

# 基于机器学习的齿轮箱故障诊断

张 嵘, 朱 寅, 李 帅

上海理工大学机械工程学院, 上海

收稿日期: 2023年2月18日; 录用日期: 2023年3月17日; 发布日期: 2023年3月24日

## 摘 要

齿轮箱是用于增加输出扭矩或改变电机速度的机械装置, 被广泛应用于如汽车、输送机、风机等机械设备中。在齿轮箱运行过程中, 可以通过加速度传感器采集振动信号来判断齿轮箱是否出现异常, 传统对数据进行分析的方式耗时大, 取而代之的机器学习诊断方式能够有效的诊断其中的问题数据判断机械故障情况。基于此问题, 本文提出了基于支持向量机(SVM)、决策树分类和BP神经网络分类的三种分类模型, 建立故障数据分类模型, 分别进行故障数据的分类和对比, 最终得出最好的分类模型。

## 关键词

故障诊断, SVM, 决策树分类, BP神经网络

# The Fault Diagnosis of Gearbox Based on Machine Learning

Rong Zhang, Yin Zhu, Shuai Li

School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Feb. 18<sup>th</sup>, 2023; accepted: Mar. 17<sup>th</sup>, 2023; published: Mar. 24<sup>th</sup>, 2023

## Abstract

The gear box is a mechanical device used to increase the output torque or change the speed of the motor, which is widely used in mechanical equipment such as automobiles, conveyors and fans. During the operation of the gearbox, the vibration signal collected by the acceleration sensor can be used to judge whether the gearbox is abnormal. The traditional method of analyzing the data takes a lot of time, and instead, the machine learning diagnosis method can effectively diagnose and predict the problem data to judge the mechanical failure. To solve this problem, this paper puts forward three classification models based on support vector machine (SVM), decision tree classification and BP neural network classification, establishes the fault data classification model, classifies and compares the fault data respectively, and finally obtains the best classification model.

## Keywords

Fault Diagnosis, SVM, Decision Tree Classification, BP Neural Network

Copyright © 2023 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

## 1. 引言

齿轮箱是用于增加输出扭矩或改变电机速度的机械装置，被广泛应用于如汽车、输送机、风机等机械设备中，它由两个或多个齿轮组成，其中一个齿轮由电机驱动。齿轮箱作为机械的核心出现故障时会造成巨大的危害，在齿轮箱运行时多采用速度传感器采集振动信号，对采集的振动信号数据进行分析判断齿轮箱是否出现故障。赵嘉玮[1]提出基于 SVM 算法对草莓的各项特征数据进行分类，判断草莓的品阶，经测试该分类方法的准确率为 84.34%。杨丽萍[2]等人使用决策树分类算法对学生的所有成绩数据进行分类算法建模，判断学生的成绩通过情况。孙嘉谦等人提出基于 BP 神经网络的矿石分类方法，使用神经网络分类模型对矿石的特征数据进行分类建模，判断矿石的类型，经测试该模型的准确率为 90.6%。本文使用基于支持向量机(SVM)、决策树分类、BP 神经网络分类的三种分类模型对采集到的齿轮箱的各类故障数据进行分类建模，并使用搭建好的模型对 12 组不同故障状态的齿轮箱采集的信号进行分类测试，对三种模型的精度进行对比，得出其中精度最高的分类模型。

## 2. 三种分类模型介绍

### 2.1. SVM 算法

#### 2.1.1. SVM 算法概述

支持向量机(Support Vector Machine, SVM)由 Vapnik 首先提出，像多层感知器网络和径向基函数网络一样，支持向量机可用于多模式分类和非线性回归[1]。支持向量机的主要思想是建立个分类超平面作为决策曲面，使得正例和反例之间的隔离边缘被最大化；支持向量机的理论基础是统计学习理论，更精确地说，支持向量机是结构风险最小化的近似实现。这个原理基于这样的事实：学习机器在测试数据上的误差率(即泛化误差率)以训练误差率和一个依赖于 VC 维数(Vapnik-Chervonenkis dimension)的项的和为界，在可分模式情况下，支持向量机对于前一项的值为零，并且使第二项最小化。因此，尽管它不利用问题的领域内部问题，但在模式分类问题上支持向量机能提供好的泛化性能，这个属性是支持向量机特有的。

支持向量机具有以下优点：

- 1) 通用性：能够在很广的各种函数集中构造函数；
- 2) 鲁棒性：不需要微调；
- 3) 有效性：在解决实际问题中总是属于最好的方法之一；
- 4) 计算简单：方法的实现只需要利用简单的优化技术；
- 5) 理论上完善：基于 VC 推广性理论的框架

支持向量机的结构体系如下图 1：

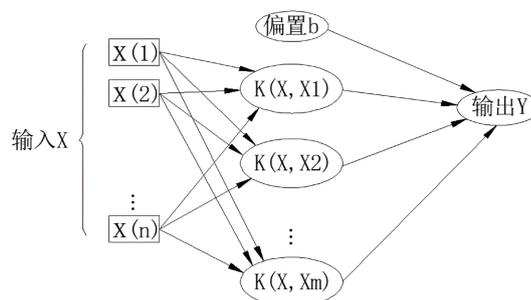


Figure 1. Support vector machine architecture  
图 1. 支持向量机结构体系

其中  $K$  为核函数，最主要的核函数为以下几种：

- 1) 线性核函数： $K(x, x_i) = x^T x_i$ ；
- 2) 多项式核函数： $K(x, x_i) = (\gamma x^T x_i + r)^p, \gamma > 0$ ；
- 3) 径向基核函数： $K(x, x_i) = \exp(-\gamma \|x - x_i\|^2), \gamma > 0$ ；
- 4) 两层感知器核函数： $K(x, x_i) = \tanh(\gamma x^T x_i + r)$ ；

