络进行对比，实证分析表明CART树拟合效果更好，模型准确率更高。而此模型也可以运用到山东省空气污染情况的预测治理上。

关键词

AQI，CART树，RBF网络，模型优劣对比

1. 引言

空气质量指数(AQI)，就是根据环境空气质量标准和各项污染物对人体健康、生态、环境的影响，将常规监测的几种空气污染物浓度简化成为单一的概念性指数值形式，它将空气污染程度和空气质量状况分级表示，适合于表示城市的短期空气质量状况和变化趋势[1]。参与空气质量评价的主要污染物为细颗粒物、可吸入颗粒物、二氧化硫、二氧化氮、臭氧、一氧化碳等六项。

空气污染指数的取值范围定为0~500，其中0~50、51~100、101~200、201~300和大于300，分别对应国家空气质量标准中日均值的I级、II级、III级、IV级和V级标准的污染物浓度限定数值，在实际应用中，又把III级和IV级分为III(1)级、III(2)级和IV(1)级、IV(2)级。I级，空气质量评估为优，对人体健康无影响；II级，空气质量评估为良，对人体健康无显著影响；III级，为轻度污染，健康人群出现刺激症状；IV级，中度污染，健康人群普遍出现刺激症状；V级，严重污染，健康人群出现严重刺激症状[2]，见表1。

Table 1. AQI air quality classification
表1. AQI空气质量类别划分

AQI	0~50	51~100	101~150	151~200	201~300	>300
级别	一级	二级	三级	四级	五级	六级
类别	优	良	轻度污染	中度污染	重度污染	严重污染

本文获取了2018年山东省各市的空气质量状况数据(共5853条数据)，基于R软件和SPSS软件运用CART分类树和径向基函数神经网络模型进行了建模，用2019年上半年的数据(共2335条数据)进行模型验证，比较两种模型的优劣。

2. CART树原理

2.1. CART树

分类与回归树模型(Classification and Regression Tree，简写为CART)由Breiman等人在1984年提出，是应用广泛的决策树学习方法。CART 假设决策树是二叉树，内部结点特征的取值为“是”和“否”，左分支为“是”，右分支为“否”，等价于递归的二分每个特征，将输入空间即特征空间划分为有限个单元，并在这些单元上确定预测的概率分布，也就是在输入给定的条件下输出的条件概率分布。分类树的输出是样本的类别，回归树的输出是一个实数[3]。

2.2. 分类树

2.2.1. Gini 指数

1) 假设有 K 个类，样本点属于第 K 类的概率为 p_k，则概率分布的基尼指数定义为

$$\mathrm{Gini}(p)=\sum_{k=1}^{K}p_k(1-p_k) \tag{1}$$

2) 对于二类分类问题，若样本点属于第 1 个类的概率为 p，则概率分布的基尼指数为

$$\mathrm{Gini}(p)=p(1-p) \tag{2}$$

3) 对于给定的样本集合 D，其基尼指数为

$$Gini(D)=1-\sum_{k=1}^{K}\left(\frac{|C_k|}{|D|}\right)^2 \tag{3}$$

其中，C_k 是 D 中属于第 k 类的样本子集，K 是类的个数。

如果样本集合 D 根据特征 A 是否取某一可能值 a 被分割成 D_1 和 D_2 两部分，即

$$D_1=\{(x,y)\in D\,|\,A(x)=a\},D_2=D-D_1 \tag{4}$$

则在特征 A 的条件下，集合 D 的基尼指数为

$$\mathrm{Gini}(D,A)=\frac{|D_1|}{|D|}Gini(D_1)+\frac{|D_2|}{|D|}\mathrm{Gini}(D_2) \tag{5}$$

$\mathrm{Gini}(D,A)$ 表示经 $A=a$ 分割后集合 D 的不确定性，基尼指数值越大，不确定性越大[4] [5]。

2.2.2. CART 树算法

输入：训练数据集 D，停止计算的条件；

输出：CART 决策树。

1) 根据训练数据集 D，从根结点开始，递归地对每个结点进行以下操作，构建二叉树：

2) 设结点的训练数据集为 D，计算现有特征对该数据集的 Gini 系数。此时，对每一个特征 A，对其可能取的每个值 a，根据样本点对 $A=a$ 的测试为“是”或“否”将 D 分割成 D_1 和 D_2 两部分，计算 $A=a$ 时的 Gini 系数。

