计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190900174

一种基于马氏距离的系统故障诊断方法

林　毅[1]　吉鸿江[2]　韩佳佳[3]　张德平[3]
1 中国人民解放军 91776 部队　北京 100084
2 北京中船信息科技有限公司　北京 100861
3 南京航空航天大学计算机科学与技术学院　南京 210000
(linyi19820804@163.com)

摘　要　针对以往系统故障诊断方法中存在的多指标相关问题以及考虑多重积分时计算复杂、效率低等问题,文中基于马氏距离(Mahalanobis Distance,MD)度量提出一种系统故障诊断方法,利用采集到的系统状态监控数据,计算观测样本与已知样本之间的马氏距离,根据距离大小的 MD 面积度量比较判断观测样本类别,对已知数据样本的马氏距离的分布与观测数据样本的马氏距离的分布的差异进行故障诊断。具体地,首先利用 MD 方法将多变量数据转换为单变量数据,排除多变量之间相关性的干扰,避免了利用多重积分求解多变量联合分布的复杂性以及不确定性;然后利用面积度量法比较单变量数据的累积分布函数之间的差异,根据定积分计算分布曲线之间的面积值,以面积值较小对应的样本故障类别作为观测数据的类别。通过将所提方法与常用故障诊断方法(BP 神经网络、朴素贝叶斯)进行比较,证明了其简单有效,故障诊断正确率高,能够大大降低计算成本,并有效地提高故障诊断的效率。

关键词:故障诊断;马氏距离(MD);面积度量;累积分布函数
中图法分类号　TP311

System Fault Diagnosis Method Based on Mahalanobis Distance Metric

LIN Yi[1],JI Hong-jiang[2],HAN Jia-jia[3] and ZHANG De-ping[3]
1 Unit 91776 of the PLA,Beijing 100084,China
2 China Shipbuilding IT CO.,LTD.,Beijing 100861,China
3 College of Computer Science and Technology,Nanjing University of Aeronautics & Astronautics,Nanjing 210000,China

Abstract　In view of the multi-index related problems in previous fault diagnosis methods and the shortcomings of the calculated complication and low efficiency when considering multiple integrals,a system fault diagnosis method based on Mahalanobis Distance (MD) metrics is proposed to improve these problems. For the system performance state data monitored on a certain device, the proposed method is to calculate the MD area metric method to compare the distribution of the Mahalanobis Distances of the known data samples with the distribution of the Mahalanobis Distances of the observed data samples. Specifically,the MD method is firstly used to convert multivariate data into univariate data,and the correlation between multivariable is eliminated,and the complexity and uncertainty of multivariate joint distribution using multiple integrals are avoided. Then the area metric is used to compare the difference between the cumulative distribution functions of the univariate data,and the area value between the distribution curves is calculated according to the definite integral,and the smaller area value is the category of the sample fault. By comparing with common fault diagnosis methods (BP neural network and Naïve Bayes),it shows that the proposed method is simple and effective,the fault diagnosis rate is high,and the calculation cost is greatly reduced,and the system fault diagnosis efficiency is improved.

Keywords　Fault diagnosis,Mahalanobis distance (MD),Area metric,Cumulative distribution function

随着科学技术的发展,技术装备的结构越来越复杂,使用环境更加多变,系统装备比以前更加复杂,对可靠性的要求越来越高,系统出现故障的机率也越来越大,由此造成的损失也越来越严重,甚至会出现灾难性的后果。如今可靠性已经成为评价系统优劣的主要性能指标,系统设备的可靠性与安全性越来越受到人们的关注,故障诊断技术作为提高和改善系统可靠性的一种有效技术手段,在学术界与工业界引起了广泛关注,故障诊断与健康管理技术的研究已达到了前所未有的高度,故障诊断技术为提高系统可靠性和降低事故风险开辟了一条新的途径[1]。

常见的故障诊断方法是基于规则的专家系统[2-3]设备故障诊断,其从专家的经验中总结出规则,用来描述故障和状态之间的关系。其优点是直观形象,推理速度快,对数据存储空间的要求相对较小,但缺点是对历史故障经验的依赖性强,当知识库中没有与系统状态匹配的相应规则时,容易造成误诊或诊断失败。基于解析模型的诊断方法主要用于难以获取历

基金项目:总装预研项目资金(41402020501)
This work was supported by the General Assembly Pre Research Project Fund(41402020501).
通信作者:吉鸿江(jihongjiang_2008@126.com)

