Journal of Sensor Technology and Application 传感器技术与应用, 2026, 14(1), 145-153
Published Online January 2026 in Hans. https://www.hanspub.org/journal/jsta
https://doi.org/10.12677/jsta.2026.141015

Hans汉斯

关于轴承故障位置判别的研究

刘　悦，刘梦茹，吕欣达，孙菊贺*

沈阳航空航天大学理学院，辽宁 沈阳

收稿日期：2025年12月13日；录用日期：2026年1月6日；发布日期：2026年1月16日

摘　要

本文基于轴承四种故障状态数据与轴承正常状态数据进行分析，从而形成对无标签轴承是否故障以及故障位置的类型划分。首先，利用Python清洗数据后，结合MATLAB将原始数据按照所选取的故障状态类型分别进行特征提取，对所提取数据进行训练集和测试集的划分，并使用一种新的基于欧氏距离和中位数的判别方法进行判别。同时，利用SPSS使用最邻近算法训练划分出的训练集，用测试集进行验证。其次，将两种算法结合聚类算法迁移到未知标签的数据，最后借鉴KNN算法的投票机制进行改进，并对数据进行具体划分，改进分类结论。

关键词

频谱能量分析，最邻近算法思路，聚类算法，分层抽样

Study on Bearing Fault Location Identification

Yue Liu, Mengru Liu, Xinda Lyu, Juhe Sun*

School of Science, Shenyang Aerospace University, Shenyang Liaoning

Received: December 13, 2025; accepted: January 6, 2026; published: January 16, 2026

Abstract

This study is based on data from four bearing fault states and normal bearing state data, aiming to classify unlabeled bearings in terms of fault presence and fault location type. First, after data cleaning using Python, feature extraction is performed on the raw data in MATLAB according to the selected fault state types. The extracted data are divided into training and test sets, and a novel discrimination method based on Euclidean distance and median is applied for classification. Meanwhile, the k-nearest neighbors (KNN) algorithm is implemented via SPSS to train the divided training set, with the test set used for validation. Subsequently, the two algorithms are integrated with a

*通讯作者。

文章引用: 刘悦, 刘梦茹, 吕欣达, 孙菊贺. 关于轴承故障位置判别的研究[J]. 传感器技术与应用, 2026, 14(1): 145-153. DOI: 10.12677/jsta.2026.141015

clustering algorithm and transferred to unlabeled data. Finally, inspired by the voting mechanism of KNN, improvements are made to refine the data partitioning and enhance the classification results.

Keywords

Spectral Energy Analysis, The Concept of K-Nearest Neighbors Algorithm, Cluster Algorithm, Stratified Sampling

1. 引言

在高铁网全面成型的当下，轴承健康状态直接联系着行车安全与运营效益。轴承在时速 350 公里工况下需承受每秒 5000 次以上交变载荷，叠加轨道激励、风沙侵蚀等干扰，故障率占走行系统故障总数的60%以上。这些故障通常进展迅速，若不及时发现并处理，可能导致相当严重的后果。目前轴承状况检测大部分依赖于人工判断，实时响应能力较差，在高铁日益发展的今天难以满足检测需求。而且实际场景中，振动信号受噪声与干扰影响、在途故障数据稀缺导致训练数据失衡，一定程度上制约了深度学习模型的识别效果与工程转化。鉴于台架实验轴承数据丰富、标签完备且故障机理相近，相关数据常被用于建立模型。

2. 基于有标签数据集的模型建立

2.1. 模型参数选取

在研究高铁轴承故障时，除了正常(N)外有三种故障是最频繁发生的[1]，分别是外圈故障(OR)、内圈故障(IR)、滚动体故障(B)，于是以这三种故障状态为研究对象。对于第一种故障状态，它涉及三种不同故障部位，把这三种不同故障部位一起纳入考虑，分别是传感器位于故障中心(OR centered)、传感器位于故障对面(OR opposite)和传感器位于故障正交(OR orthogonal)。在本研究中假定故障模式是单一存在，不存在其他复合故障模式，且传感器位置影响故障发生。本文所选用数据信号为离散的信号，需要对其进行数据清洗，如降噪和去除极端数值，且为了配合后续研究需要对数据进行处理，如连续化和数据模式转换。由于轴承的几何参数包括轴承参数、转速和频率等对采集出的数据都有影响，因此本研究将轴承的几何参数与转速视为模型的前置输入，仅用于构建阶段计算核心特征，在后续的故障诊断与状态评估中，模型将直接输出分析结果，从而规避了对这些底层参数影响的直接归因分析。