3) 在所有可能的特征 A 以及它们所有可能的切分点 a 中，选择 Gini 系数最小的特征及其对应的切分点作为最优特征与最优切分点。依最优特征与最优切分点，从现结点生成两个子结点，将训练数据集依特征分配到两个子结点中去。

4) 对两个子结点递归地调用步骤 l~2，直至满足停止条件。

5) 生成 CART 决策树。

2.3. 剪枝

输入：CART 算法生成的决策树 T_0；

输出：最优决策树 T_α

1) 设 $k=0,T=T_0,\alpha=+\infty$；

2) 自上而下地对各内部结点 t 计算 $C(T_t),|T_t|$ 以及 $g(t)=\dfrac{C(T)-C(T_t)}{|T_t|-1},\alpha=\min(\alpha,g(t))$；这里，$T_t$ 表示以 t 为根结点的子树，$C(T_t)$ 是对训练数据的预测误差，$|T_t|$ 是 T_t 的叶节点个数；

3) 自上而下地访问内部结点 t，如果有个 $g(t)=a$，进行剪枝，并对叶结点 t 以多数表决法决定其类，得到树 T；

4) 设 $k=k+1,\alpha_k=\alpha,T_k=T$；

5) 如果 T 不是由根节点单独构成的树，则回到步骤(4)；

6) 采用交叉验证法在子树序列 $T_0,T_1,\cdots,T_n$ 中选择最优子树 T_α [6]-[11]。

3. 径向基神经网络

3.1. RBF 神经网络

径向基(Radial Basis Function)网络是由 Powell M.J.D.于 1985 年提出的，以函数逼近理论为基础构造的一类前向型网络，具有自学习、自组织和自适应等特点，相较于 BP 神经网络和灰色关联度，RBF 神经网络具有学习速度快、精度高以及建立网络和训练网络时间少等优点。径向基函数网络是一个只有两层的网络，在中间层，它以对局部响应的径向基函数代替传统的全局响应的激发函数。由于局部相应的特性，它对函数的逼近是最优的，而且训练过程很短，它具有简单的结构、快速的训练过程及与初始权值无关的优良特性。

RBF 神经网络的基本思想：用 RBF 作为隐单元的“基”构成隐藏层空间，隐藏层对输入矢量进行变换，将低维的模式输入数据变换到高维空间内，使得在低维空间内的线性不可分问题在高维空间内线性可分。就是用 RBF 的隐单元的“基”构成隐藏层空间，这样就可以将输入矢量直接(不通过权连接)映射到隐空间。当 RBF 的中心点确定以后，这种映射关系也就确定了。

3.2. RBF 算法

采用径向基函数(RBF)神经网络，是具有单隐层的 3 层前向网络。

1) 输入层 X：由信号源节点构成，仅起到数据信息的传递作用，对输入信息不作任何变换。

2) 隐藏层 H：节点数视需要而定. 隐含层神经元核函数(作用函数)是高斯函数，对输入信息进行空间映射的变换。

3) 输出层 Y：对输入模式作出响应. 输出层神经元的作用函数为线性函数，对隐含层神经元输出的信息进行线性加权后输出，作为整个神经网络的输出结果。

径向基神经网络的数学模型为

$$y_i=\sum_{i=1}^{n_c}w_i g\left(\|x-c_i\|/\sigma_i\right)+b \tag{6}$$

式中：x 为神经网络输入的 n 维向量；w_i 为输出层权重；$g(*)$ 为径向基函数；c_i 为径向基函数中心；σ_i 为宽度；b 为输出层阈值；n_c 为隐藏层神经元数目；$\|x-c_i\|$ 为向量 $x-c_i$ 的范数，通常表示 x 与 c_i 间的距离。

通常选择高斯基函数为径向基函数，输出层阈值为 0，该层神经元 i 的输出为

$$R_i(x)=\exp\left[-\|x-c_i\|^2/\left(2\sigma_i^2\right)\right] \tag{7}$$

则隐藏层与输出层的映射关系为

$$Y=f(x)=WR=\sum_{i=1}^{n_c}w_i R_i(x) \tag{8}$$

式中：Y 是输出向量，$Y=\left[y_1,y_2,\cdots,y_q\right]^{\mathrm{T}}$，其中，$q$ 是输出层的单元数，W 为输出层的权值，R 为隐藏层神经元的输出值。