### 2.1.2. 多分类 SVM

标准 SVM 是基于二元分类问题设计算法，无法直接处理多分类问题。利用标准 SVM 的计算流程有序地构建多个决策边界以实现样本的多分类，通常的实现为“一对多(one-against-all)”和“一对一(one-against-one)” [3]。一对多 SVM 对  $m$  个分类建立  $m$  个决策边界，每个决策边界判定一个分类对其余所有分类的归属；一对一 SVM 是一种投票法(voting)，其计算流程是对  $m$  个分类中的任意 2 个建立决策边界，即共有  $\frac{m(m-1)}{2}$  个决策边界，样本的类别按其对所有决策边界的判别结果中得分最高的类别选取。

一对多 SVM 通过对标准 SVM 的优化问题进行修改可以实现一次迭代计算所有决策边界。

## 2.2. 决策树分类算法

### 决策树分类算法概述

决策树算法是一种逼近离散函数值的方法。它是一种典型的分类方法，首先对数据进行处理，利用归纳算法生成可读的规则和决策树，然后使用决策对新数据进行分析。本质上决策树是通过一系列规则对数据进行分类的过程[4] [5]。

决策树算法构造决策树来发现数据中蕴涵的分类规则。如何构造精度高、规模小的决策树是决策树算法的核心内容。决策树构造可以分两步进行。第一步：决策树的生成；由训练样本集生成决策树的过程。一般情况下，训练样本数据集是根据实际需要有历史的、有一定综合程度的，用于数据分析处理的数据集。第二步：决策树的剪枝；决策树的剪枝是对上一阶段生成的决策树进行检验、校正和修下的过程，主要是用新的样本数据集(称为测试数据集)中的数据校验决策树生成过程中产生的初步规则，将那些影响预测准确性的分枝剪除[5]。

决策树算法的基本思想为树以代表训练样本的单个结点开始，如果样本都在同一个类，则该结点成为树叶，并用该类标记。否则，算法选择最有分类能力的属性作为决策树的当前结点。根据当前决策结点属性取值的不同，将训练样本数据集  $t_{11}$  分为若干子集，每个取值形成一个分枝，有几个取值形成几个分枝。针对上一步得到的一个子集，重复进行先前步骤，递归形成每个划分样本上的决策树。一旦一个属性出现在一个结点上，就不必在该结点的任何后代考虑它。

决策树优点在于分类精度高，生成的模式简单，对噪声数据有很好的健壮性。

### 2.3. BP 神经网络分类算法

#### BP 神经网络分类算法概述

BP 神经网络是前向神经网络的重要组成部分，同时也是核心部分，其主要的特点是：前后两层的神经元间通过权值相连，而且在前向传播的过程中，只有上一层的神经元输出会产生对下层神经元的影 响，并不会产生反馈信号。且上一层的一个神经元只有一种输出，并把输出结果送到下层的每个神经元。还有同层之间神经元的输入与输出是互不干扰，互不影响的[6]。根据这一特点可以得到前向型神经网络的一般模型，如下图 2 所示：