本研究采用的数据[2]仅有部分高铁数据，其源域数据有数据编号、基座/驱动端/风扇端采样振动信号(BA/DE/FE)和轴承转速(RPM)，目标域为位置标签的十六个振动信号数据，其所有标签均未知，仅有振动信号数据，其编号设为 A 到 P。

2.2. 傅里叶变换与奈奎斯特频率处理能量相关数据

在处理数据时，基于轴承的几何参数如下：

滚动体数量：$n = 9$。

滚动体直径：d (SKF6205：0.3126 英寸；SKF6203：0.2656 英寸)。

节圆直径：D (SKF6205：1.537 英寸；SKF6203：1.22 英寸)。

转速：rpm。

轴频：

$$f_r = \frac{rmp}{60}$$

外圈故障频率(BPFO)：

$$f_{\mathrm{BFRO}} = \frac{n}{2} \cdot f_r \cdot \left(1 - \frac{d}{D}\right)$$

内圈故障频率(BPFI)：

$$f_{\mathrm{BPFI}} = \frac{n}{2} \cdot f_r \cdot \left(1 + \frac{d}{D}\right)$$

滚动体故障频率(BSF)：

$$f_{\mathrm{BSF}} = \frac{D}{d} \cdot f_r \cdot \left[1 - \left(\frac{D}{d}\right)^2\right]$$

对于清洗后的数据进行时域频域、谱特征等相关特征提取。对于能量相关数据[3] [4]，由于样本数据均为离散的值，而计算能量需要用可计算的连续数据频率，于是采用傅立叶变换将其转换成连续的可计算数据，从而进行处理。最终可以得出样本时域频域特征以及能量相关数据。

快速傅里叶变换(FFT)公式为：

$$X[k] = \sum_{n=0}^{\infty} x[n] \cdot \mathrm{e}^{-j\frac{2\pi}{N}kn}, k = 0,1,\cdots,N-1$$

奈奎斯特频率与采样定理的频域分析：

奈奎斯特频率(Nyquist frequency)是离散信号处理系统中的关键参数，定义为采样频率 f_s 的一半 $2f_N = f_s$，同时也是信号最高频率 $f_{\max}$ 的两倍($f_s \geq 2f_{\max}$)。在频域中，奈奎斯特频率是信号频谱周期性延拓的边界。若信号频率超过 f_N，会导致频谱混叠。

采样信号为：

$$x_s(t) = x_a(t) \cdot \sum_{n=-\infty}^{\infty} \delta(t - nT)$$

其频谱为：

$$x_s(f) = \frac{1}{T} + \sum_{k=-\infty}^{\infty} x_a\left(f - \frac{k}{T}\right)$$

其中，$Tf_s = 1$，f_s 为采样频率。当 $f_s \geq 2f_{\max}$，各个周期频谱不重叠。若 $f_s < 2f_{\max}$，相邻频谱周期重叠，导致高频成分被误认为低频：

$$f_{alias} = \left|f - kf_s\right| (k \text{ 为整数})$$

在确定奈奎斯特点时，依据是信号在单边谱上的最高频率分量。

单边谱是指仅包含正频率部分的频谱表示，通常用于实信号。通过傅里叶级数三角形式(正弦/余弦项)或 FFT 变换后的正频率部分生成，直观反映信号的主要频率成分及其能量分布。单边谱通过三角形式展开，仅保留正频率分量：

$$x(t)=a_0+\sum_{n=1}^{\infty}\left[a_n\cos\left(n_0^{\omega}t\right)+b_n\sin\left(n_0^{\omega}t\right)\right]$$

其中，幅值为：

$$A_n=\sqrt{a_n^2+b_n^2}$$

相位为：

$$\phi_n=\arctan\left(-\frac{b_n}{a_n}\right)$$

2.3. 建立评估模型

在通过傅里叶变换对数据进行处理后，取故障频率前后 10%带宽作为故障考察段，可以得出故障频带能量，进而计算出频谱总能量、故障能量占比(即故障能量相对大小)和频谱重心频率。根据这个流程可以求出不同部位故障时的频率、故障位置附近能量、频谱总能量、故障能量相对大小、能量集中的频率位置这些相关参数。本研究主要研究故障部位，故只计算了在故障部位的数据。此外，还计算了极大值、极小值、峰度、偏度、平均值和标准偏差进行辅助判断，增强结果的可信性。

