基于 PCA-ABBP 强分类器的矿区采空塌陷危险性预测*

李 旭** 刘 剑
(辽宁工程技术大学安全科学与工程学院,葫芦岛 125105)

摘 要:为了准确、快速地预测矿区采空塌陷危险性,针对矿区采空塌陷影响因素之间存在信息重叠以及利用单一 BP 神经网络进行预测时存在的局部极值等问题,提出了一种 PCA-ABBP 强分类器模型。以北京西山某地的 24 组采空塌陷数据为样本,选取了采空区空间叠置层数等 7 个变量作为矿区采空塌陷的影响因素,以前 17 组数据作为训练样本,建立基于 PCA-ABBP 强分类器的矿区采空塌陷危险性预测模型。利用该模型对后 7 组数据进行预测,预测结果与实际完全相符,而单一 BP 神经网络预测的平均误差为 17.14%,验证了所提出模型的有效性和可靠性。

关键词:采空塌陷危险性;集成学习;BP 神经网络;预测;采空区
中图分类号:X936 **文献标识码**:A doi:10.16507/j.issn.1006 – 6055.2016.05.008

Prediction of Underground Goaf Collapse Risk Based on PCA-ABBP Strong Classifier*

LI Xu** LIU Jian
(College of Safety Science and Engineering, Liaoning Technical University, Huludao 125105)

Abstract: In order to forecast the underground goaf collapse risk accurately and quickly, a PCA-ABBP strong classifier model is proposed. And the proposed method is mainly focus on the problems of information overlap between underground goaf collapse influencing factors and defects of single BP neural network. Based on 24 historical collapse information of Beijing Xishan Mine, and seven variables, such as the number of overlapping layers, are chosen as the influencing factors. The PCA-ABBP strong classifier of underground goaf collapse risk is established using the first 17 training samples. On this basis, the last 7 samples are predicted using the established model, and the predicted results are in complete agreement with the actual results. However, the average error of single BP neural network prediction is 17.14%, which verifies the validity and reliability of the proposed model.

Key words: goafcollapse risk; ensemble learning; BP neural network; prediction; goaf

1 引言

采空塌陷是指由于地下挖掘空间上部的岩土层受到外部应力、自身重力等作用造成岩土层失稳而引起的地面塌陷现象。采用空场采矿法进行开采时,随着矿山资源不断采出,会形成大量的采空区[1]。这些采空区上方岩土层的稳定性会逐渐变差,进而可能引发采空塌陷。由于带有不确定性和突发性,矿区一旦发生采空塌陷事故,往往会带来非常严重的危害[2]。随着矿山资源的不断开采,我国矿区采空区的面积不断增加,矿区采空塌陷灾害也愈发严重,矿区采空塌陷已成为煤矿安全生产领域中亟需解决的重要课题[3]。矿区采空塌陷的预测及其防治具有重要的现实意义和理论价值[4]。

目前,已经有许多学者对矿区采空塌陷预测进行研究,建立了许多预测模型,例如,神经网络模型[5]、决策树模型[6]、贝叶斯判别模型[2]、支持向量机模型[7,8]、模糊评价模型[9]、Fisher 判别模型[10]等。这些模型无疑为矿区采空塌陷预测及其防治做出了重要贡献,同时也存在一定局限,例如:贝叶斯判别模型和 Fisher 判别模型为基于统计学理论的线性算法模型,对样本要求较高,对非线性数据预测结果较差;神经网络模型的结构较难确定而且容易陷入局部极值;支持向量机模型的核函数及参数较难确定等。

矿区采空塌陷的发生与否受到多种因素的共同影响,具有多维度、非线性的动力学特征。BP 神经网络在解决多维度、非线性问题具有很强的优势,无疑是矿区采空塌陷预测的强有力的工具。然而,单一 BP 神经网络对于多噪声样本和小样本问题预测结果相对较差,而且预测时可能会出现局部极值等问题。针对此问题,本文以单一 BP 神经网络作为

2016-03-09 收稿,2016-04-11 接受,2016-10-25 网络发表
* 国家自然科学基金(51374121)资助
** 通讯作者,E-mail:lixulunwen@sina.com