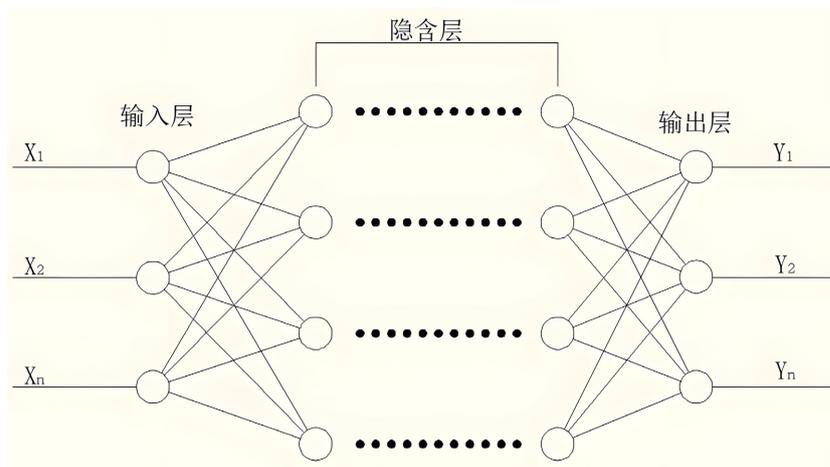


Figure 2. Multilayer neural network topology  
图 2. 多层神经网络拓扑结构

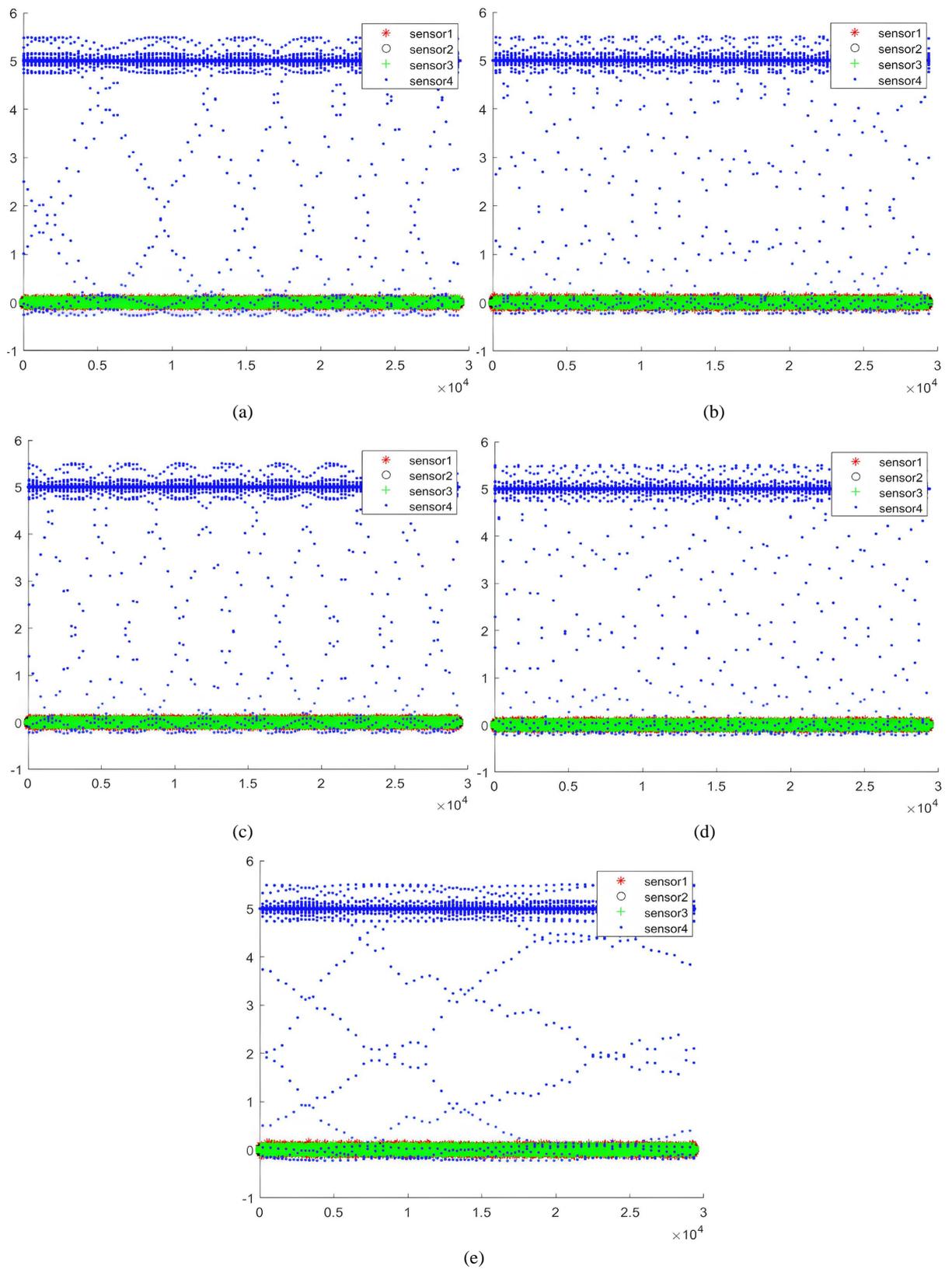
BP 神经网络是一种单向传播的多层神经网络，该神经网络除了输入层与输出层以外，在输出层与输入层之间还有一层或者多层的隐含层，同时同层之间的输入与输出都互不影响[7]。输入的信号通过输入层传入然后依次经过隐含层最后在输出层接受到输出信号，而且每一层的输出只会影响下一层的输入，其在传递到下一层的过程中的一般转移函数，为 Sigmoid 型转移函数， $\frac{1}{1+e^{-Bx}}$  ( $B > 0$ )，但在输出层输出结果时，结点的单元特性有时也可能是线性的。

### 3. 三种分类模型训练测试结果

使用三种分类模型前，需对传感器采集到的振动数据进行分析，研究正常情况和故障状态下振动数据的变化和差异，利用 Matlab 的 plot 功能分别绘制正常状态和故障状态下的散点数据图，观察不同状态下不同传感器采集到的数据的变化规律，如下图 3 所示，利用 EXCEL 对不同状态下不同传感器采集到的数据进行排序，寻找不同状态下各传感器采集到的振动数据的最大值和最小值，如表 1 所示。

从图中可以看出，不同状态下，传感器 1，传感器 2，传感器 3 数据变化不明显，不同状态下传感器 4 数据变化明显。

对五类故障状态数据进行分类算法建模后，使用搭建好的分类模型对 12 组齿轮箱的故障状态进行分类，分析不同模型检测精度。12 组齿轮箱的故障状态如下表 2。



**Figure 3.** (a) Normal conditions; (b) Fault state 1; (c) Fault state 2; (d) Fault state 3; (e) Fault state 4  
**图 3.** (a) 正常状态; (b) 故障状态 1; (c) 故障状态 2; (d) 故障状态 3; (e) 故障状态 4

**Table 1.** Sensor data analysis  
**表 1.** 传感器数据分析

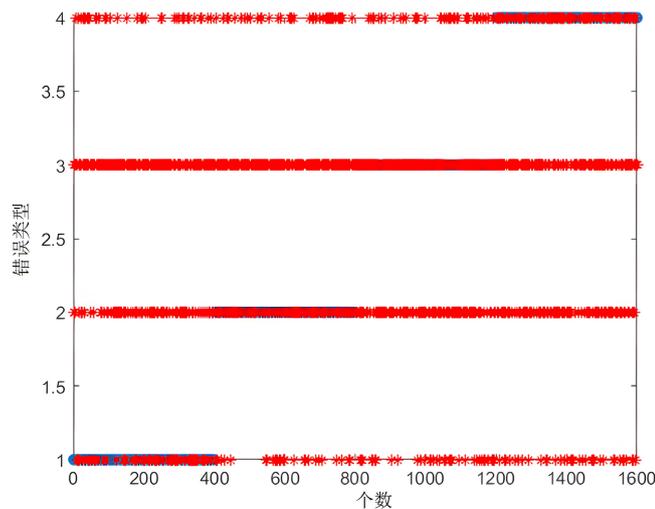
$\begin{matrix} 1 \\ 2 \end{matrix}$	传感器 1	传感器 2	传感器 3	传感器 4
正常	Max = 0.121	Max = 0.049	Max = 0.094	Max = 5.501
	Min = -0.132	Min = -0.053	Min = -0.097	Min = -0.284
故障一	Max = 0.152	Max = 0.005	Max = 0.127	Max = 5.509
	Min = -0.133	Min = -0.061	Min = -0.119	Min = -0.235
故障二	Max = 0.127	Max = 0.057	Max = 0.105	Max = 5.511
	Min = -0.129	Min = -0.049	Min = -0.102	Min = -0.237
故障三	Max = 0.130	Max = 0.053	Max = 0.102	Max = 5.51
	Min = -0.126	Min = -0.061	Min = -0.114	Min = -0.229
故障四	Max = 0.159	Max = 0.074	Max = 0.126	Max = 5.505
	Min = -0.151	Min = -0.062	Min = -0.114	Min = -0.231