本研究采用两种方法进行，一种是对于标签内的数据，根据故障时的频率、故障位置附近能量、频谱总能量、故障能量相对大小、能量集中的频率位置参数进行判别，对于每组数据按故障进行分类，将处理后数据的中位数作为判断基准线，依照极值差额大小的 80%为波动范围大小，从而得到各个参数波动范围表，因此即可判断该数据是否大体在这一波动范围，从而得出该数据大概率是否是这一故障。另一种是直接采用 KNN 算法，根据故障发生时的频率、故障位置附近能量、频谱总能量、故障能量相对大小、能量集中的频率位置等参数随机划分训练集和测试集，通过训练集预测出测试集数据的分类，从而得出该模型是否有效。

前一种方法需要针对标签内的数据进行分层抽样。分层抽样的要点在于确保数据集中每个类别在训练集和测试集中的比例一致，在此基础上按照标签进行抽样，因此本研究在每个集合均取 70%的值做训练集，剩余 30%做测试集，从而得出该模型是否有效。后一种方法按照同样比例则直接使用 SPSS 得出结果。

2.4. 算法的建立与评估模型

进行数据训练前要先进行标准化，本研究选用 Z-score 标准化方法：

$$x'=\frac{x-\bar{x}}{\sigma}$$

其中，$\bar{x}$ 为训练集的均值，σ 为训练集的标准差。测试集利用训练集的均值和标准差进行转换，确保训练集与测试集的相互独立性。

使用最邻近算法[5]进行数据训练，最邻近算法是一种基于实例的监督学习算法，其核心思想是通过计算待分类样本与训练集中各个样本的距离，选取 k 个邻近样本，根据这些邻近样本的类别进行多数表决或加权投票来决定待分类样本的类别。

1) 距离度量：欧式距离公式为：

$$d(x,y)=\sqrt{\sum_{i=1}^{n}\left(x_i+y_i\right)^2}$$

2) 决策规则：新样本的类别由 k 个最近邻中占比最大的类别决定。本文取 $k=3$。除此之外，由于在

判定无标签数据时有许多影响因素，假定无标签数据的参数均为 DE 轴承的参数(因为样本源域数据占比大)，并结合中位数和判定的区间范围等因素考虑。

3) 决策规则 k 的选取：将源域能量数据基于 KNN 算法求解，并根据后文的成功概率进行计算。由于样本数据较少，且为了保持 KNN 算法稳定 k 取奇数，且最好不取 1，我们将所有数据以七三划分进行测试，得出成功率如下表 1 所示。

Table 1. k-value selection table
表 1. k 值选择表

类别	样本数	$k=3$ 准确数	$k=5$ 准确数	$k=7$ 准确数
B	36	30	23	27
IR	36	36	36	32
OR centered	30	24	20	4
OR orthogonal	27	16	21	23
OR opposite	20	18	15	10

根据表中数据我们可以看出，k 为 3 时最好，且使用 SPSS 时在选择参数 k 时勾选自动选择和 V 折交叉验证，结论依旧为 3，故本文取 $k=3$。

建立跨域诊断轴承故障模型，通过源域数据(12 Hz 和 48 kHz 混合采样数据)中百分之七十的数据分为训练集，百分之三十的数据分为测试集，识别四种轴承故障状态：Normal (正常状态)、OR (外圈故障状态)、IR (内圈故障状态)、B (滚动体故障状态)，其中假设不存在其他混合故障状态，且 OR 故障分为三种不同故障部位。使用最邻近算法训练测试集，并将训练出的模型对测试集进行测试。统计轴承状态成功率和精确率模型如下：

