

文章编号:1000-8055(2013)10-2181-13

基于支持向量机的航空发动机转静碰摩部位 诊断规则提取

李 爱, 陈 果, 于明月

(南京航空航天大学 民航学院, 南京 210016)

摘 要: 引入基于支持向量机(SVM)的数据挖掘技术,提出了基于 SVM 的转静碰摩部位诊断知识获取. 首先,基于带机匣的航空发动机转子实验器,模拟了 4 个碰摩部位的碰摩实验,利用机匣 4 个部位的应变测试,获取了 4 个碰摩部位和 4 个测点的大量实验数据;然后提出了一种基于支持向量聚类(SVC)的诊断知识规则提取方法. 在该方法中,利用 SVC 算法得到特征选取后样本的聚类分配矩阵,最后根据聚类分配矩阵构建超矩形规则. 为使规则更加简洁,易于解释,采用规则合并、维数约简、区间延伸等方法对超矩形规则进行进一步简化. 利用基于 SVM 的数据挖掘方法,从大量的碰摩部位实验数据中提取出了转静碰摩部位诊断的知识规则,并进行了相应解释和验证,规则识别率达到了 99%以上,表明了该方法的正确有效性.

关 键 词: 转静碰摩; 支持向量机; 故障诊断; 规则提取; 知识获取

中图分类号: V263. 6

文献标志码: A

Aero-engine rotor-stator rubbing positions diagnosis rule acquisition based on support vector machine

LI Ai, CHEN Guo, YU Ming-yue

(College of Civil Aviation,

Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: The data mining technology based on support vector machine (SVM) was introduced, and aero-engine rotor-stator rubbing positions diagnosis rule acquisition was proposed based on SVM. Firstly, the rubbing experiment of 4 rubbing parts was simulated based on aero-engine rotor tester with the casing, and a large number of experimental data were obtained by using the strain test of 4 parts of the casing. A new approach was proposed to extract knowledge rules from support vector clustering (SVC). Then SVC algorithm was adopted to get the clustering distribution matrix of the sample data with chosen features. Secondly, hyper-rectangle rules were constructed on the basis of the clustering distribution matrix. In order to make the rules more concise and explainable, hyper-rectangle rules should further simplified by using such means as rules merger, dimension reduction and interval extension. Finally, using the data mining method based on SVM, aero-engine rotor-stator rubbing positions diagnosis rules were extracted from a large number of rubbing position experiment data, then explanation and validated accordingly. The recognition rates

收稿日期:2012-09-21

网络出版地址:

基金项目:基础研究项目(613139); 国家自然科学基金(61179057)

作者简介:李爱(1987—),女,山东章丘人,博士生,主要研究方向为专家系统与故障诊断.

were more than 99%, showing that the method is corrective and effective, and embodies great practical values.

Key words: rubbing rotor-stator; support vector machine; fault diagnosis; rules extraction; knowledge acquisition

航空发动机转静间的碰摩^[1]是一种严重故障,会导致转静间隙增大、轴承磨损、叶片折断,直至机械失效.有效识别碰摩部位,避免由于碰摩故障而造成的严重事故,不但具有很大的经济效益,同时对于发现碰摩原因和改进设计,提高航空发动机的工作安全性,防止重大事故的发生,有着重要的意义.目前对碰摩部位的识别方法主要有:①基于声发射技术的方法^[2-3];②基于有限元模型的方法^[4];③动刚度法^[5].但这些方法存在两方面问题:首先是方法所采用的测试手段在航空发动机中不常用,因而难于应用于工程实际;其次是未提取和总结出碰摩部位识别的规律,同样也导致方法难于推广和使用.由此可见,总结出碰摩部位识别规律,提取和挖掘出易于理解和解释的诊断规则是碰摩部位识别方法应用于工程实际的重要途径.