史经验的故障诊断中，通过对系统设备不同的分系统和部件等构建数学模型并求解来进行故障诊断。近年来，基于解析模型的非线性故障诊断方法受到国内外学者的广泛关注，文献[4]对非线性未知输入观测器和自适应非线性观测器方法进行了综述；文献[5]主要对 NUIO 和自适应学习方法进行了系统介绍，且侧重于诊断方法的鲁棒性。解析模型的优点是可以诊断未预知的故障，不需要历史经验知识。其缺点是模型较为复杂庞大、诊断速度慢，对模型精度的依赖性较强，并且模型的不确定性因素有可能导致错误的报警。故障树方法[6-7]则是一种体现故障传播关系的有向图，它以诊断对象最不希望发生的事件作为顶事件，按照对象的结构和功能关系逐层展开，直到不可分事件（底事件）为止。它的优点是可以较好地表达设备故障的层次关系及关联关系，赋予各个事件故障概率后也可以进行诊断策略优化的研究，诊断速度快，能保持一致性，并且应用领域广，只要给定相应的故障树就可以实现诊断。它的缺点是故障树一旦建造好就不容易更改，难以将与设备故障无关但可以用以诊断故障的相关信息纳入故障诊断过程之中；而且诊断结果严重依赖故障树信息的完整性，不能诊断不可预知的故障。而贝叶斯网络模型[8-10]通过实践积累可以随时进行学习，改进网络结构和参数，提高了故障诊断能力。贝叶斯网络具有很强的学习能力，在接收了新信息后会立即更新网络中的概率信息。在设备故障诊断应用中，贝叶斯网络将人机交互信息、与设备诊断相关的所有信息以适当的节点变量表示后，可以进行统一处理，即建造贝叶斯网络时变量的广义性可以将与设备故障诊断有关的所有信息来源纳入网络结构中，适合于表达更为复杂的、不确定性的问题。现有研究更倾向于把故障树和贝叶斯网络结合起来[11]。概率神经网络[12]（PNN）是在径向基函数神经网络的基础上发展而来的一种前馈型神经网络，近年来，PNN 以其良好的数据分类性能被广泛地应用于故障诊断领域，并在大量的实际工程中获得了成功应用[13-15]。目前对 PNN 的研究主要集中在算法学习理论及改进模型上，并获得了丰硕的成果[16-17]。然而，PNN 的隐层单元采用高斯函数作为激活函数，这就默认了训练数据的独立同分布假设。但实际上训练数据之间往往是相关的，这限制了 PNN 的应用。

模型验证度量作为一种数学运算符，可用于测量从仿真结果获得的模型预测与从实验得到的物理观测之间的差异[18]。在度量模型之间的差异时，当模型预测的真实联合累积分布函数（CDF）由经验联合 CDF 近似时，需要大量的随机样本来构建模型的多元联合 CDF 以及需要大量时间用于分类，计算成本较高；并且数据量少时，获得的联合 CDF 不能表示出模型的真实分布。Ferson 等提出了基于面积度量的方法[19]来测量预测分布和观测分布之间的整体差异，通过引入马氏距离（MD）[20-22]，将在特定位置收集的多变量实验观察结果和模拟的多个模型响应分别通过 MD 转换为单变量数据序列。MD 面积度量提供了模型的 MD 的 CDF 与实验观测的 MD 的经验 CDF 之间的比较。由于 MD 的单变量性质，所提出的度量方法适用于验证具有多个响应的模型。此外，MD 仅与模型的均值向量和协方差矩阵有关，而没有考虑模型的联合 CDF，这显着地降低了计算成本。

本文基于 MD 面积度量验证技术，提出了一种新的故障诊断方法。对于在系统设备监测的故障数据，采用 MD 面积度量方法对已知数据样本与观测数据样本的分布进行比较，首先利用 MD 方法将多变量数据转换为单变量数据，接着利用面积度量法比较单变量数据的累积分布函数之间的差异，面积值较小的则为设备故障的类别。此方法具有算法简单、计算量小、所需测点少、故障检测正确率高的特点，并且大大降低了计算成本。

1 相关技术

1.1 马氏距离

马氏距离（MD）是由印度统计学家马哈拉诺比斯（P. C. Mahalanobis）提出的，表示数据的协方差距离。它是一种有效计算一个样本和一个样本集“重心”的最近距离，或者有效计算两个未知样本集的相似度的方法。与欧氏距离不同的是，它考虑到各种特性之间的联系，可以排除变量之间的相关性的干扰，并且马氏距离是尺度无关的，即独立于测量尺度。当 $\boldsymbol{\Sigma}$ 是单位矩阵的时候，马氏距离即为欧氏距离。综上所述，马氏距离能够很方便地度量观测样本与已知样本集间的距离，因而很适合用于故障诊断。

假设 $\boldsymbol{Y}=(Y_1,\cdots,Y_m)$ 是 m 维随机向量，其均值向量为 $\boldsymbol{\mu}$，协方差矩阵为 $\boldsymbol{\Sigma}$，$F_Y=(y_1,\cdots,y_m)$ 是 $\boldsymbol{Y}$ 的联合累积分布函数（CDF）。$y=(y_1,\cdots,y_m)$ 是 Y 的随机样本。则从 $\boldsymbol{Y}$ 到均值向量 $\boldsymbol{\mu}$ 的马氏距离（MD）为：

$$\boldsymbol{R}(\boldsymbol{Y},\boldsymbol{\mu})=\sqrt{(\boldsymbol{Y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})} \tag{1}$$

从 y 到均值向量 $\boldsymbol{\mu}$ 的马氏距离（MD）为：

$$\boldsymbol{R}(y,\boldsymbol{\mu})=\sqrt{(y-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(y-\boldsymbol{\mu})} \tag{2}$$

其中：

$$\boldsymbol{\mu}=\frac{1}{m}\sum_{i=1}^{m}y_i \tag{3}$$

$$\boldsymbol{\Sigma}=\frac{1}{m}\sum_{i=1}^{m}(y_i-\mu)(y_i-\mu)' \tag{4}$$

由式（1）可以看出，$\boldsymbol{R}(\boldsymbol{Y},\boldsymbol{\mu})$ 是随机变量 $\boldsymbol{Y}$ 的函数，且 $\boldsymbol{R}(\boldsymbol{Y},\boldsymbol{\mu})$ 是一个标量随机变量。$\boldsymbol{R}(\boldsymbol{Y},\boldsymbol{\mu})$ 的 CDF 由 $F_R(r)$ 表示，$F_R(r)=P(\boldsymbol{R}\leqslant r)$。$F_R(r)$ 表示了关于 $\boldsymbol{Y}$ 的联合 CDF 的相关结构的信息。$F_R(r)$ 的精确数值是难以计算的，但是可以利用 Monte Carlo 模拟来近似。图 1 给出了二维随机向量的 CDF 的图示，图 2 给出了 $\boldsymbol{R}(\boldsymbol{Y},\boldsymbol{\mu})$ 的经验 CDF $F_R(r)$ 的图示。当收集的 $\boldsymbol{R}(\boldsymbol{Y},\boldsymbol{\mu})$ 样本充分多时，则可以平滑地近似 $\boldsymbol{R}(\boldsymbol{Y},\boldsymbol{\mu})$ 的 CDF，如图 2 所示。