① 成功率：

$$Q=\frac{\mathrm{TRUE}}{\mathrm{ALL}}$$

其中，TURE 为正确的数量，ALL 为该分类中参与判别的数量，分别判断各种故障的判别成功概率，并进行比较，从而得出结论是否符合。

② 精确率：

$$P=\frac{\sum_{i=1}^{n}Q_iN_i}{\sum_{i=1}^{n}N_i}$$

其中，n 为总类数，N_i 为第 i 组数据样本数量，Q_i 为第 i 组数据成功率，在判断各种故障的判别成功率的基础上求出该模型的精确率，从而得出结论是否符合。由于 KNN 算法选取数据时是随机选取的，无法判断究竟选取了哪些数据，因此选取总体的精确率来诊断。本研究中模型总体(包括训练集)成功率约为 83.2%，根据成功率公式可以得出综合精确率为 45%，假设模型基本成立。同理，根据中位数和能量得出三个故障对应四种能量共 12 组数据，大于 7 组在范围内视为成功，成功率如下表 2 所示。

B 和 IR 故障诊断本故障概率较高，推测是同组内差异较大以及判别范围选取过大的影响，后文将利用其他方法综合判别，综合来看基于中位数的欧氏距离范围可以进行辅助判别。

Table 2. Success rate summary table
表 2. 成功率汇总表

故障类别	成功率
B	100%
IR	100%
OR centered	60%
OR orthogonal	71.4%
OR opposite	60%

3. 未知故障标签数据的分类

基于最邻近算法相关思想的评估模型在未知故障标签数据中的运用与拓展

未知标签数据是十六个以 A 到 P 字母命名的数据集，没有详细地标注数据所属的工作状态，只是提供了列车不同部位振动的信号数据、采集时间和采样频率等前置条件，无法直观判断数据所属的部位状态是否故障。因此，本研究将根据训练集训练出的模型进行优化迁移和标签分类。根据建立的模型将目标域的数据进行数据清洗和傅里叶变换，并基于双边频谱幅度与相位谱结合单边谱奎奈斯特频率与采样定理进行计算能量相关数据，得出目标域的能量数据后进行分析。除此之外，还用了上文提到的峰度和偏度等相关数据。类似前面所用到的算法，不同的是，在初次判别分类后，首先运用 KNN 算法进行粗略估算，其次使用 k-means 算法[6]进行辅助分类，最后采用一种简单的基于中位数的判别方法进行最终分析。

首先对数据进行清洗后使用 SPSS 对模型进行 KNN 求解(k 取 3)，得出结论如下表 3 所示。

Table 3. Initial analysis table of k-nearest neighbors algorithm
表 3. 最邻近算法初次分析表

类别	编号
OR orthogonal	ABDFGHKLMNOP
B	CEIJ

继续进行聚类算法(下表采用 k-means 算法，直接使用 SPSS 软件，数据在使用前需要先使用 Z-score 方法进行归一化)进行模型分析，由于观察数据大致分为七类，本次聚类使用 $k = 7$，利用软件进行求解，可得出的初次分类如下表 4 所示。

Table 4. k-means clustering analysis table
表 4. k 均值聚类分析表

类别	编号
1	A
2	BGLO
3	CIJ
4	DMP
5	E
6	F
7	HKN

从表 4 可以看出，KNN 方法精度并不是很高。但对比聚类分析的结果，可以发现 KNN 的分类与聚类分析的结果是相对吻合的，所以可以将其用于粗糙的初步判别。但为了得出更精确的结果，考虑用以下方法进行进一步验证。

对于数据，我们根据不同条件(DE、FE 以及频率)进行分组，对于每个部位(DE 或 FE)，其未知标签数据计算出的能量相关数据也有很多不同之处，需要根据部位进行分别计算，求出能量相关数据后才能进行使用。对于不同条件，比较在该条件下故障发生时各未知标签数据与原数据的中位数距离。

以中位数为中点，借鉴 KNN 的投票机制，以极值差百分之四十大小为上(下)波动范围，对于在判别范围内的数据，进行数据投票，将其分类判别情况记为 1，对于不在此范围内的数据，将其分类判别情况为 0，可以得到是否在此判别范围的 0~1 矩阵，从而对数据进行直观分析。

对于无标签数据得出的 0~1 矩阵，将其视作算法判别的辅助工具，若 0~1 矩阵中 0 和 1 较多，且最终得出的数据每一个部位每行均小于 3 个 1，说明这组数据不在选择的故障范围内，则将其视为正常标签数据。另外，由于故障能量相对大小数据非常接近且非常小，误差较大，因此将去除故障能量相对大小作为首要判别数据，取投票数最高的数据，将这一项可能作为最终结果，且逐步结合总故障票数和故障能量相对大小分别区分存在平票情况的数据结果，从而提高最终判别结果的准确性。