**Table 2.** 12 groups of gearbox fault states  
**表 2.** 12 组齿轮箱故障状态

样本编号	故障状态	样本编号	故障状态	样本编号	故障状态	样本编号	故障状态
齿轮箱 1	正常工况	齿轮箱 4	正常工况	齿轮箱 7	故障状态 4	齿轮箱 10	故障状态 3
齿轮箱 2	故障状态 1	齿轮箱 5	故障状态 4	齿轮箱 8	故障状态 4	齿轮箱 11	故障状态 2
齿轮箱 3	正常工况	齿轮箱 6	故障状态 2	齿轮箱 9	正常工况	齿轮箱 12	故障状态 3

### 3.1. 多分类 SVM 训练及测试结果

本文将各错误类型的第 1 行至第 29,000 行数据代入 LIBSVM 工具箱进行训练, 将第 29,001 行至 29,400 行作为测试数据进行精度检验。训练完成结果精度为 30%, 利用测试数据进行检验, 检验结果如下图 4:



**Figure 4.** SVM prediction graph  
**图 4.** SVM 预测图

从上可以看出,预测精度并不高,主要是 SVM 主要适合小样本的训练和预测,本次的训练数据太大,增加了空间复杂度和时间复杂度,导致训练结果差,精度不高。可以通过优化训练参数集,通过减少训练参数来提高训练精度。

由于训练时 SVM 对大数据样本的精度较差,不使用 SVM 进行比较测试。

## 3.2. 决策树分类算法训练及测试结果

### 3.2.1. 决策树分类算法训练结果

本文使用 Matlab 中的分类学习器工具箱,将五种典型故障状态的数据作为训练数据导入,建立相应的分类模型,将另外的五组故障状态数据作为验证数据,判断该分类模型的精度。采用决策树中的可优化树,当最大分裂数为 30,优化器选择贝叶斯优化时,训练精度达到最高为 68%,训练得出的散点图,最小分类误差,ROC 曲线,如下图 5~7 所示:

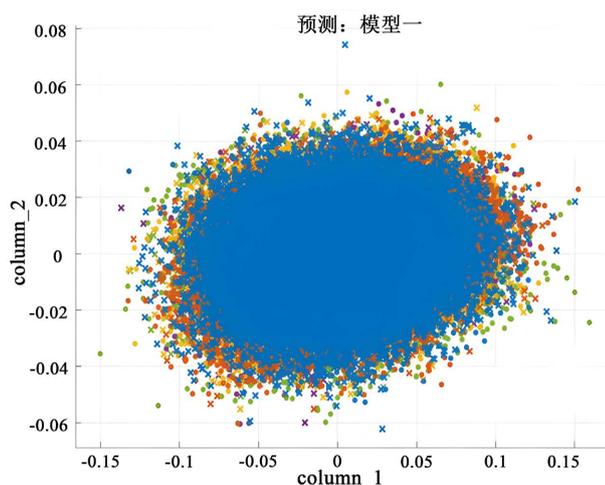


Figure 5. Scatter diagram

图 5. 散点图

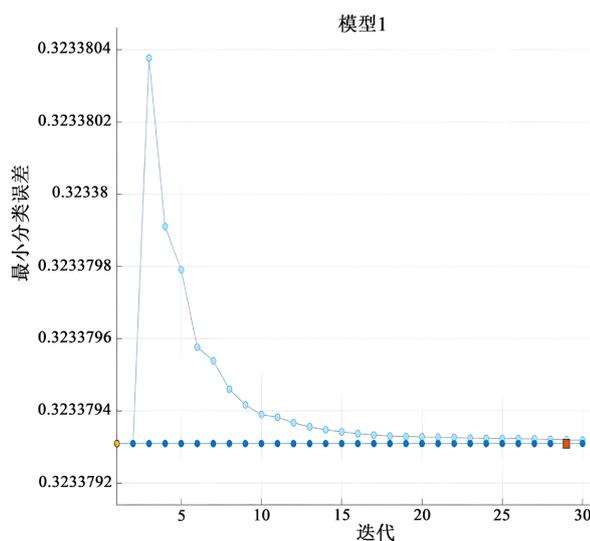


Figure 6. Minimum classification error

图 6. 最小分类误差

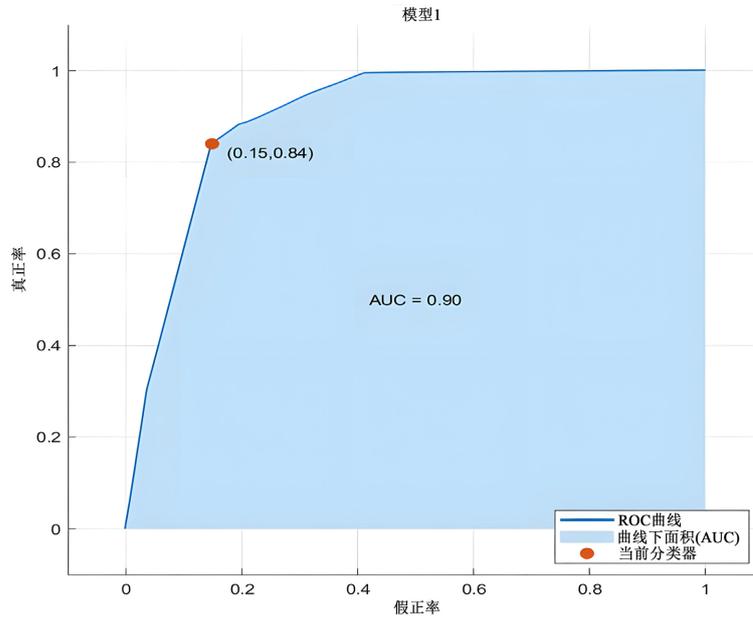


Figure 7. ROC curve  
图 7. ROC 曲线

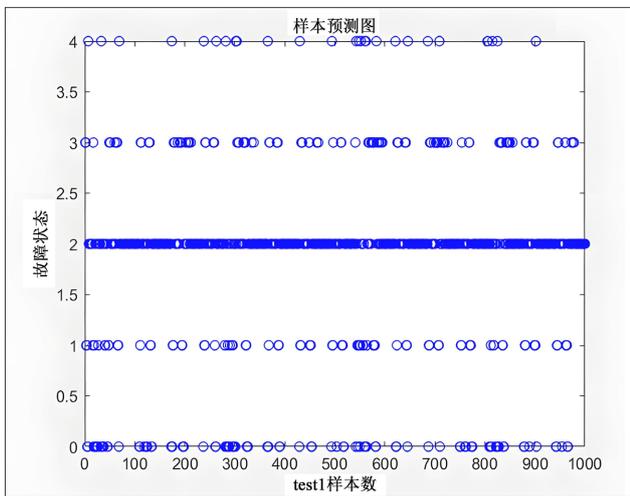
### 3.2.2. 决策树分类算法分类测试结果

使用搭建好的决策树分类模型对 12 组齿轮箱进行分类，对每个齿轮箱采集的数据进行分类，取其中数值最大的状态数据作为齿轮箱的状态数据。故障状态用数字 0, 1, 2, 3, 4 表示(0: 正常, 1: 故障 1, 2: 故障 2, 3: 故障 3, 4: 故障 4)，每个齿轮箱数据分类结果如下表 3:

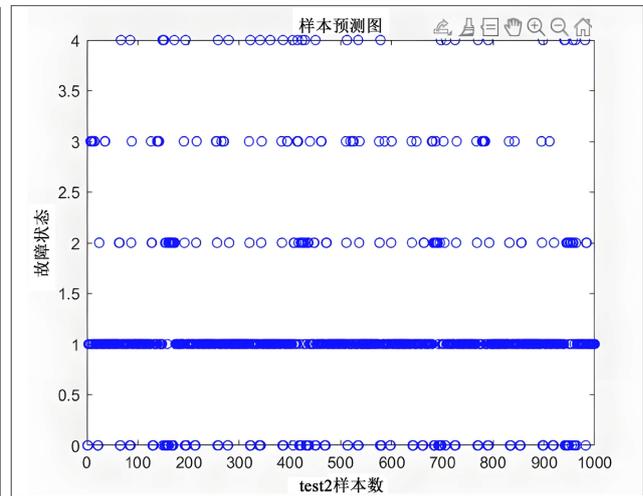
Table 3. Decision tree data classification of gearbox  
表 3. 齿轮箱决策树数据分类

齿轮箱	0 正常	1 故障 1	2 故障 2	3 故障 3	4 故障 4
齿轮箱 1	91	77	682	119	31
齿轮箱 2	100	735	73	58	31
齿轮箱 3	791	64	70	28	49
齿轮箱 4	759	74	44	32	39
齿轮箱 5	95	731	61	80	33
齿轮箱 6	97	60	708	106	30
齿轮箱 7	161	321	63	27	428
齿轮箱 8	189	61	71	20	659
齿轮箱 9	72	154	120	633	20
齿轮箱 10	91	157	134	590	27
齿轮箱 11	186	61	86	24	642
齿轮箱 12	228	551	75	47	99

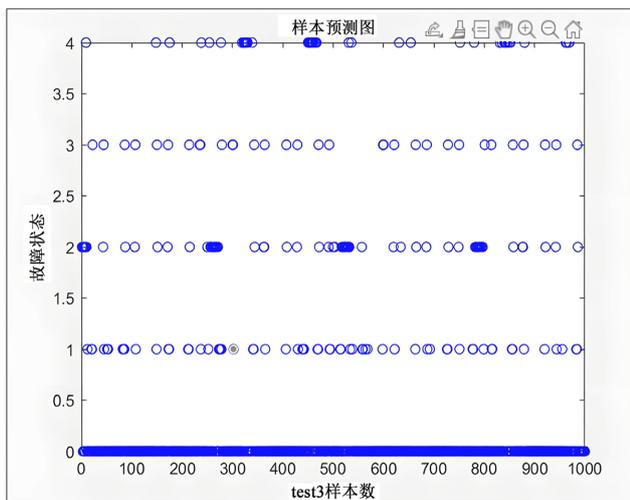
将上述的数据进行可视化处理如下图 8:



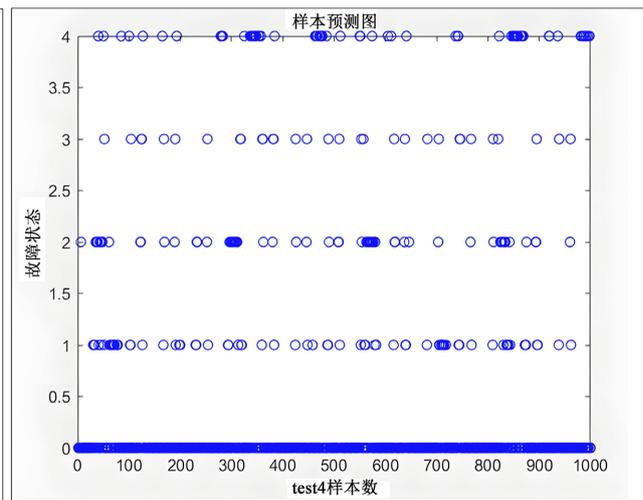
(a) 齿轮箱 1 数据分类



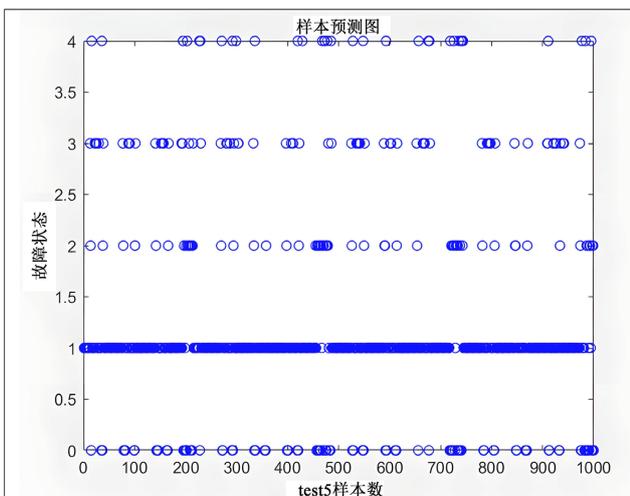
(b) 齿轮箱 2 数据分类



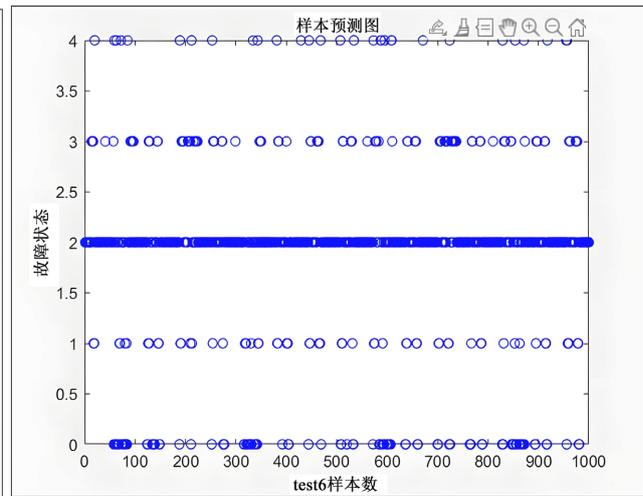
(c) 齿轮箱 3 数据分类



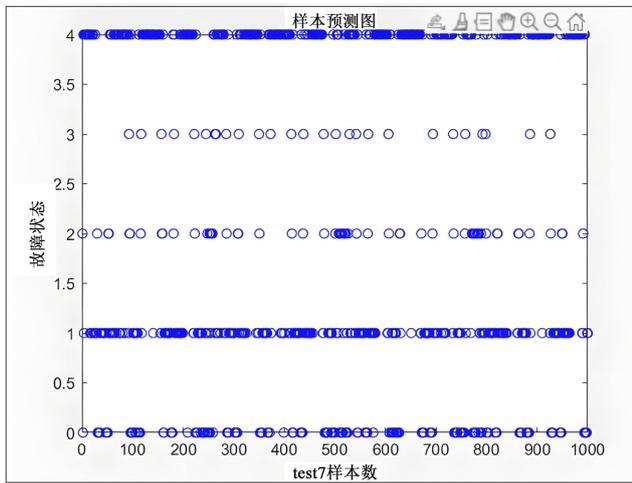
(d) 齿轮箱 4 数据分类



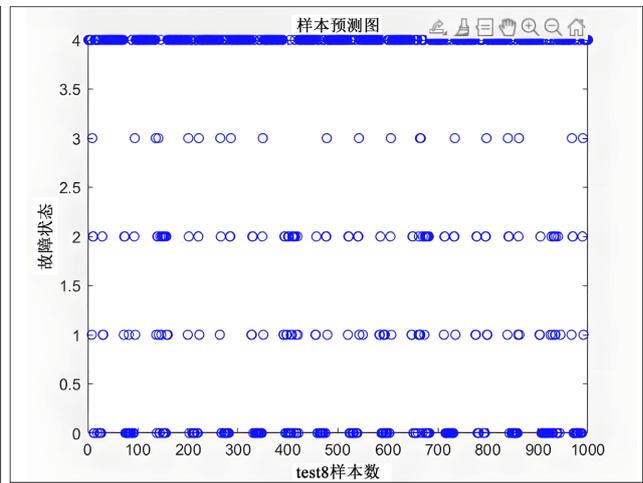
(e) 齿轮箱 5 数据分类



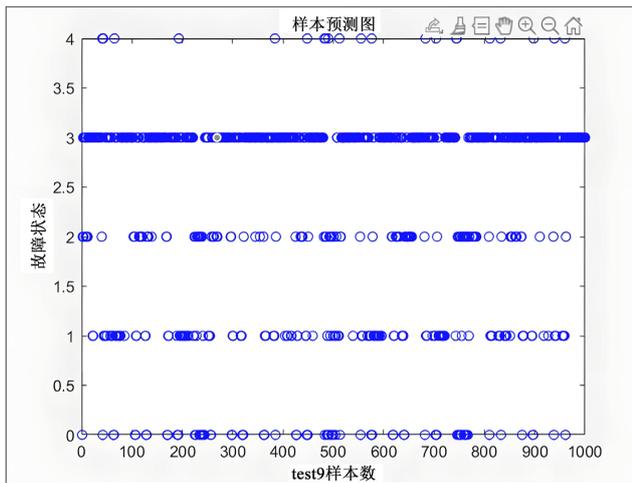
(f) 齿轮箱 6 数据分类



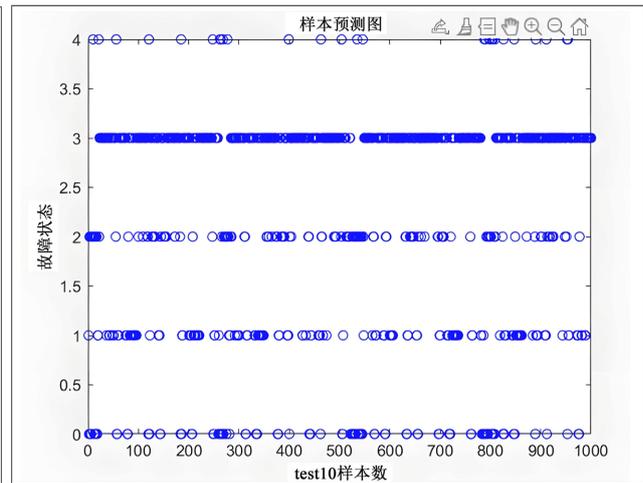
(g) 齿轮箱 7 数据分类



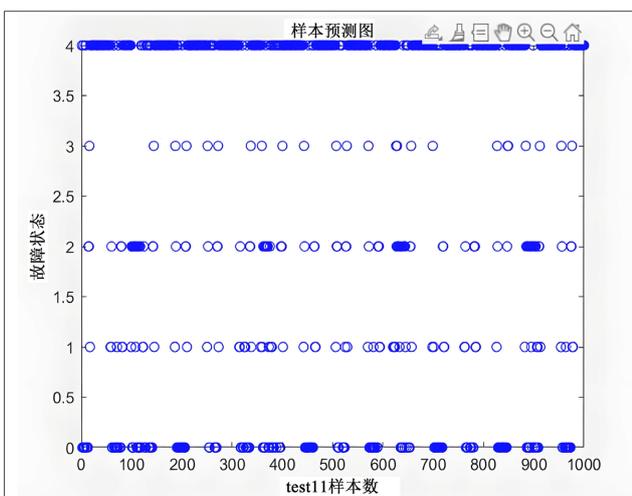
(h) 齿轮箱 8 数据分类



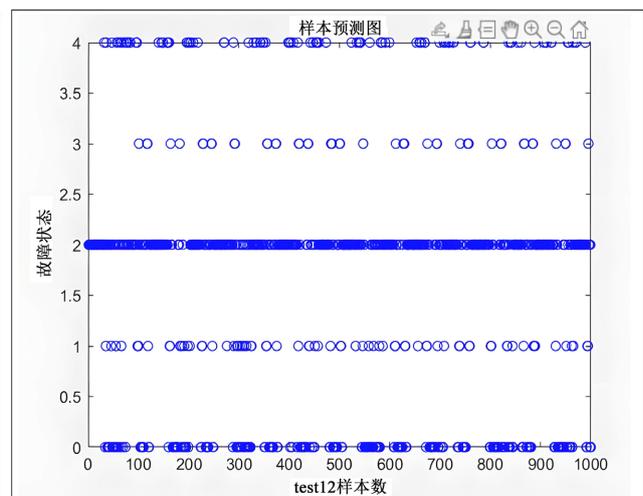
(i) 齿轮箱 9 数据分类



(j) 齿轮箱 10 数据分类



(k) 齿轮箱 11 数据分类



(l) 齿轮箱 12 数据分类

Figure 8. 12 groups of gearbox data classification diagram

图 8. 12 组齿轮箱数据分类图

据上图可知：使用决策树算法对 12 组齿轮箱进行分类后，各齿轮箱的故障状态如下表 4：

**Table 4.** Classification of decision tree algorithm for gearbox fault state

**表 4.** 齿轮箱故障状态决策树算法分类

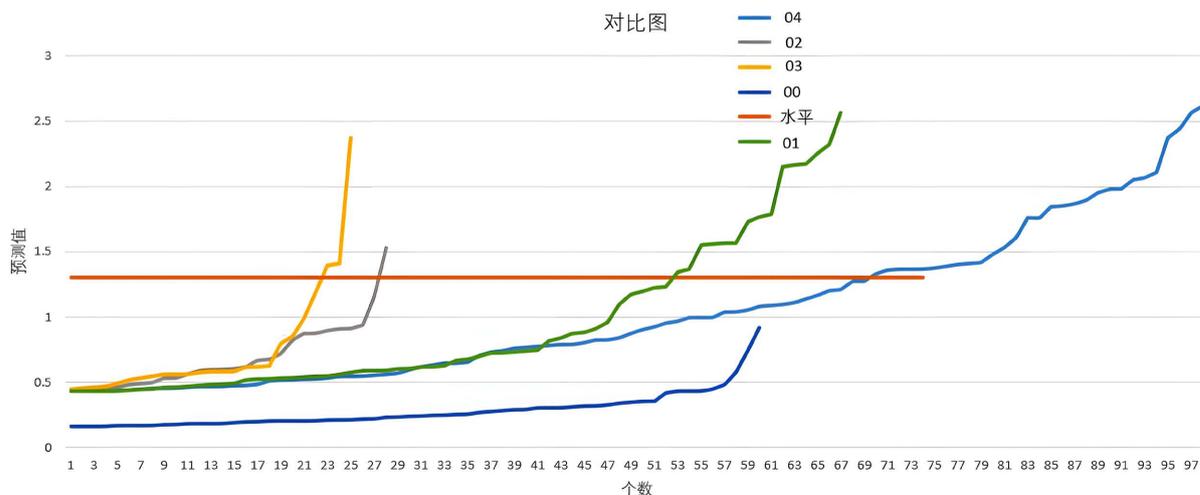