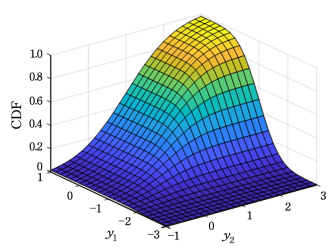


图 1 二维随机向量的 CDF

Fig. 1 CDF of two-dimensional random vector

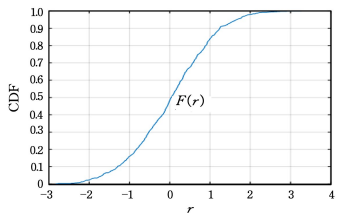


图 2 $\boldsymbol{R}(\boldsymbol{Y},\boldsymbol{\mu})$ 的经验 CDF

Fig. 2 Empirical CDF of $\boldsymbol{R}(\boldsymbol{Y},\boldsymbol{\mu})$

根据图 1 和图 2 可以看到，MD 距离度量实质是一种将多维空间映射到低维空间的方法，通过利用 MD 方法可以将随机向量 $\boldsymbol{Y}$ 变换成标量随机变量 $\boldsymbol{R}(\boldsymbol{Y},\boldsymbol{\mu})$，并且 $\boldsymbol{Y}$ 的主要统计特征可以由 $\boldsymbol{R}(\boldsymbol{Y},\boldsymbol{\mu})$ 表示。

1.2　面积度量

在数学上，MD 面积度量指标主要表示为不同曲线之间围成的面积。在某个验证点上的标量响应为 y，模型和物理实验的总体差异可以通过对模型响应的 CDF $F_Y^m(y)$ 与观测数据的经验 CDF $S_n^e(y)$ 之间的绝对值差进行积分来计算，如式(5)所示：

$$d(F_Y^m,S_n^e)=\int_{-\infty}^{+\infty}|F_Y^m(y)-S_n^e(y)|\,\mathrm{d}y \tag{5}$$

模型与实验结果的总体不一致可以通过面积差异来衡量，如图 3 中的阴影面积是模型响应的标准均匀分布和实验的经验 CDF 之间的差值，积分值为 0 说明完全匹配，为 0.5 说明差异达到最大。

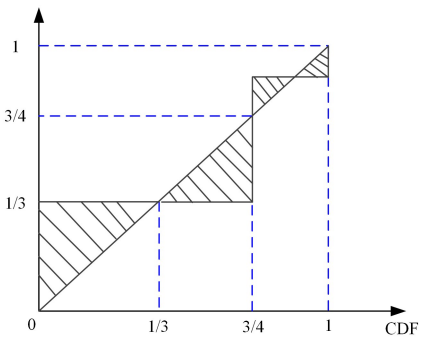


图 3　面积度量的比较

Fig. 3　Comparison of area measures

2　MD 面积度量诊断方法

本节通过扩展面积度量概念，提出一种新的验证度量标准来评估具有多个数据特征的模型的准确性。单监测点的 MD 面积度量用于将多个相关的数据转换为单变量数据序列。

MD 面积度量用于度量观测数据与已知数据之间的差异。上面提到的物理实验是多变量已知数据样本，表示为 $y_i^e(\boldsymbol{x})(i=1,\cdots,d)$，$\boldsymbol{x}$ 是监测点可控输入的向量，d 是变量的个数。观测数据由 $y_i^m(\boldsymbol{x})(i=1,\cdots,d)$ 表示。考虑到多个变量之间的不确定性和相关性，观测数据由具有均值向量 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$ 的联合 CDF $F_Y^m(y_1,\cdots,y_d)$ 表示。

MD 面积度量过程如图 4 所示，其包含如下 6 个步骤。

步骤 1　在流程图左侧，数据集 $\{y_i^e\}_{i=1}^k$ 是通过在特定监测点 $\boldsymbol{x}$ 处获得的历史数据样本，其中 $y_i^e=(y_1^e,\cdots,y_d^e)(i=1,\cdots,k)$ 是第 j 个数据集，k 是数据集的大小。

步骤 2　在流程图右侧，观测数据集 $\{y_h^m\}_{h=1}^n$ 是在相同监测点处获得的，其中 $y_h^m=(y_1^m,\cdots,y_d^m)_h(h=1,\cdots,n)$ 是第 h 个监测数据，n 是数据样本的大小。

步骤 3　对观测数据样本集进行归一化，将归一化后的数据样本通过式(6)转换为随机样本集。r_h^m 是从第 h 个监测数据 y_h^m 到 $\boldsymbol{\mu}$ 的随机 MD 样本，则基于 $\{r_h^m\}_{h=1}^n$ 来估计监测数据的 MD 的 CDF $F_R^m(r)$。

$$r_h^m(y_h^m,\boldsymbol{\mu})=\sqrt{(y_h^m-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(y_h^m-\boldsymbol{\mu})} \tag{6}$$