4. 描述性统计

首先对山东省 2018 年的空气质量数据进行了简单的描述性统计，得出 2018 年间各月份的空气污染状况。由图 1 可以看出，各月份空气质量类别为良的天数占的比重最大，其次为轻度污染，说明山东省整体的空气质量较为良好。各月份中出现空气质量类别为优的月份主要为七月、八月和九月，占比分别为 25.2%、21.77%和 12.29%，即山东省夏季的空气质量状况较好。各月份中出现严重污染的月份依次为一月、十一月和四月，占比分别为 2.42%、1.25%和 0.21%，各月份中出现重度污染的月份依次为一月、十一月、十二月等，占比分别为 13.51%、9.17%和 7.26%，即较为严重的空气污染主要集中在冬季和春季。

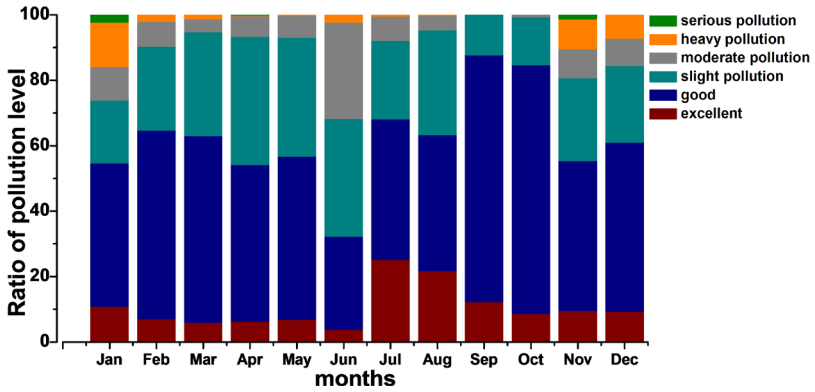

Figure 1. Air quality category for each month
图 1. 各月份空气质量类别

5. 预测模型

5.1. CART 决策树

5.1.1. 模型建立

空气质量预测模型的建立使用了空气质量等级作为最终的预测变量，该变量为离散型。选取 $PM_{2.5}$、PM_{10}、SO_2、NO_2、O_3_8h、CO，一共 6 个自变量进行预测模型的训练与测试，得到图 2 CART 树：

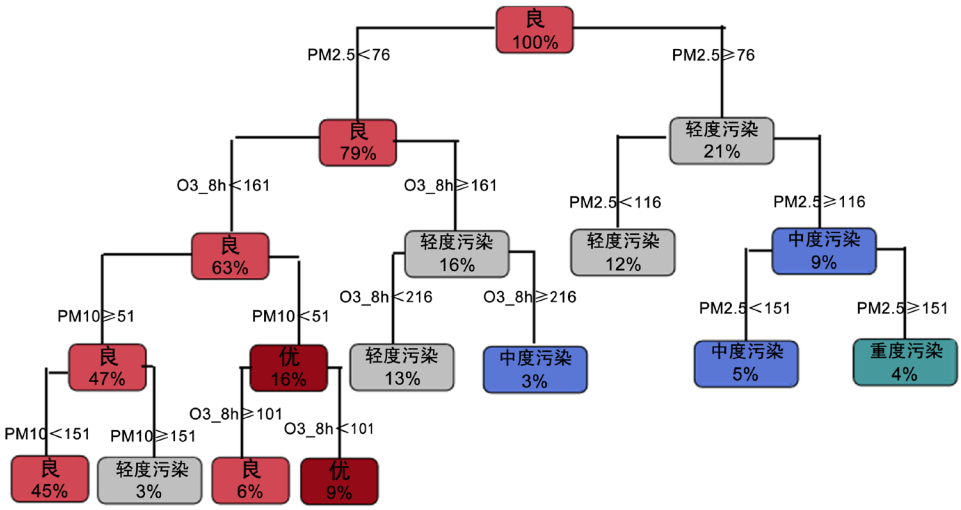

Figure 2. CART-tree model
图 2. 决策树模型

可以看到，训练之后，采用了 $PM_{2.5}$、O_3_8h 和 PM_{10} 三个指标作为分支节点来建立决策树，而忽略了很多与 AQI 相关性不高的的特征。