样本编号	故障状态	样本编号	故障状态	样本编号	故障状态	样本编号	故障状态
齿轮箱 1	故障状态 2	齿轮箱 4	正常工况	齿轮箱 7	故障状态 4	齿轮箱 10	故障状态 3
齿轮箱 2	故障状态 1	齿轮箱 5	故障状态 1	齿轮箱 8	故障状态 4	齿轮箱 11	故障状态 4
齿轮箱 3	正常工况	齿轮箱 6	故障状态 2	齿轮箱 9	故障状态 3	齿轮箱 12	故障状态 2

由给出的齿轮箱故障状态表可看出：齿轮箱 1、齿轮箱 5、齿轮箱 9、齿轮箱 11、齿轮箱 12 状态判断错误，建立的决策树算法模型的分类精度为 58.3%。

### 3.3. BP 神经网络分类算法训练及测试结果

本次采用 Matlab 中的分类学习器工具箱，将训练数据导入，将五种状态数据的前 29,000 (train\_error) 组作为训练，后 400 组作为测试验证[8] [9]。采用 BP 神经网络，当层数为 60，优化步长为 17 时，相关性达到 0.93，有较强的可信度[10]。

将五种状态数据作为验证测试组，利用训练好的神经网络进行反预测，将预测数据与训练数据对比，画出故障趋势折线图进行对比[11]，得到故障状态，故障趋势折线图和部分预测折线图如下图 9、图 10：



**Figure 9.** Fault trend line chart

**图 9.** 故障趋势折线图

经过各数据对比后确定每个齿轮箱的故障状态如下表 5。

由给出的齿轮箱故障状态表可看出：齿轮箱 7 状态判断错误，建立的决策树算法模型的分类精度为 91.7%。

## 4. 总结

本文使用了三种分类算法对齿轮箱各故障状态的传感器数据进行分类建模，由于 SVM 算法主要适合小样本的训练和预测，本次的训练数据太大，增加了空间复杂度和时间复杂度，导致训练结果差，精度