步骤 4　对历史数据样本集进行归一化，将归一化后的数据样本通过式(7)转换成一维数据序列。r_i^e 是从历史数据集 y_i^e 的第 j 个 MD 到 $\boldsymbol{\mu}$ 的 MD，则历史数据的 MD 的经验 CDF $S^e(r)$ 基于 $\{r_i^e\}_{i=1}^k$ 来估计。

$$r_i^e(y_i^e,\boldsymbol{\mu})=\sqrt{(y_i^e-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(y_i^e-\boldsymbol{\mu})} \tag{7}$$

步骤 5　根据度量算子 $d(F,S)=\int_0^{+\infty}|F_R^m(r)-S^e(r)|\,\mathrm{d}r$ 比较两个分布曲线 $F_R^m(r)$ 和 $S^e(r)$ 之间的面积差异。

步骤 6　故障分析诊断。

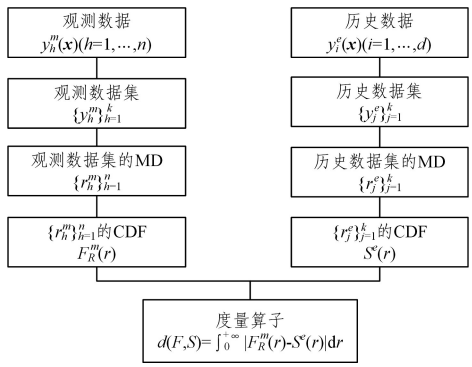


图 4　面积度量过程

Fig. 4　Process of area measurement

如果预测结果与历史数据分布相一致，则 $\{r_i^e\}_{i=1}^k$ 是来自监测数据的 MD 的样本，并且具有与 $F_R^m(r)$ 相同的分布。如果两者之间存在显著差异，且通过增加观测数据不能减少差异值，则观测数据与历史数据之间存在不一致。

3　实例分析

本节进行两组实验，实验采用的数据来源于预测与健康管理国际会议(PHM08)的预测挑战赛。数据集由多个多元时间序列组成。在本实验中，把每个数据集分为历史数据和观测数据。每个时间序列来自不同的发动机，即数据可以认为是由一组相同类型的发动机产生。每个发动机都以不同程度的初始磨损和变化开始运转，这是用户未知的。这种磨损和变化对发动机来说是正常情况，即非故障状态。发动机在每个时间序列开始时正常运行，并且在序列期间的某个点开始退化。在故障数据集中，退化的幅度不断增大，直到达到预定义的阈值，若超过阈值，则为故障状态。在正常数据集中，时间序列在完全退化前一段时间结束。数据的每行是单个操作周期内发动机记录的数字序列；每列是不同的变量，这些列对应于振动速率、温度、电压等。

3.1　评价指标

为检验实验模型的有效性，本文采用常用的评估分类器性能的评价指标进行评价，分别为真阳性、真阴性、假阳性、假阴性、正确率，根据表 3 给出的各评价指标的定义及计算，画出 ROC 曲线，求出 AUC。

表 1　分类器分类结果

Table 1　Result of classifier

		真实结果	
		正	负
预测结果	正	TP	FP
	负	FN	TN

(1)真阳性(True Positive)，也称召回率，对应真阳性率(TPR)，指的是按分类器分类标准正确判断正类为正类的样本数占该类实际样本总数的百分比，公式为：

$TP/(TP+FN)$

(2)假阳性(False Positive),对应假阳性率(FPR),指的是按分类器分类标准把不是正类的样本判为正类的样本数占非该类样本总数的百分比,公式为:

$FP/(TN+FP)$

(3)真阴性(True Negative),对应真阴性率(TNR),指的是按分类器分类标准将负类样本正确判为负类样本的样本数占非该类样本的总数的百分比,公式为:

$TN/(TN+FP)$

(4)假阴性(False Negative),对应假阴性率(FNR),指的是按分类器分类标准将正类样本误判断为反类的样本数占该类实际样本总数的百分比,公式为:

$FN/(TP+FN)$

(5)正确率(PRE),指的是预测为正例的样本中的真正正例的百分比,公式为:

$TP/(TP+FP)$

构建一个同时使正确率和召回率都很大的分类器是很难的,如果将所有样本都判断为正类,则召回率达到百分之百而正确率很低,所以在实际应用中应根据具体情况来调节两者的大小。

3.2 实例 1

在本实验中,所用数据集的指标变量是 23 个,每个周期对这 23 个指标进行测量,采集正常历史数据和故障历史数据各 1000 个作为训练集,采集正常观测数据和故障观测数据各 100 个作为测试集。首先对数据进行归一化处理,计算观测样本与历史样本之间的马氏距离,根据距离大小判断观测样本类别。

记故障测试样本到故障训练样本的 MD 为 MD1,故障测试样本到正常训练样本的 MD 为 MD2,正常测试样本到故障训练样本的 MD 为 MD3,正常测试样本到正常训练样本的 MD 为 MD4,计算结果如表 2 所列。

表 2 测试样本到训练样本的 MD

Table 2 MD from test sample to training sample

序号	MD1	MD2	MD3	MD4
1	1.5164	1.5276	1.1294	0.6741
2	1.4849	1.4994	1.1197	0.6820
3	1.5164	1.5276	1.1197	0.6820
4	1.5163	1.5275	0.5255	0.5011
5	1.4601	1.4971	4.1946	1.5818
6	1.5215	1.5311	1.7090	0.8645
7	1.4790	1.5068	4.1875	1.5769
8	1.4849	1.4993	1.7103	0.8645
9	1.4807	1.5091	1.7114	0.8661
…	…	…	…	…
100	1.4807	1.5091	1.1212	0.6817