目前,基于数据挖掘^[6-7]的知识获取主要是通过机器学习或数理统计方面的一些算法从已有的一些数据中获取知识.其中关联分析法^[8]、人工神经网络^[9]、粗糙集^[10]和决策树^[11]等在数据挖掘中的应用很广泛.近年来,SVM^[12]作为数据挖掘中的一种新兴的分类技术,良好的泛化性能和非线性映射能力使 SVM 和神经网络一样,具有逼近任意连续有界非线性函数的能力,并且还具有神经网络所不具有的许多优点,如泛化能力强,学习问题不存在局部极小,可以自动确定学习机的结构,不存在维数灾难问题,以及处理小样本能力强等.因此,基于 SVM 的数据挖掘技术已受到数据挖掘界的重视,对它的研究不断深入.迄今发表的 SVM 规则提取算法^[13-15]不仅简单而且具有广泛的适用性.

鉴于此,本文提出一种基于 SVC 的知识规则提取方法,在该方法中,为了提高知识规则的解释性.采用 SVC 算法得到样本的聚类分配矩阵,最后在聚类分配矩阵的基础上构建超矩形,当这些超矩形映射到坐标轴上后,就得到超矩形规则.最后利用规则合并和约简对超矩形规则进行进一步简化,使其更简洁,易于解释和理解.

本文将所提出的基于 SVM 的数据挖掘方法应用于航空发动机碰摩部位识别和诊断规则获取研究.首先对带机匣的航空发动机转子实验器进

行不同碰摩部位实验,由于现代航空发动机普遍采用薄壁结构,碰摩易于导致机匣变形,因此,本文采用广泛用于航空发动机的应变测试技术,直接测取航空发动机机匣的应变信号,并以此提取碰摩部位识别特征.本文模拟了 4 个碰摩部位的碰摩实验,利用机匣 4 个部位的应变测试,获取了 4 个碰摩部位和 4 个测点的大量实验数据;最后,利用本文基于 SVM 的数据挖掘方法,从大量的碰摩部位实验数据中提取出了转静碰摩部位诊断的知识规则,并进行了相应解释和验证.

1 不同碰摩部位的碰摩实验

由于现代航空发动机普遍采用薄壁结构,在碰摩作用下产生一定的变形.鉴于此,本文提出基于机匣应变信号的航空发动机碰摩部位识别技术.从转子实验器机匣获取应变信号,提取应变信号的平均值.

本文利用沈阳发动机设计研究所设计研制的航空发动机转子实验器进行碰摩实验,该实验器可以通过调节碰摩螺钉挤压碰摩环实现不同程度的碰摩.实验器如图 1 所示.

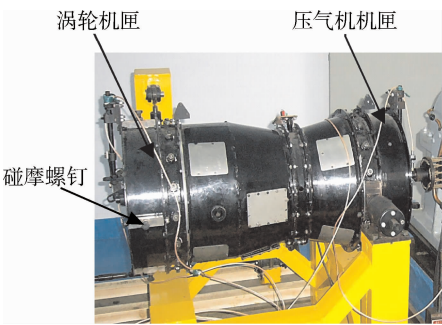


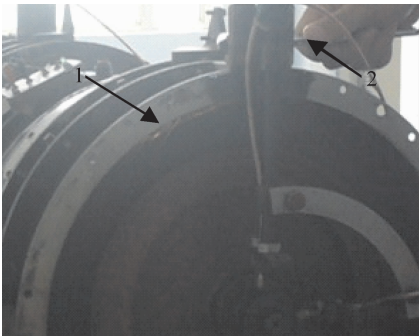
图 1 航空发动机转子实验器

Fig.1 Experiment rig of aero-engine rotor

在涡轮机匣处设计了 4 个碰摩螺钉,实现 4 个部位的碰摩实验.碰摩位置分别为涡轮机匣垂直上、水平右、垂直下、水平左 4 个方向.碰摩实验如图 2 所示.实验时用扳手拧碰摩点位置的碰摩螺钉,使碰摩环产生变形,从而与旋转的涡轮叶片产生碰摩,当碰摩严重时,将产生碰摩火花.