弱分类器,利用 Adaptive Boosting 集成学习算法对 BP 神经网络进行集成,建立一种 ABBP 强分类器模型。同时,考虑到矿区采空塌陷的众多影响因素之间可能存在信息重叠,会增加建模难度并影响模型的预测精度,引入 PCA(Principal Components Analysis)方法可解决这一问题。基于以上分析,本文将 PCA 方法与 ABBP 强分类器相结合,提出一种基于 PCA-ABBP 强分类器的矿区采空塌陷危险性预测模型,并将其应用到实际预测中,以验证所提出模型的有效性和准确性。

2 原理与方法

2.1 PCA 方法

设原始指标体系数据矩阵为

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}_{n\times p}$$

求解矩阵 X 协方差矩阵的 P 个非负特征值对应的特征向量,即

$$C^{(i)} = (C_1^{(i)}, C_2^{(i)}, \cdots, C_p^{(i)}),\ i = 1,2,\cdots,P \tag{1}$$

依据式(1),可以得到矩阵 X 对应的 P 个新因子,即

$$\begin{cases} z_1 = C_1^{(1)}\tilde{x}_1 + C_2^{(1)}\tilde{x}_2 + \cdots + C_p^{(1)}\tilde{x}_p \\ z_2 = C_1^{(2)}\tilde{x}_1 + C_2^{(2)}\tilde{x}_2 + \cdots + C_p^{(2)}\tilde{x}_p \\ \cdots\cdots \\ z_p = C_1^{(p)}\tilde{x}_1 + C_2^{(p)}\tilde{x}_2 + \cdots + C_p^{(p)}\tilde{x}_p \end{cases} \tag{2}$$

其中,$\tilde{x}_i$ 表示 x_i 的标准化变换。

通过上述变换,可以消除原始自变量之间的信息重叠(信息相关),即得到的新因子 $z_1, z_2, \cdots, z_P$ 之间线性无关。新因子的方差贡献率决定了其所保留原始信息的多少,即

$$\alpha_i = \lambda_i / \sum_{i=1}^{p} \lambda_i,\ i = 1,2,\cdots,P \tag{3}$$

式中,α_i 表示新因子 z_i 的方差贡献率,λ_i 表示矩阵 X 的协方差矩阵对应的第 i 个非负特征值。

根据式(3)和实际预测要求,可以求得前 $m(m \leqslant P)$ 个新因子的方差累计贡献率,即

$$\alpha = \sum_{i=1}^{m} \alpha_i \tag{4}$$

依据式(4),可以对主成分进行选取。通常选取 $\alpha > 85\%$ 所对应的前 m 个主成分即可,这样既能消除原始变量之间的信息重叠,又能最大限度的包含原始标体系的信息[11-13]。

2.2 ABBP 强分类器模型

ABBP(Adaptive Boosting Back Propagation)强分类器是以 BP(Back Propagation)神经网络作为弱分类器,利用 AB(Adaptive Boosting)集成学习算法对 BP 神经网络进行集成而建立的一种强分类器模型。Adaptive Boosting 是 Boosting 的一个分支,其最先由 Freund[14] 提出。已有研究表明 Adaptive Boosting 能够有效地提高弱分类器预测精度[15-17]。Adaptive Boosting 的主要思想是:首先,对每一组训练样本赋予一个相同的权重,利用这一组样本通过训练建立一个弱分类器;然后,利用建立好的弱分类器进行预测,并依据预测结果来调整训练样本的权重。权重调整的原则是,降低预测精度较高的样本的权重,增加预测精度低的样本的权重。经过不断地训练和权重调整,可以得到一系列弱预测器及其权值;最后,依据得到的弱预测器和其对应的权值,可以将弱预测器进行集成形成强分类器,以达到提高分类精度的目的。Adaptive Boosting 通过不断地调整样本的权重,可以将预测训练放在关键数据上面。Adaptive Boosting 非常适用于实际问题的求解,因为其采用的是加权投票机制对弱预测器进行集成。具体算法如下:

1)数据选择及 BP 神经网络初始化。从样本数据中选择 ∂ 组训练样本,初始化迭代次数 $t = 1$,初始化训练样本分布权值 $D_1(\vartheta) = \frac{1}{\partial}$,$\vartheta = 1,2,\cdots,\partial$。根据训练样本确定 BP 神经网络结构,并初始化 BP 神经网络权值和阈值;