上述决策树的分支过程如下：

首先，将 $PM_{2.5}$ 作为节点的第一特征，分为左支 D_1—$PM_{2.5} < 76$；右支 D_2—$PM_{2.5} \geq 76$；

对于 D_1，将 O_3_8h 作为节点的第二特征，进一步分为左支 C_1—O_3_8h < 161；右支 C_2—O_3_8h ≥ 161。对于 D_2，将 $PM_{2.5}$ 继续作为节点的第二特征，进一步划分为左支 C_1—$PM_{2.5} < 116$；右支 C_2—$PM_{2.5} \geq 116$。

如此进行下去，得到最终的 CART 树。

由上述 CART 树可以得出如下结论：

1) $PM_{2.5}$、PM_{10} 和 O_3_8h 是影响空气质量级别的主要因素。

2) 当 $PM_{2.5} \geq 151$ 时空气质量级别直接划分为重度污染；

3) $PM_{2.5} < 151$，$161 \leq O_3$_8h < 215.5 或 $116 \leq PM_{2.5} < 156$ 时，空气质量级别划分为中度污染；

4) $76 \leq PM_{2.5} < 116$ 或 $PM_{2.5} < 76$，$161 \leq O_3$_8h < 216 或 $PM_{2.5} < 76$，O_3_8h < 161，$51 \leq PM_{10} < 151$ 时，空气质量级别划分为轻度污染；

5) 当 $PM_{2.5} < 76$，O_3_8h < 161，$PM_{10} < 151$ 或 $PM_{2.5} < 76$，O_3_8h < 161，O_3_8h ≥ 101 时，空气质量级别划分为良；

6) 当 $PM_{2.5} < 76$，O_3_8h < 101，$PM_{10} < 51$ 时，空气质量级别划分为优。

5.1.2. 决策树的剪枝

剪枝是决策树学习算法解决模型“过拟合”的主要手段，在决策树学习中，为了尽可能正确分类训练样本，结合划分过程将不断重复，有时会造成决策树分支过多，这时就可能因训练样本拟合的准确度很高，以致于把训练集自身的一些特点当作所有数据都具有的一般性质而导致过拟合。因此，可通过主动去掉一些分支来降低过拟合的风险[12]。

建立树模型要权衡两方面问题，一个是要拟合得使分组后的变异较小，另一个是要防止过度拟合，而使模型的误差过大，前者的参数是 CP，后者的参数是 Xerror。CP 是参数复杂度(complexity parameter)作为控制树规模的惩罚因子，简而言之，就是 CP 越大，树分裂规模(nsplit)越小。输出参数(rel error)指示了当前分类模型树与空树之间的平均偏差比值。Xerror 为交叉验证误差，Xstd 为交叉验证误差的标准差[13]。所以要在 Xerror 最小的情况下，也使 CP 尽量小。如果认为树模型过于复杂，我们需要对其进行修剪，下面列出了 CP 值与 Xerror 值。

Table 2. CART - tree complexity parameter table
表 2. 决策树的复杂性参数表

CP	nsplit	rel error	xerror	xstd
0.249251	0	1.000000	1.000000	0.0112859
0.094311	2	0.501497	0.501497	0.0097164
0.089696	3	0.407186	0.419162	0.0091172
0.072106	5	0.227794	0.227794	0.0071062
0.065868	6	0.155689	0.155689	0.0059903
0.047904	7	0.089820	0.089820	0.0046287
0.010000	8	0.041916	0.041916	0.0032006

由表 2 可以看出，可以看到，当 nsplit 为 8 的时候，即有四个叶子结点的树，要比 nsplit 为 7，即八个叶子结点的树的交叉误差要小。而决策树剪枝的目的就是为了得到更小交叉误差(xerror)的树。因为本模型较为简单，所以不需要修剪。

5.1.3. 决策树的模型预测

从表 3 可以看出，模型的预测准确率为(1163 + 637 + 0 + 148 + 218 + 141)/2335 = 92.46%。

Table 3. Confusion matrix
表 3. 混淆矩阵

预测值 / 真实值	良	轻度污染	严重污染	优	中度污染	重度污染
良	1163	0	0	3	0	0
轻度污染	1	637	0	0	0	0
严重污染	0	0	0	0	0	11
优	0	0	0	148	0	0
中度污染	0	10	0	0	218	0
重度污染	0	1	0	0	2	141