以无标签数据 D 为例，把整列全为零的数据删除，只留下有用的数据。所得 0~1 矩阵如下表 5 所示。

Table 5. 0~1 matrix of A
表 5. A 的 0~1 矩阵

部位	BA 故障能量相对大小	BA 能量集中的频率位置	峰度	偏度	平均值
B_DE	1	0	1	1	0
B_DE2	0	1	1	1	1
B_FE	0	0	0	1	0
IR_DE2	1	1	1	1	1
IR_FE	1	0	0	1	0
OR cen_DE	0	0	1	1	0
OR cen_DE2	0	0	1	1	1
OR cen_FE	0	0	0	1	1
OR op_DE	0	0	0	0	0
OR op_DE2	0	1	0	0	0
OR op_FE	0	0	1	0	0
OR or_FE	0	0	1	1	1
OR or_DE	0	0	0	0	0
OR or_DE2	0	0	0	1	0

由表 5 可以看出每个故障都跟部位有关，DE 与 DE2 是不同频率下的 DE 故障数据。表中 1 有一定占比，其杂乱无章没有规律，为了进一步提高分类准确性，由于故障能量相对大小各组数据差异较小，故误差较大，本文将去除故障能量相对大小的 0~1 矩阵作为首要判别数据。如表所示，B 和 IR 在 DE_2 位置每行存在 4 个 1 情况，则本数据为这两种故障之一，因此结合这两种故障总故障票数以及考虑故障能量相对大小票数进行判别，可以得出 IR 故障票数多，故该部位是 IR 故障。

根据以上说明的评价方法，利用 Python 结合 Excel 进行求解。最终得出位置标签数据所对应的不同故障分类，如下表 6 所示。

Table 6. Final analysis table
表 6. 最终分析表

类别	编号
正常	CEIJ
B	M
IR	ANHD
OR centered	GOBKLP
OR orthogonal	-
OR opposite	-

将上述表 6 与上文 KNN 分析和聚类分析出的结果相结合可以推测，B 故障与正常数据有些接近以至于混淆可能导致 KNN 分析比较粗糙。该方法与 KNN 分析和聚类分析相结合有一定可靠性，因此根据上述表格，进行数据可视化处理。