2)BP 神经网络预测。令迭代总次数为 T,对于 $t = 1,2,\cdots,T$ 进行迭代,可以得到 T 个不同的 BP 神经网络弱预测器。在迭代第 t 次时,可以第 t 个 BP 神经网络弱预测器,利用其对训练样本进行拟合,建立回归模型 $g_t \to y$,可以得到期预测误差率,

$$\xi_t = \sum_{\vartheta} D_t(\vartheta),\vartheta = 1,2,\cdots,\partial \tag{5}$$

3)BP 神经网络弱分类器权重计算。依据 $g_t(x)$ 和 ξ_t 可以计算弱预测器序列的权重 a_t,

$$a_t = \frac{1}{2}\ln\left(\frac{1-\xi_t}{\xi_t}\right),t = 1,2,\cdots,T \tag{6}$$

4)更新训练样本权重。依据得到的弱预测器序列的权重 a_t,可以计算下一轮训练样本权重 $D_{t+1}(\vartheta)$,

$$D_{t+1}(\vartheta) = \frac{D_t(\vartheta)}{B_t}\exp[-a_t y_t g_t(x_\vartheta)] \tag{7}$$

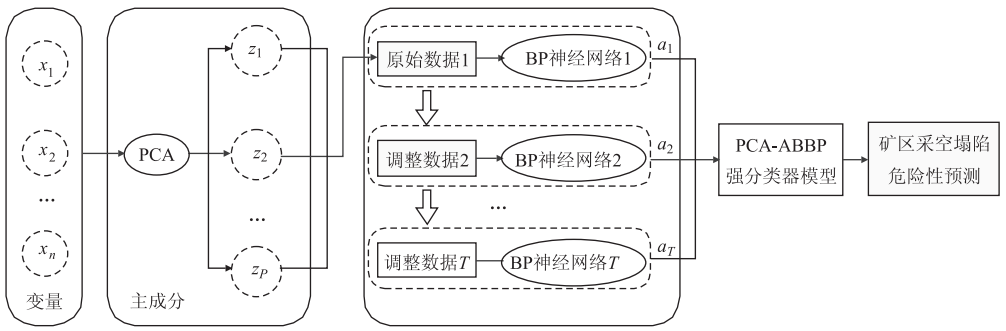

图 1　模型预测流程图

式中，B_t 为归一化因子，目的是使权重求和等于 1，即 $\sum_{\vartheta=1}^{\partial} D_{t+1}(\vartheta) = 1$。

5）建立 ABBP 强分类器。经过 T 轮训练，可以得到 T 个 BP 弱预测器函数 $f(g_t, a_t)$ 和其对应的权重 a_t，利用加权求和方法可以得到 ABBP 强分类器函数 $h(x)$。

$$h(x) = sign\left[\sum_{t=1}^{T} a_t \cdot f(g_t, a_t)\right] \quad (8)$$

2.3　矿区采空塌陷危险性预测的 PCA-ABBP 强分类器

如图 1 所示，利用 PCA-ABBP 强分类器对矿区采空塌陷危险性进行预测的具体步骤如下：

1）建立矿区采空塌陷危险性预测指标体系，并依据建立的指标体系收集样本数据。

2）利用 PCA 方法对样本数据进行分析，消除样本数据间的信息重叠，并选取有效主成分。

3）利用选取的主成分对 ABBP 强分类器进行训练，建立矿区采空塌陷危险性预测的 PCA-ABBP 强分类器模型。

4）利用建立好的 PCA-ABBP 强分类器模型对待识别的矿区采空塌陷的危险性进行预测。

3　矿区采空塌陷危险性预测 PCA-ABBP 强分类器模型的建立及应用

3.1　预测指标体系的建立及样本数据的获取

矿区采空塌陷的发生与否受到多种因素的共同影响，而且这些影响因素之间又存在着很复杂的关系。本文通过对大量矿区采空塌陷案例进行分析并参考相关文献，选取了 7 个变量作为矿区采空塌陷的影响因素，如图 2 所示。由于变量 x_5 和 x_7 为非数值变量，这里对其进行量化处理，量化方式见表 1。以北京西山某地采空塌陷数据为样本进行试验分析，其中前 17 组样本数据用于训练和建立 PCA-ABBP 强分类器模型，后 7 组数据用于模型检验。样本数据见表 2。