由于篇幅有限,本实验在表中仅列出了 10 组数据,但可从表 2 中看出,MD1 均小于 MD2,MD4 均小于 MD3,即正常测试样本均判断为正常,故障测试样本均判断为故障,这表明根据马氏距离很好地完成了故障与否的识别,正确率达到了 100%,对于小样本数据,本方法达到了理想的效果,但若数据量过大,则计算成本较高,不适合实际应用。

3.3 实例 2

本实验采用的训练数据样本与实例 1 相同,测试样本扩大了 10 倍,克服了实例 1 的缺点。首先利用 MD 方法将多变量数据转换为单变量数据,避免了求解多变量联合分布的复杂性以及不确定性。再利用面积度量法比较单变量数据的累积分布函数之间的差异,面积值较小的则为设备故障的类别。通过计算历史样本之间及观测样本之间的马氏距离,并求解所有马氏距离的累积分布函数,根据面积度量公式求出分布函数之间的面积差,进而判断故障类别。

记故障训练样本之间的 MD 为 MD5,故障测试样本之间的 MD 为 MD6,正常训练样本之间的 MD 为 MD7,正常测试样本之间的距离为 MD8。根据 MD 分别求出其累计分布函数(CDF),如图 5—图 8 所示。

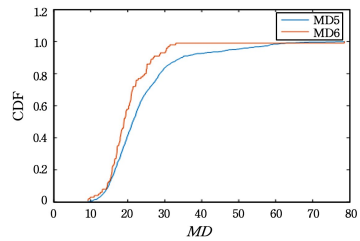


图 5 故障训练样本与故障测试样本的 MD 的 CDF

Fig. 5 CDF of MD of fault training samples and fault test samples

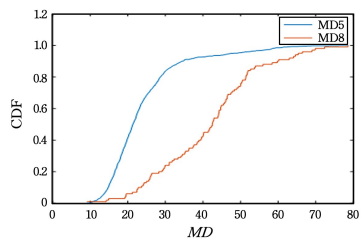


图 6 故障训练样本与正常测试样本的 MD 的 CDF

Fig. 6 CDF of MD of fault training samples and normal test samples

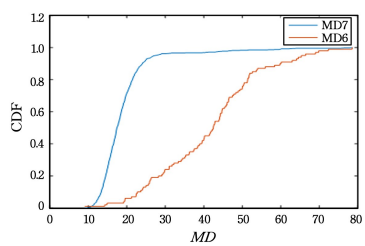


图 7 正常训练样本与故障测试样本的 MD 的 CDF

Fig. 7 CDF of MD of normal training samples and fault test samples

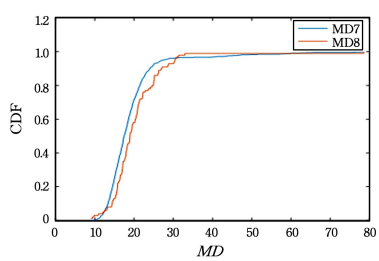


图 8 正常训练样本与正常测试样本的 MD 的 CDF

Fig. 8 CDF of MD of normal training samples and normal test samples

从图 5—图 8 可以直观地看出,故障测试样本与故障训练样本分布之间的面积差异明显小于故障测试样本与正常训练样本分布之间的面积,正常样本也是如此,为具体说明,给出计算结果如表 3 所列。

表 3　MD 面积度量结果

Table 3　Result of MD area measurement

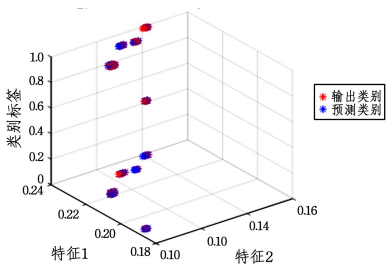
样本	MD5	MD7
MD6	0.1071	0.4364
MD8	0.3572	0.1037

从表 3 可以看出，正常测试样本与正常训练样本的面积差(0.1037)远小于其与故障训练样本的面积差(0.4364)，故障测试样本与故障训练样本的面积差(0.1071)远小于其与正常训练样本的面积差(0.3572)。这说明根据 MD 的累积分布函数可以有效地识别出设备地故障与否，准确率较高，且避免了求解多变量联合分布的复杂性以及不确定性，大大降低了计算复杂度。

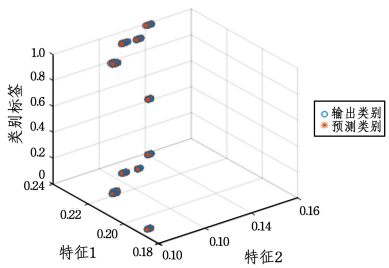
3.4　实例 3

本实例采用 BP 神经网络作为实例 1 的对比实验，所用数据与实例 1 相同，并把故障类型设置为 0，正常类型设置为 1。对已知训练样本集进行归一化，训练构建的 BP 神经网络以调整网络的权值和阈值，最终得到期望的故障结果输出。神经网络设置为输入层、隐含层、输出层 3 层，节点数分别为 24，10，1，隐含层、输出层函数均设为 logsig 函数，隐含层节点数 10 是经过多次训练比较得出的最优值，最大训练次数为 1 000，误差精度为 0.000 04，学习率为 0.02。训练完毕后，利用训练好的网络模型进行故障诊断。输入测试数据样本，输出结果大于 0.5 时判断为 1 类，即正常；结果小于 0.5 时为 0 类，即故障。