5.2. 径向基函数神经网络

5.2.1. 模型建立

在本模型中，训练集采用 5853 个样本，占总样本量的 64%，测试集采用 2335 个样本，占总样本量的 36%。RBF 神经网络模型的输入参数和输入层的神经元数量根据实验因素确定，输出参数和输出层的神经元数量根据评价指标确定。在本文之中，输入参数为 $PM_{2.5}$、PM_{10}、SO_2、NO_2、CO、O_3_8h，输入层的神经元有 6 个，输出层的参数为优、良、轻度污染、中度污染、重度污染、严重污染，输出层的神经元有 6 个，隐藏层的隐藏函数为 Softmax 函数。建立如下图 3：

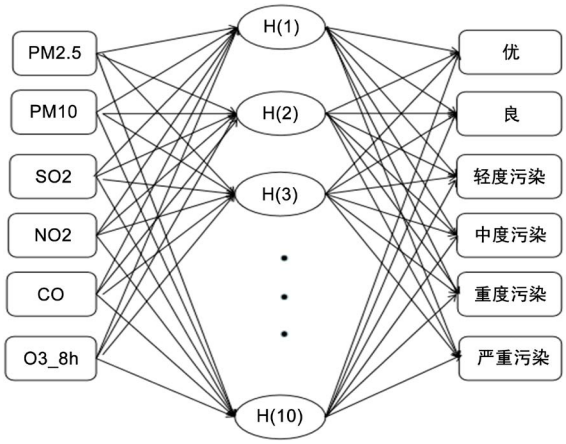

Figure 3. RBF neural network model
图 3. RBF 神经网络模型

5.2.2. 模型评价

ROC 曲线指受试者工作特征曲线(Receiver Operating Characteristic Curve)，是反映敏感性和特异性连续

变量的综合指标，是用构图法揭示敏感性和特异性的相互关系，它通过将连续变量设定出多个不同的临界值，从而计算出一系列敏感性和特异性，再以敏感性为纵坐标、特异性为横坐标绘制成曲线，曲线下面积越大，诊断准确性越高。在 ROC 曲线上，最靠近坐标图左上方的点为敏感性和特异性均较高的临界值。

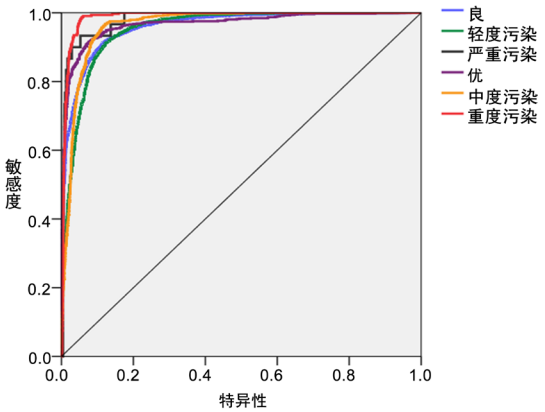

Figure 4. ROC curve
图 4. ROC 曲线

由上述 ROC 曲线图 4 可知，径向基神经网络模型对空气质量类别的拟合效果较好[14] [15] [16] [17] [18]。

由表 4 和表 5 可知，训练集的预测准确率为 83.5%，测试集的预测准确率为 84.2%，模型准确率相较于 CART 树模型的准确率不高。

Table 4. Confusion matrix of training set
表 4. 训练集的混淆矩阵

预测值 / 真实值	良	轻度污染	严重污染	优	中度污染	重度污染
良	2545	286	0	95	2	1
轻度污染	118	1286	0	1	78	9
严重污染	0	0	0	0	0	16
优	110	6	0	426	12	0
中度污染	0	106	0	0	339	38
重度污染	0	1	0	0	65	165

Table 5. Confusion matrix of testing set
表 5. 测试集的混淆矩阵

预测值 / 真实值	良	轻度污染	严重污染	优	中度污染	重度污染
良	1102	116	0	31	1	1
轻度污染	53	619	0	0	34	1
严重污染	0	0	0	0	0	14
优	52	2	0	165	1	0
中度污染	0	49	0	0	140	11
重度污染	0	0	0	0	26	66

5.3. 模型优劣对比

CART 树既可以做分类算法，也可以做回归。其优点为：1) 可以生成可以理解的规则。2) 计算量相对来说不是很大。3) 决策树可以清晰的显示哪些字段比较重要。缺点为：1) 当类别太多时，错误可能就会增加的比较快。2) 一般的算法分类的时候，只是根据一个字段来分类。