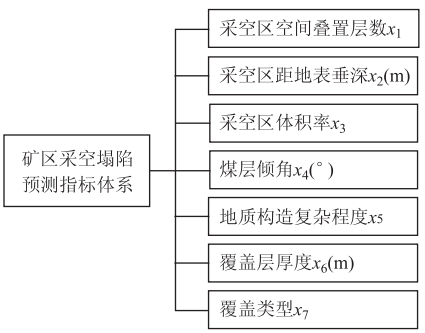

图 2　矿区采空塌陷预测指标体系

表 1　变量 x_5 和 x_7 量化方式

变量		量化值
x_5	x_7	
复杂	砂砾夹粉土、碎石或者黄土	3
一般	黏土、粉尘和砂砾	2
简单	少量砂砾、以黏土为主	1

表 2　样本数据

序号	x_1	x_2	x_3	x_4	x_5	x_6	x_7	实际类别[1)]
1	3	10.4	18	28	2	7.5	3	1
2	3	22	18	45	2	11.5	3	1
3	3	16	14	55	3	14.5	2	1
4	2	16.7	15	70	3	14	3	−1
5	3	15.4	1.5	50	2	13.5	3	−1
6	1	26	6	35	2	19	2	−1
7	2	22.5	4	50	2	10	1	−1
8	1	18.2	5	35	1	15.5	3	−1
9	2	25	7	40	2	12	2	−1
10	2	20.2	20	80	3	17	3	1
11	1	16.5	2	40	2	15	3	−1
12	1	16.4	2.5	45	2	10	2	−1
13	2	30	5.5	25	1	15	1	−1
14	3	12.7	12	75	3	9.5	2	1
15	4	14.5	11	55	3	12.5	2	1
16	3	17.5	10	50	2	15	2	1
17	3	13.5	10	50	3	12	3	1
18	3	10.4	18	28	2	7.5	3	1
19	3	22	18	45	2	11.5	3	1
20	3	16	14	55	3	14.5	2	1
21	4	14.5	11	55	3	12.5	3	1
22	3	17.5	10	50	2	15	3	1
23	1	18.2	5	35	1	15.5	2	−1
24	2	25	7	40	2	12	1	−1

1）1 表示发生塌陷，−1 表示稳定。

3.2　模型建立

在 SPSS 20.0 环境下对样本数据进行主成分分析,得到各主成分方差贡献率和累计贡献率。结果如表 3 所示,前 4 个主成分的累计方差贡献率为 88.664%(>85%),原则上提取前 4 个主成分即可,为了提高预测精度,本文选取前 5 个主成分。提取的原始数据的前 5 个主成分结果见表 4。

表 3　各主成分及方差贡献

主成分	初始特征值	方差贡献率/%	累计方差贡献率/%
z_1	3.091	44.157	44.157
z_2	1.378	19.679	63.837
z_3	1.009	14.411	78.247
z_4	0.729	10.417	88.664
z_5	0.425	6.074	94.739
z_6	0.219	3.125	97.864
z_7	0.150	2.136	100.000

表 4　原始数据的前 5 个主成分

序号	z_1	z_2	z_3	z_4	z_5	实际类别
1	0.85548	-2.42645	0.44651	-0.02116	-0.9794	1
2	0.44043	-0.46506	1.3462	1.37381	-0.30913	1
3	0.60558	0.74156	-0.55878	0.47994	0.17626	1
4	1.19807	0.07923	-0.18617	-0.1507	1.65888	-1
…	…	…	…	…	…	…
21	1.19807	0.07923	-0.18617	-0.1507	1.65888	1
22	0.2471	0.23385	0.99145	-0.58535	0.95258	1
23	-1.38388	-0.23663	0.60838	-1.2876	-0.68546	-1
24	-1.07041	-0.04619	-1.1194	1.59267	0.1166	-1

依据表 4 和上述 ABBP 强分类器的相关理论,在 Matlab2014 环境下进行仿真实验,建立矿区采空塌陷危险性预测的 PCA-ABBP 强分类器模型。其中,相关参数设置如下:1)BP 神经网络结构为 5-11-1;2)BP 神经网络学习率为 0.1;3)最大训练步数为 100;4)弱分类器(BP 神经网络)数量 $T=10$,即训练样本迭代次数为 10。利用表 2 中前 17 组样本数据对模型进行训练,得到弱预测器序列的权重和训练样本权重见表 5 和表 6。利用训练好的模型对原始样本数据进行回代,预测结果与实际结果一致。