为了更直观地展示故障类别预测结果，本文采用 PCA 方法对特征指标进行了降维，选取第一主成分和第二主成分分别作为 x 轴和 y 轴，类别标签为 z 轴，BP 神经网络与实例 1 的实验结果展示如图 9 所示。



(a)BP 神经网络预测类别与实际类别比对



(b)MD 预测类别与实际类别比对

图 9　BP 神经网络与 MD 方法诊断结果

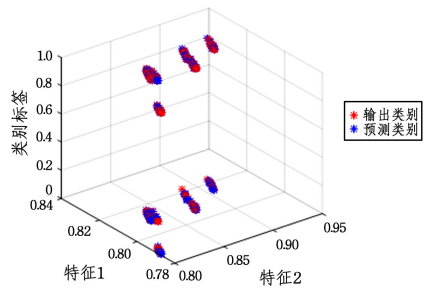
Fig. 9　Diagnosis results of BP neural network and MD method

从图 9 可以看出，BP 神经网络预测类别与实际输出类别大部分是相同的，但还是能明显地看出差异，而 MD 方法预测类别与实际输出类别是重合的，说明 MD 方法的预测效果更好。在实验中采用 100 个故障样本，100 个正常样本，BP 神经网络预测正确的故障样本有 93 个，正常样本有 82 个，正确率为 0.875，说明 BP 神经网络对故障分类也有较好的识别能力，但不如 MD 方法。

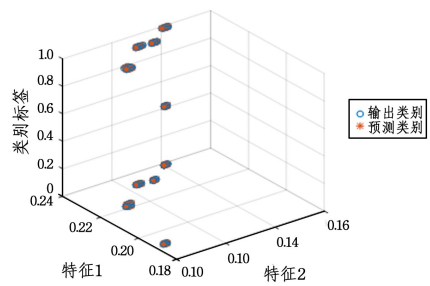
3.5　实例 4

本实例采用朴素贝叶斯方法作为 MD 方法的对比实验，所用数据与实例 1 相同，并把故障类型设置为 0，正常类型设置为 1。贝叶斯分类器的分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率估计，并记录结果。其输入是特征属性和训练样本，输出是分类器。对已知训练样本集进行归一化，训练构建的贝叶斯分类器，最终得到期望的故障结果输出。

为了更直观地展示故障类别预测结果，本文采用 PCA 方法对特征指标进行了降维，选取第一主成分和第二主成分分别作为 x 轴和 y 轴，类别标签为 z 轴，BP 神经网络与实例 1 的实验结果展示如图 10 所示。



(a)朴素贝叶斯预测类别与实际类别比对



(b)MD 预测类别与实际类别比对

图 10　朴素贝叶斯与 MD 方法诊断结果

Fig. 10　Diagnosis results of Naive Bayes and MD method

从图 10 可以看出，朴素贝叶斯方法预测类别与实际输出类别大部分是相同的，但还是能明显地看出差异，而 MD 方法预测类别与实际输出类别是重合的，说明 MD 方法的预测效果更好。实验采用 100 个故障样本，100 个正常样本，预测正确的故障样本有 86 个，正常样本有 95 个，正确率为 0.905，说明朴素贝叶斯方法对故障分类也有较好的识别能力，但不如 MD 方法。

3.6　假设检验分析

因为上述实验所用数据较少，所得结论并不能完全让人信服，所以需要进行假设检验。假设检验又称统计假设检验，是一种基本的统计推断形式，用来判断样本与样本、样本与总体的差异是由抽样误差引起还是本质差别造成的统计推断方法。其基本原理是先对总体的特征作出某种假设，再通过抽样研究的统计推理，对此假设应该被拒绝还是接受作出推断。

假设检验过程为：

(1)提出原假设 H。样本与总体是一致的;备择假设 H_1:样本与总体间存在本质差异。

(2)选用 Z 检验,计算统计量 Z 值的公式为:

$$Z=\frac{\overline{X}-\mu}{s/\sqrt{n}} \tag{8}$$

其中,$\overline{X}$是检验样本的平均数;μ 是已知总体的平均数;S 是样本的标准差;n 是样本容量。

(3)计算 p 值,根据统计量的大小及其分布确定检验假设成立的可能性 p 的大小并判断结果。若 $p>\alpha$ 则,接受 H_0;若 $p\leqslant\alpha$,则接受 H_1。

设定显著性水平 α 为 0.05,即当检验假设为真,但被错误拒绝的概率。表 4 给出了 MD 方法、BP 神经网络以及朴素贝叶斯方法进行 Z 检验的结果,从表 4 可以看出 3 种方法的检验 p 值均大于 0.05,则均接受原假设,说明样本与总体是一致的,实验结果具有可信度。

表 4 各方法假设检验结果

Table 4 Hypothesis testing results of each method

样本	Z 值	p 值
MD	0.6428	0.4364
BP	1.2642	0.1647
Bayes	1.5742	0.2563

3.7 实验结果分析

本文从精确度和时间复杂度两个方面对结果进行分析。精确度通过 ROC 曲线下的面积来表示,时间复杂度通过计算不同样本量所花费的时间来刻画。

ROC 分析工具是一个画在二维平面上的曲线——ROC curve。平面的横坐标是 *FPR*,纵坐标是 *TPR*。对分类器而言,我们可以根据其在测试样本上的表现得到一个 *TPR* 和 *FPR* 点对。这样,此分类器就可以映射成 ROC 平面上的一个点。虽然,用 ROC curve 来表示分类器的性能很直观,但是人们总希望能有一个数值来标志分类器的好坏。于是 Area Under roc Curve(AUC)被提出。顾名思义,*AUC* 的值就是处于 ROC curve 下方的面积的大小。通常,*AUC* 的值介于 0.5 到 1.0 之间,较大的 *AUC* 代表了较好的性能。