考虑转静碰摩过程中切向摩擦力引起的机匣

变形,沿机匣周向粘贴应变片,其示意图如图 3 所示.图 4 表示面向涡轮机匣,沿顺时针方向应变片的粘贴位置及碰摩位置示意图.本文实验数据来自 3 次独立实验,如表 1 所示.



1 碰摩火花;2 拧碰摩螺钉产生碰摩

图 2 碰摩实验

Fig. 2 Rubbing experiment

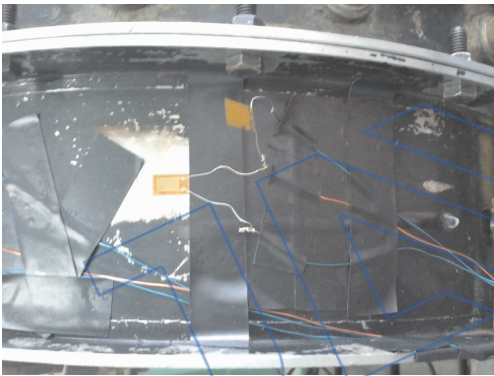


图 3 应变片周向粘贴示意图

Fig. 3 Strain gauges pasted along casing circumferential

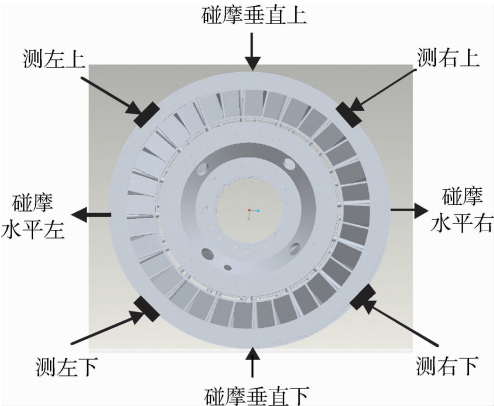


图 4 应变片位置及碰摩位置

Fig. 4 Position of strain gauges and rubbing

表 1 碰摩实验数据
Table 1 Rubbing experimental data

编号	实验日期	转速/(r/min)	样本数
1	2012-5-15	1500	395
2	2012-5-16	1500	395
3	2012-5-17	1500	395

2 基于应变信号的特征提取

2.1 应变均值特征提取

当转子部件发生碰摩的位置不同时,机匣同一位置的变形不同,而碰摩位置相同时,机匣同一位置的变形又有一定的一致性,因此本文提取粘贴于转子实验器 4 个部位应变片应变信号的均值作为特征参数输入至 SVM 中,其具体实现过程如下:采集转子实验器正常运转及碰摩位置为涡轮机匣的垂直上、垂直下、水平左、水平右,实验转速为 1500 r/min 时机匣 4 个位置所粘贴的应变片的应变值;提取应变信号的均值特征;将转子实验器碰摩状态下的应变均值与正常状态的应变均值对应相减.

2.2 不同碰摩部位的应变特征分析

由实验结果提取的样本特征值如图 5~图 8 所示.图 5~图 8 分别表示碰摩涡轮机匣垂直上、水平右、垂直下、水平左时 4 个位置应变片应变信号的均值特征.图 5~图 8 中的图(a)表示 2012 年 5 月 15 日实验时涡轮机匣各位置应变信号的均值,图 5~图 8 中的图(b)表示 2012 年 5 月 16 日实验时涡轮机匣各位置应变信号的均值.

从图 5~图 8 可以很明显地看出采用沿机匣周向粘贴应变片的实验方案时,当实验器碰摩位置不同时,机匣同一位置的应变信号具有一定的差异性.以该方案的第一次实验为例,碰摩位置为涡轮机匣垂直上时,机匣左上方粘贴的应变片拉伸变形较大(如图 5(a)的“+”),而碰摩涡轮机匣垂直下时,粘贴于机匣左上方的应变片压缩变形最大(如图 7(a)的“+”).当碰摩位置相同时同一位置的应变信号也表现出较好的一致性,例如同为碰摩涡轮机匣垂直下时,均为靠近涡轮左下方应变片受拉伸变形最大(如图 7(a)“○”及图 7(b)“○”),而且该方案下各次实验的应变值较接近,这为碰摩部位识别提供了良好的基础.