3.3　模型应用

利用建立好的矿区采空塌陷危险性预测 PCA-ABBP 强分类器模型对后 7 组数据进行预测,模型预测结果见图 3。为了便于比较和分析,同时统计了 10 个 BP 神经网络弱分类器的 10 次分类结果(表 7)。由图 3 和表 7 可知,PCA-ABBP 强分类器模型对 7 组测试样本的预测结果与实际情况完全相符,而利用单一 BP 神经网络模型预测的平均误差为 17.14%,验证了所提出模型的有效性、可靠性及准确性。

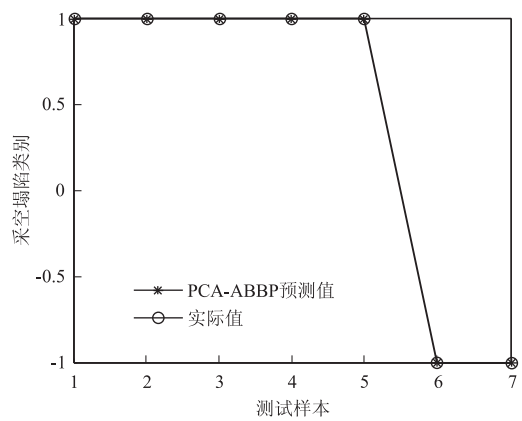

图 3　预测值及比较

表 5　弱预测器序列的权重

a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
1.0075	1.1387	1.7645	0.9154	0.7641	0.782	0.7559	1.359	1.325	0.1218

表 6　训练样本权重

序号	$D_1(\vartheta)$	$D_2(\vartheta)$	$D_3(\vartheta)$	$D_4(\vartheta)$	$D_5(\vartheta)$	$D_6(\vartheta)$	$D_7(\vartheta)$	$D_8(\vartheta)$	$D_9(\vartheta)$	$D_{10}(\vartheta)$	$D_{11}(\vartheta)$
1	0.0588	0.0581	0.0462	0.0626	0.057	0.0594	0.0631	0.0577	0.0506	0.0178	0.0121
2	0.0588	0.0581	0.0566	0.0637	0.0542	0.0636	0.0729	0.0667	0.0775	0.0965	0.0861
3	0.0588	0.051	0.075	0.0796	0.0688	0.0752	0.0774	0.0674	0.061	0.0759	0.08
4	0.0588	0.0203	0.0242	0.0369	0.0666	0.0679	0.0567	0.0518	0.0475	0.0591	0.0649
5	0.0588	0.0581	0.0478	0.0741	0.1067	0.1134	0.099	0.0994	0.0903	0.1125	0.1214
6	0.0588	0.0581	0.0578	0.0327	0.0295	0.0328	0.0327	0.0128	0.0118	0.0147	0.0137
7	0.0588	0.0581	0.0483	0.0564	0.0524	0.0498	0.0567	0.0519	0.0414	0.0516	0.0573
8	0.0588	0.0581	0.024	0.037	0.0285	0.0257	0.0223	0.049	0.039	0.0073	0.0071
9	0.0588	0.0581	0.059	0.0865	0.0803	0.0739	0.0736	0.0673	0.0357	0.0445	0.0403
10	0.0588	0.0602	0.0928	0.0619	0.0597	0.0721	0.0732	0.1366	0.1373	0.0357	0.0326
11	0.0588	0.1374	0.1467	0.1042	0.0925	0.0897	0.0844	0.0772	0.1659	0.2065	0.218
12	0.0588	0.0551	0.0618	0.0476	0.0236	0.0295	0.0273	0.025	0.027	0.0336	0.0291
13	0.0588	0.0581	0.0729	0.0742	0.151	0.1309	0.1391	0.1272	0.116	0.171	0.1792
14	0.0588	0.0369	0.0232	0.0116	0.0058	0.0055	0.0052	0.0036	0.0033	0.005	0.0044
15	0.0588	0.0581	0.0639	0.0703	0.0605	0.0468	0.0476	0.0435	0.0362	0.0451	0.0333
16	0.0588	0.0581	0.0571	0.0408	0.0067	0.0064	0.0081	0.0074	0.0068	0.0085	0.008
17	0.0588	0.0581	0.0426	0.0598	0.0562	0.0573	0.0606	0.0555	0.0525	0.0148	0.0126