本文将上述实验方法的测试结果分成 10 个个数相同的部分,分别计算每部分的 *FPR* 及 *TPR*,映射成 ROC 平面上的点,并绘制 ROC 曲线,如图 11 所示。

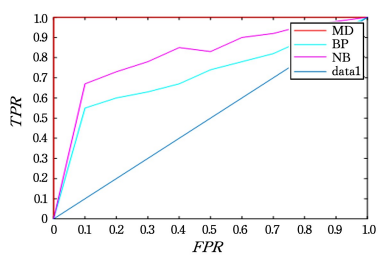


图 11 ROC 曲线

Fig. 11 ROC curve

从图 11 可以直观地看出 MD 方法优于 BP 神经网络优于朴素贝叶斯,且 MD 过左上角的点(TPR=1,FPR=0),为完美分类,说明 MD 方法对于故障诊断的分类具有很高的可信度。为了用数值进行有力的说明,利用积分计算 *AUC* 值,如表 5 所列。

表 5 实验性能对比

Table 5 Comparison of experimental performance

	MD	BP	NB
TPR	1	0.82	0.97
FPR	0	0.07	0.14
TNR	1	0.93	0.86
FNR	0	0.18	0.03
PRE	1	0.92	0.874
AUC	1	0.725	0.793

本文通过上述评价指标的计算结果来比较所提方法与 BP 神经网络及朴素贝叶斯方法的性能。从表 5 可以看出,基于 MD 的方法在每个评价指标下都优于 BP 神经网络和朴素贝叶斯分类。MD 的 *TPR* 是 1,*FPR* 是 0,*AUC* 是 1,分类完全正确。朴素贝叶斯的 *TPR* 是 0.97,比 BP(0.82)的要高,而 *FPR* 是 0.14,比 BP(0.07)的低,这说明朴素贝叶斯比 BP 把更多的样本判断为了正常。朴素贝叶斯的 *AUC* 为 0.793,高于 BP(0.725),说明贝叶斯的整体性能要优于 BP。本实验所用的测试集仅有 200 个,数据量较小,MD 方法的准确率达到 100%,对于大样本数据,采用 MD 面积度量方法进行判断,故障检测的正确率仍然优于 BP 神经网络和朴素贝叶斯方法,且 MD 方法计算量小,所需测点也较少,通过运行时间对比图可以得到 MD 方法降低了计算成本。

时间复杂度可以通过算法运行的具体时间以及时间变化的趋势两个方面来表示。本文通过选取不同的样本量进行测试,通过表格及图形展示了各样本的具体运行时间以及时间变化的趋势,样本量分别为[10,50,100,500,600,800,1000],运行时间对比表及对比图如表 6、图 12 所示。

表 6 运行时间对比表

Table 6 Comparison of running time

	MD	BP	NB
10	0.001	0.001	0.001
50	0.001	0.001	0.001
100	0.001	0.001	0.001
500	0.001	0.005	0.006
600	0.002	0.010	0.020
800	0.003	0.015	0.030
1000	0.0035	0.025	0.050

从表 6 可以看出,在数据量较小时,BP 和 NB 以及 MD 的运行时间基本相同。而数据量超过 500 时,MD 方法的运行时间缓慢增长,BP 神经网络和朴素贝叶斯的运行时间增长很快,且是 MD 方法的 10 倍左右,说明在数据量大的情况下 MD 方法很好地降低了计算开销。

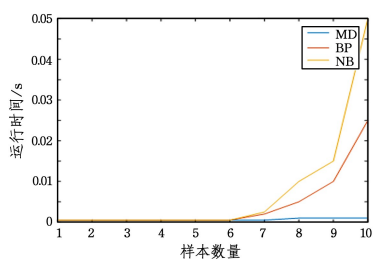


图 12 运行时间对比图

Fig. 12 Diagram of running time comparison

从图 12 可以看出,MD 方法的运行时间一直较低,且变化趋势很平缓,速度基本不变,BP 和 NB 在数据量较小时运行速度与 MD 基本相同,而数据量超过 5000 时,BP 神经网络和朴素贝叶斯运行速度变慢,时间增长速度快速增加,说明

MD方更适用于大样本数据集，其能很好地解决时间复杂度问题，大大降低了计算成本。

结束语　针对以往故障诊断方法中存在的多指标相关问题以及考虑多重积分时计算复杂、效率低等缺点，本文提出两种基于马氏距离(MD)度量的故障诊断方法。采用正常历史数据和故障历史数据各1 000个作为训练集，正常观测数据和故障观测数据各100个作为测试集，并对数据进行归一化处理。计算观测样本与已知样本之间的马氏距离，根据距离大小判断观测样本类别，此方法直观有效，适用于小样本数据。第二个实验采用MD面积度量方法，首先利用MD方法将多变量数据转换为单变量数据，避免了求解多变量联合分布的复杂性以及不确定性，再求出单变量数据样本的马氏距离的分布，利用面积度量法比较单变量数据的累积分布函数之间的差异，根据积分计算曲线之间的面积值，面积值较小的则为设备故障的类别。

本文在真实数据集的基础上，通过与BP神经网络及朴素贝叶斯故障诊断方法进行比较，从ROC曲线图以及TPR，FPR，TNR，FNR，PRE，AUC 6个性能评价指标上得出，使用MD方法进行故障诊断是完全可行的，且方法简单有效，准确率很高；从时间复杂度的对比结果可以得出，本文方法大大降低了计算成本，提高了计算效率。

参 考 文 献

[1] CHEN Y. Theories and methods of automobile engine fault diagnosis[J]. Industry Press, 2016(6): 56-57.