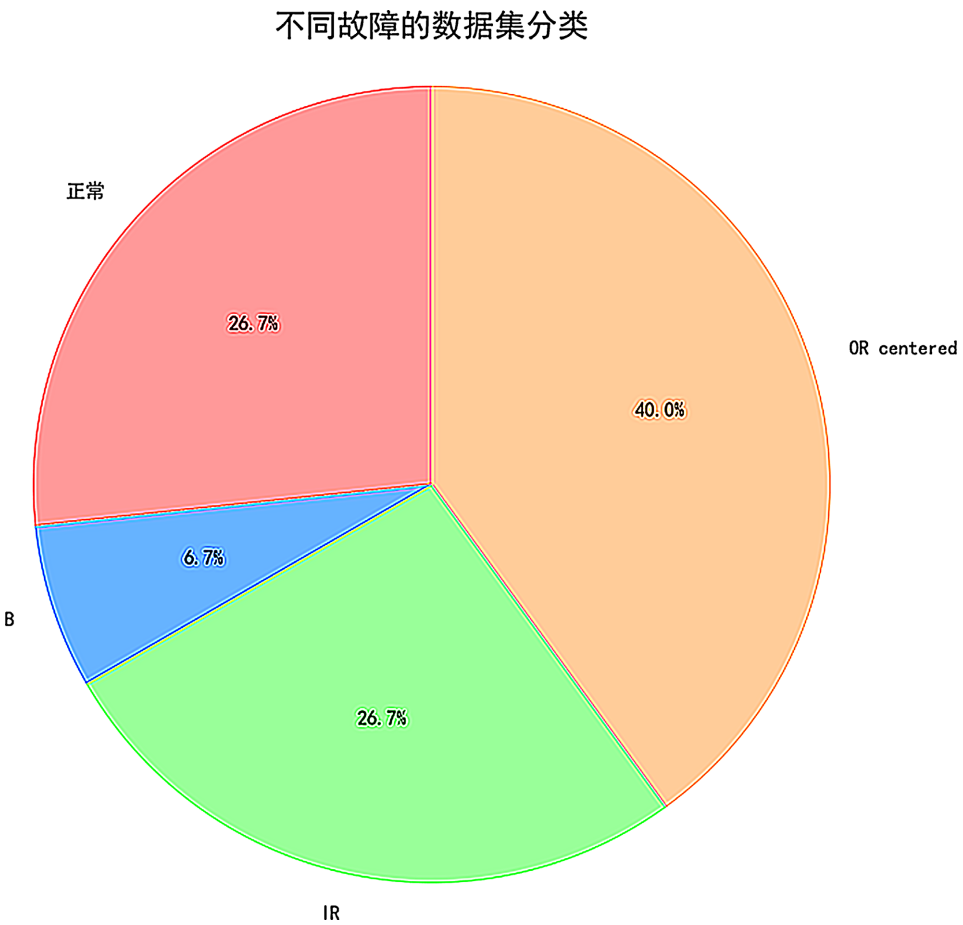

Figure 1. Classification of datasets with different faults
图 1. 不同故障的数据集分类

饼图能够直观地展示各部分在整体之中的占比，可以比较容易地观察和理解数据在样本中的分布情况。由图 1 可以看到，样本中 IR 和 OR 故障最多，正常数据也占有可观的一部分，B 故障样本中较少，说明 IR、OR 故障和正常数据确实很容易观测到。

4. 结论

本研究以轴承三个故障状态为研究对象(其中 OR 状态为三个不同故障部位)，基于傅里叶变换和奈奎斯特定理，利用最邻近算法和基于分层抽样和欧氏距离的中位数法进行有故障标签数据分类的模型建立，并将其迁移到无标签数据的分类运用。并加入了 k-means 算法进行验证，提出了一种多角度的算法判别分类结果。

数据结果表明，轴承故障在这些分类中都很常见，其中 IR 和 OR 故障最为常见，正常的数据也较多，体现了故障数据十分难收集。B 故障比起其余故障相对少一点。综上所述，这三种故障确实是轴承故障中值得分析的三种状态。

本研究给出了一种分析轴承无标签故障数据集分类的方法，对于高铁的故障判断提供一种可行方案，这有助于减轻传统人工判断故障压力，第一时间报告故障发生位置，为及时响应和故障处理提供助力。

5. 评价与改进

1) 对于初步分析后的统计数据，还需进一步验证是否有一定的可靠性。

2) 本文 0~1 矩阵的判别具有较强主观性，分类有较大误差，可以替换为更有说服力的数据分析。

3) 对于故障数据的分类，本文采用单一故障假设，且用的是实验室模拟数据，即实验数据与实际高铁运行数据这种真实情况下的数据有差异。

4) 本文十分依靠数据，且使用的能量计算方法较为简单，误差较大，可以在计算时添加窗函数减小噪声数据。由于中位数的选取十分简单，故对于数据结论影响较大，且不同频率之间的差异几乎忽略不计，可以采用更高级的机器学习相关算法。

参考文献

[1] 肖裕君, 高帆. 双高棒产线关键设备智能运维设计及应用[J]. 自动化仪表, 2024, 45(3): 103-107, 112.

[2] 中国学位与研究生教育学会, 中国科协青少年科技中心. “华为杯”第二十二届中国研究生数学建模竞赛参赛邀请函[EB/OL]. https://cpipc.acge.org.cn/cw/hp/4, 2025-09-21.

[3] 齐鹏宇, 郑近德, 潘海洋, 等. 波形自适应小波分解及其在滚动轴承故障诊断中的应用[J/OL]. 振动工程学报: 1-9. https://link.cnki.net/urlid/32.1349.tb.20250331.1017.004, 2025-09-25.

[4] 杨荣杰, 陈开超, 刘显著. 水电站机组振动监测及故障诊断系统设计与应用[J]. 电站辅机, 2025, 46(3): 78-81.

[5] 万珊, 苟文博. 计及神经网络信息检索算法的数字教学能力提升研究[J]. 自动化与仪器仪表, 2025(8): 199-202, 207.

[6] 肖宜滨. 聚类分析的理论及其应用[J]. 江苏统计, 2001(11): 11-13.

[7] 张慕宇, 杨智春, 丁燕, 等. 采用主成分分析与最邻近法的复合材料板损伤检测实验[J]. 西北工业大学学报, 2010, 28(5): 786-791.