径向基神经网络的优点是：1) 分类能力好，学习过程收敛速度快。2) 具有唯一最佳逼近特性，且无局部极小问题存在。缺点是：RBF 神经网络的非线性映射能力体现在隐层基函数上，而基函数的特性主要是由基函数的中心确定的，从数据点中任意选取中心构造出来的 RBF 神经网络的性能显然不能令人满意。

在上述对 2018 年 1 月至 2019 年 6 月的山东省空气质量类别预测的模型建立过程中可以看到，CART 树模型的预测准确率为 92.46%，而径向基函数神经网络模型的预测准确率为 84.2%，显然，CART 树模型的建立更有效。

参考文献

[1] Kampa, M. and Castanas, E. (2008) Human Health Effects of Air Pollution. *Environmental Pollution*, **151**, 362-367. https://doi.org/10.1016/j.envpol.2007.06.012

[2] Zhan, D.S., Kwan, M.-P., Zhang, W.Z., *et al*. (2018) The Driving Factors of Air Quality Index in China. *Journal of Cleaner Production*, **197**, 1342-1351. https://doi.org/10.1016/j.jclepro.2018.06.108

[3] 张松林. CART 分类与回归树方法介绍[J]. 火山地质与矿产, 1997(1): 67-75.

[4] Kim, B. and Kim, J. (2016) Stochastic Ordering of Gini Indexes for Multivariate Elliptical Risks. *Insurance Mathematics and Economics*, **68**, 84-91.

[5] 刘云翔, 吴浩. 基于改进 CART 决策树建立水华预警模型[J]. 中国农村水利水电, 2018(1): 26-28.

[6] 蔡丽清. 基于 CART 算法的高校超市服务应用研究[J]. 电脑知识与技术, 2016, 12(13): 261-263.

[7] 黄晓君. 基于变化检测 CART 决策树模式自动识别沙漠化信息[J]. 灾害学, 2017, 32(1): 36-42.

[8] 孔颖. 基于 CART 算法的垃圾邮件过滤模型设计与实现[J]. 计算机应用, 2009, 29(2): 374-376.

[9] 钱揖丽. 基于分类回归树 CART 的汉语韵律短语边界识别[J]. 计算机工程与应用, 2006, 44(6): 169-171.

[10] 刘玉茹. CART 分析及其在故障趋势预测中的应用[J]. 计算机应用, 2017(S2): 57-59.

[11] 冯洁. CART 算法在银行 CRM 中的应用研究[J]. 高效理科研究, 2011(26): 111-112.

[12] Shang, Z.G., Deng, T., He, J.Q. and Duan, X.H. (2019) A Novel Model for Hourly PM2.5 Concentration Prediction Based on CART and EELM. *Science of the Total Environment*, **651**, 3043-3052. https://doi.org/10.1016/j.scitotenv.2018.10.193

[13] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) Classification and Regression Trees, Wadsworth.

[14] Bai, Y., Li, Y., Wang, X.X., Xie, J.J., *et al*. (2016) Air Pollutants Concentrations Forecasting Using Back Propagation Neural Network Based on Wavelet Decomposition with Meteorological Conditions. *Atmospheric Pollution Research*, **7**, 557-566. https://doi.org/10.1016/j.apr.2016.01.004

[15] Zhu, S.L., Lian, X.Y., Liu, H.X., Hu, J.M., Wang, Y.Y. and Che, J.X. (2017) Daily Air Quality Index Forecasting with Hybrid Models: A Case in China. *Environmental Pollution*, **231**, 1232-1244. https://doi.org/10.1016/j.envpol.2017.08.069

[16] He, Q.F., Shahabi, H. and Shirzadi, A. (2019) Landslide Spatial Modelling Using Novel Bivariate Statistical Based Naïve Bayes, RBF Classifier, and RBF Network Machine Learning Algorithms. *Science of the Total Environment*, **663**, 1-15. https://doi.org/10.1016/j.scitotenv.2019.01.329

[17] Park, J. and Sandberg, I.W. (1993) Approximation and Radial-Basis-Function Networks. *Neural Computation*, **5**, 305-316. https://doi.org/10.1162/neco.1993.5.2.305

[18] Dong, J., Zhao, Y.X. and Liu, C. (2019) Orthogonal Least Squares Based Center Selection for Fault-Tolerant RBF Networks. *Neurocomputing*, **339**, 217-231. https://doi.org/10.1016/j.neucom.2019.02.039