表 7 10 个 BP 神经网络弱分类器预测错误样本数量[1)]

分类器编号	1	2	3	4	5	6	7	8	9	10
预测错误样本数	0	0	2	1	2	4	1	0	0	2

1)平均误差 17.14%。

4 结论

针对矿区采空塌陷影响因素之间可能存在信息重叠的问题,引入 PCA 方法来消除采空塌陷的众多影响因素之间的信息重叠,这不仅能降低预测体系变量维度,而且能提高模型的训练速度和预测精度。针对单一 BP 神经网络对于多噪声样本和小样本问题预测结果相对较差,且在利用 BP 神经网络预测时可能会出现局部极值等问题,利用 Adaptive Boosting 集成学习算法对 BP 神经网络进行集成,建立一种 ABBP 强分类器模型。将 PCA 方法与 ABBP 强分类器相结合,建立基于 PCA-ABBP 强分类器的矿区采空塌陷危险性预测模型。将该模型应用到北京西山某地采空塌陷数据预测中,模型预测结果与实际完全相符,而单一 BP 神经网络预测平均误差率为 17.14%,验证了所提出模型的有效性、可靠性及准确性,为矿区采空塌陷预测提供了一种新的途径,为矿区采空塌陷预测及其防治提供科学依据。

参考文献

[1]冯长根,李俊平,于文远,等. 东桐峪金矿空场处理机理研究[J]. 黄金,2002,(10):11-15.

[2]宫凤强,刘科伟,李志国. 矿区采空塌陷危险性预测的 Bayes 判别分析法[J]. 采矿与安全工程学报,2010,27(1):30-34.

[3]杜坤,李夕兵,刘科伟,等. 采空区危险性评价的综合方法及工程应用[J]. 中南大学学报(自然科学版),2011,42(9):2802-2811.

[4]付玉华,肖国喜. 采空区塌陷预测的多元识别模型[J]. 有色金属科学与工程,2012,(6):61-64.

[5]慎乃齐,杨建伟,郑惜平. 基于神经网络的采空塌陷预测[J]. 煤田地质与勘探,2001,29(3):42-44.

[6]GAO Y,JR E C A. Sinkhole risk assessment in Minnesota using a decision tree model[J]. Environmental Geology,2008,54(5):945-956.

[7]温廷新,孙红娟,徐波,等. 矿区采空塌陷危险性预测的 RS-SVM 模型[J]. 中国安全科学学报,2015,25(10):16-21.

[8]王海峰,李夕兵,董陇军,等. 基于支持向量机的采空区稳定性分级[J]. 中国安全生产科学技术,2014,(10):154-159.

[9]张长敏,董贤哲,祁丽华,等. 采空区地面塌陷危险性两级模糊综合评判[J]. 地球与环境,2005,33(z1):99-103.

[10]朱胜利. 矿山采空区塌陷预测方法研究[J]. 价值工程,2010,29(25):124-125.

[11]杨拉蒂,毕建武,贾进章. 基于主成分回归分析的瓦斯含量预测[J]. 世界科技研究与发展,2013,35(6):694-696.

[12]张瑾,卢国斌. 基于主成分 Fisher 判别分析的矿井通风系统安全评价[J]. 世界科技研究与发展,2013,35(4):501-504.

[13]毕建武,贾进章. 基于 SPSS 的 PCA-MRA 回采工作面瓦斯涌出量预测[J]. 安全与环境学报,2014,14(5):54-57.

[14]FREUND Y. Boosting a weak learning algorithm by majority[J]. Information and computation,1995,121(2):256-285.

[15]CORTES E A,MARTINEZ M G,RUBIO N G. Multiclass corporate failure prediction by Adaboost. M1[J]. International Advances in Economic Research,2007,13(3):301-312.

[16]ALFARO E,GARCIA N,GAMEZ M,et al. Bankruptcy forecasting: An empirical comparison of AdaBoost and neural networks[J]. Decision Support Systems,2008,45(1):110-122.

[17]ACHARYA U R,FAUST O,KADRI N A,et al. Automated identification of normal and diabetes heart rate signals using nonlinear measures[J]. Computers in biology and medicine,2013,43(10):1523-1529.