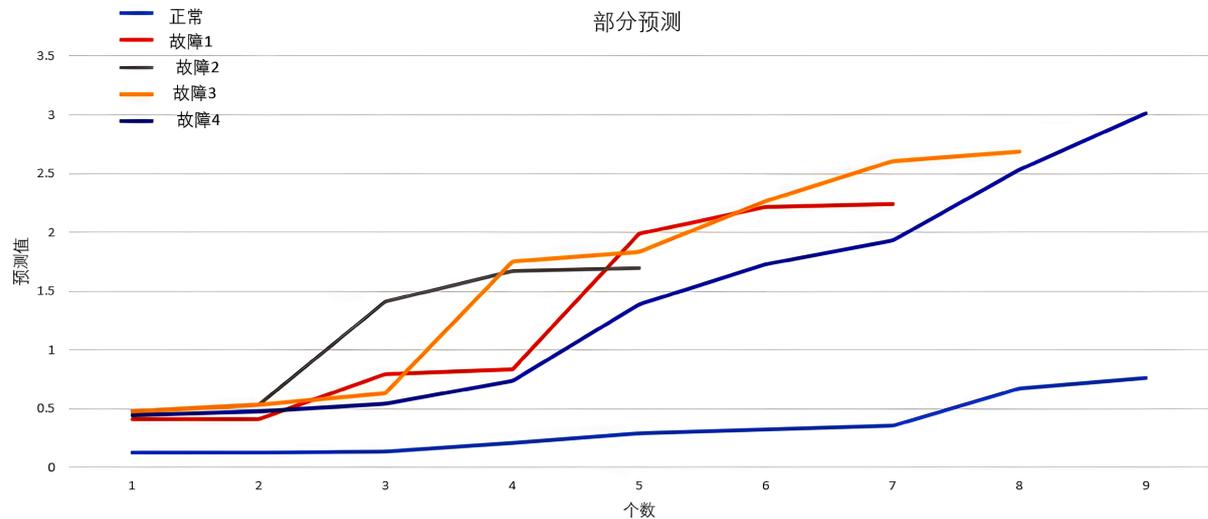


Figure 10. Partial forecast line chart

图 10. 部分预测折线图

Table 5. Classification of gearbox fault state by BP neural network

表 5. 齿轮箱故障状态 BP 神经网络分类

样本编号	故障状态	样本编号	故障状态	样本编号	故障状态	样本编号	故障状态
齿轮箱 1	正常工况	齿轮箱 4	正常工况	齿轮箱 7	故障状态 3	齿轮箱 10	故障状态 3
齿轮箱 2	故障状态 1	齿轮箱 5	故障状态 4	齿轮箱 8	故障状态 4	齿轮箱 11	故障状态 2
齿轮箱 3	正常工况	齿轮箱 6	故障状态 2	齿轮箱 9	正常工况	齿轮箱 12	故障状态 3

不高。决策树算法训练精度为 68%，在进行 12 组齿轮箱分类时，其精度为 58.3%，BP 神经网络分类算法训练精度为 93%，在进行 12 组齿轮箱分类时，其精度为 91.7%。经对比可得出，BP 神经网络分类算法拥有较高的训练精度，且实际使用时分类精度较高。本研究为齿轮箱故障检测数据的分类及判断齿轮箱故障类型提供了一个较好的方案，对于辅助人员进行齿轮箱故障诊断具有重要的现实意义。

## 参考文献

- [1] 赵嘉玮. 基于 SVM 算法的目标分类筛选方法研究[J]. 无线互联科技, 2022, 19(23): 123-125.
- [2] 杨丽萍, 郭宏升. 决策树分类算法在课程成绩预测中的应用[J]. 电子测试, 2022, 36(17): 56-58. <https://doi.org/10.16520/j.cnki.1000-8519.2022.17.022>
- [3] 芦伟东. 基于 SVM 的边境频谱监测数据分类预测[J]. 中国无线电, 2022(9): 44-46.
- [4] 孙佩, 汪权方, 易洁伟, 康全国, 张驰, 尹伟, 袁知洋. 基于决策树分类方法的小麦油菜种植范围提取[J]. 农业与技术, 2022, 42(24): 7-11. <https://doi.org/10.19754/j.nyyjs.20221230002>
- [5] 申明尧, 韩萌, 杜诗语, 孙蕊, 张春砚. 数据流决策树集成分类算法综述[J]. 计算机应用与软件, 2022, 39(9): 1-10.
- [6] 孙嘉谦, 李志刚, 皮晓宇, 杨永辉, 高闯. 基于神经网络的矿石分类方法[J]. 有色金属(选矿部分), 2023(1): 22-27.
- [7] Zhi, X., Ye, J.S., et al. (2009) BP Neural Network with Rough Set for Short-Term Load Forecasting. *Expert Systems with Applications: An International Journal*, **36**, 273-279. <https://doi.org/10.1016/j.eswa.2007.09.031>
- [8] Toqeer, R.S. and Bayindir, N.S. (2003) Speed Estimation of an Induction Motor Using Elman Neural Network. *Neurocomputing*, **55**, 727-730. [https://doi.org/10.1016/S0925-2312\(03\)00384-9](https://doi.org/10.1016/S0925-2312(03)00384-9)
- [9] Li, A.P., Li, A.Y., Xiong, B.Q., et al. (2014) Application of a Hybrid Quantized Elman Neural Network in Short-Term

---

Load Forecasting. *International Journal of Electrical Power & Energy Systems*, **55**, 749-759.  
<https://doi.org/10.1016/j.ijepes.2013.10.020>

- [10] 李萍, 曾令可, 税安泽, 等. 基于 MATLAB 的 BP 神经网络预测系统的设计[J]. 计算机应用与软件, 2008, 25(4): 3.
- [11] 蔚治国. 基于改进 BP 神经网络的样本分类方法[J]. 自动化与仪器仪表, 2016(6): 237-239.  
<https://doi.org/10.14016/j.cnki.1001-9227.2016.06.237>