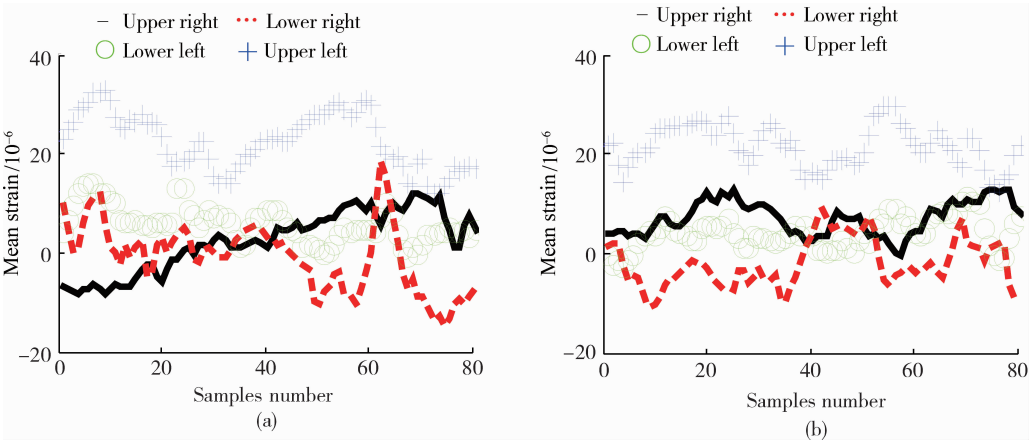


图 5 碰摩涡轮机匣垂直上应变信号的平均值

Fig. 5 Mean strain of the vertical upper position on the rubbing turbine casing

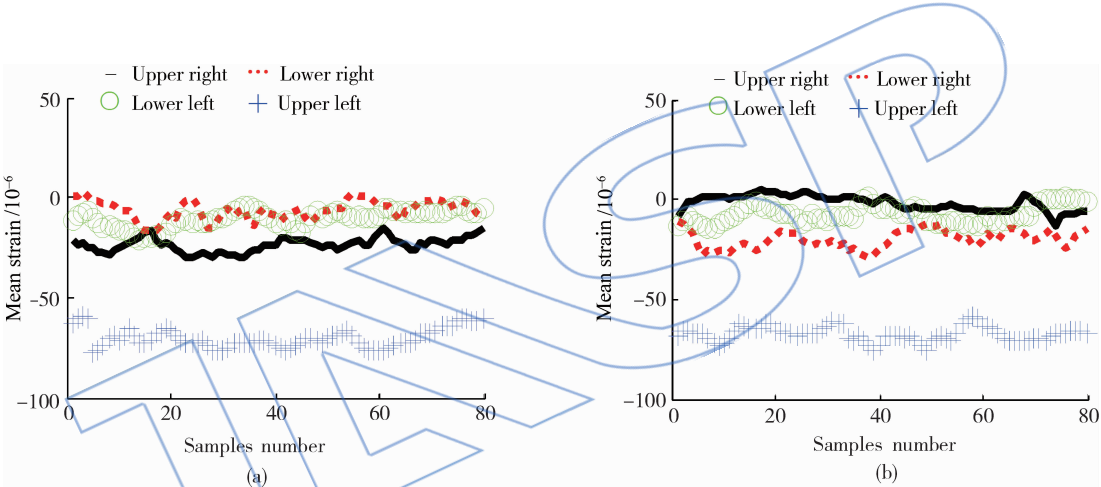


图 6 碰摩涡轮机匣水平右应变信号的平均值

Fig. 6 Mean strain of the horizontal right position on the rubbing turbine casing