[2] MENG X P, LI J L, ZHANG Y W. Fault Diagnosis of Building Automation System Based on Expert System[J]. Computer Engineering, 2011, 37(21): 273-275.

[3] ZHAO L, LU Z, YUN W, et al. Validation metric based on Mahalanobis distance for models with multiple correlated responses [J]. Reliability Engineering & System Safety, 2017, 159: 80-89.

[4] SIDHU A, IZADIAN A, ANWAR S. Adaptive Nonlinear Model-Based Fault Diagnosis of Li-Ion Batteries[J]. Industrial Electronics IEEE Transactions on, 2014, 62(2): 1002-1011.

[5] KARGAR S M, SALAHSHOOR K, YAZDANPANAH M J. Integrated nonlinear model predictive fault tolerant control and multiple model based fault detection and diagnosis[J]. Chemical Engineering Research & Design Transactions of the Inst, 2014, 92(2): 340-349.

[6] ZHENG Q, WANG Y, UNIVERSITY Q N. Fault Diagnosis of Generator Sets Based on Fault Tree Analysis[J]. Marine Electric & Electronic Engineering, 2017.

[7] REN Y, YAXIONG B I, WANG D, et al. Fault tree intelligent diagnosis technology for wind turbine drivetrain[J]. Journal of Drainage & Irrigation Machinery Engineering, 2016.

[8] ENRIQUE SUCAR L, BIELZA C, MORALES E F, et al. Multilabel classification with Bayesian network-based chain classifiers [J]. Pattern Recognition Letters, 2014, 41(1): 14-22.

[9] HE S, WANG Z, WANG Z, et al. Fault Detection and Diagnosis of Chiller Using Bayesian Network Classifier with Probabilistic Boundary[J]. Applied Thermal Engineering, 2016, 107: 37-47.

[10] CAI B, HUANG L, XIE M. Bayesian Networks in Fault Diagnosis[J]. IEEE Transactions on Industrial Informatics, 2017, PP(99): 1-1.

[11] XUE S S, LI X C, XU X Y. Fault Tree and Bayesian Network Based Scraper Conveyer Fault Diagnosis[M] // Proceedings of the 22nd International Conference on Industrial Engineering and Engineering Management 2015. Atlantis Press, 2016.

[12] KOWALSKI P A, KULCZYCKI P. Interval probabilistic neural network[J]. Neural Computing & Applications, 2017, 28(4): 1-18.

[13] 王子健. 基于概率神经网络的发动机失火故障诊断[D]. 长春：吉林大学，2016.

[14] XU S, LIU D, LIU B. Application of fuzzy algorithm-based multiple cmac neural networks in coagulant dosing system [J]. Computer Applications & Software, 2016(3): 23-28.

[15] WANG D, ZHAO X J. A simple and fast guideline for generating enhanced/squared envelope spectra from spectral coherence for bearing fault diagnosis[J]. Mechanical Systems and Signal Processing, 2019, 122: 754-768.

[16] ZONG M, MENG H, GU W, et al. Rolling Bearing Fault Diagnosis Method Based on LMD Multi-scale Entropy and Probabilistic Neural Network[J]. China Mechanical Engineering, 2016.

[17] FERNÁN D M, CISNEROSRUIZ A J, CALLEJÓNGIL Á. Applying a probabilistic neural network to hotel bankruptcy prediction[J]. Tourism & Management Studies, 2016, 12(1): 40-52.

[18] LIU D, ZENG H, XIAO Z, et al. Fault diagnosis of rotor using EMD thresholding-based de-noising combined with probabilistic neural network[J]. Journal of Vibroengineering, 2017, 19(8).

[19] ZHAO L, LU Z, YUN W, et al. Validation metric based on Mahalanobis distance for models with multiple correlated responses [J]. Reliability Engineering & System Safety, 2017, 159: 80-89.

[20] LEI Y, LI N. Machinery health prognostics: a systematic review from data acquisition to RUL prediction[J]. Mech. Syst. Signal Process, 2018, 104: 799-834.

[21] MELLIT A, TINA G M, KALOGIROU S A. Fault detection and diagnosis methods for photovoltaic systems: A review[J]. Renewable and Sustainable Energy Reviews, 2018, 91: 1-17.

[22] COWLING B J, HEDLEY A J. The Mahalanobis Distance[M]. BMJ, 2017.

[23] MEI J, LIU M, WANG Y F, et al. Learning a Mahalanobis Distance-Based Dynamic Time Warping Measure for Multivariate Time Series Classification[J]. IEEE Transactions on Cybernetics, 2016, 46(6): 1363-1374.

LIN Yi, born in 1982, master, assistant research fellow. His main research interests include military big data, military modeling and simulation, and evaluating effectiveness of military system.

JI Hong-jiang, born in 1987, master, engineer. His main research interests include military and industrial big data and artificial intelligence, information integration and modeling.