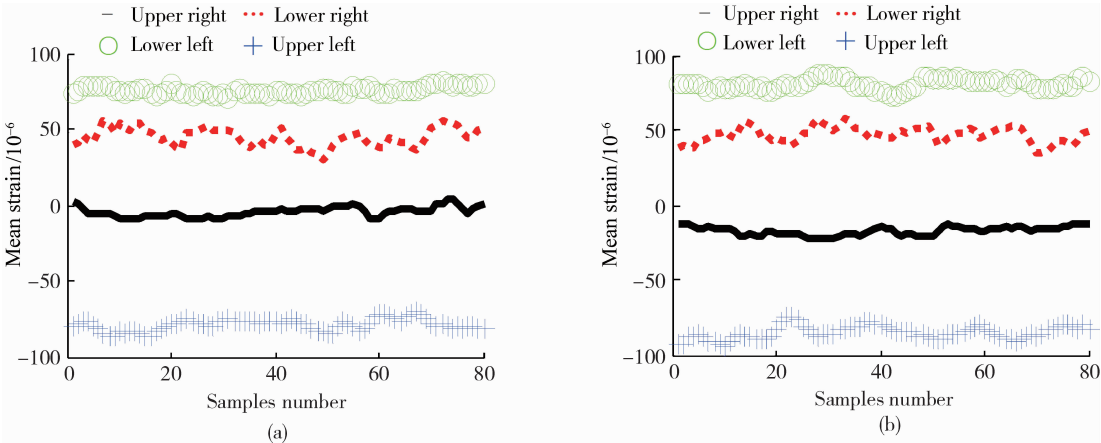


图 7 碰摩涡轮机匣垂直下应变信号的平均值

Fig. 7 Mean strain of the vertical down position on the rubbing turbine casing

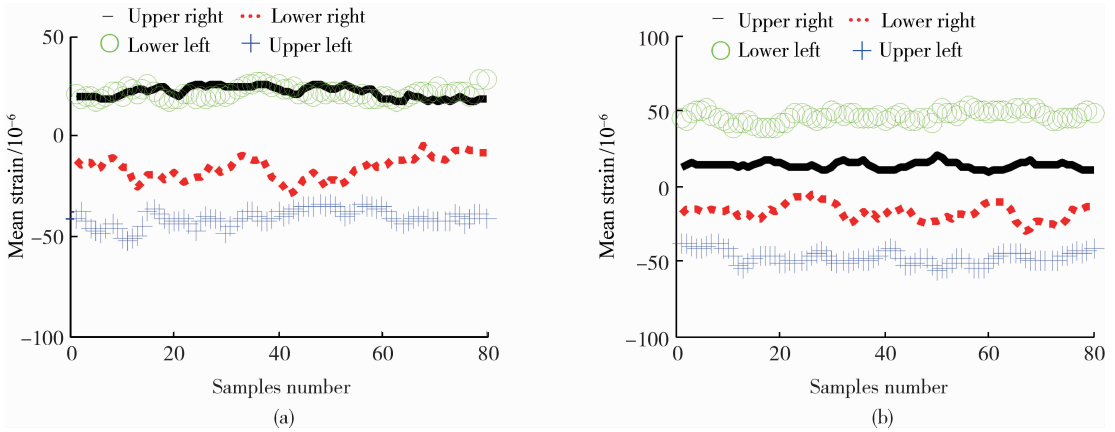


图 8 碰摩涡轮机匣水平左应变信号的平均值

Fig. 8 Mean strain of the horizontal left position on the rubbing turbine casing

3 基于 SVC 的知识规则提取方法

3.1 支持向量聚类算法 SVC

SVC^[16]是一种新颖的聚类算法,其基本思想是:将训练样本经过核函数的非线性变换映射到高维的特征空间中,然后在此空间中找到一个具有最小半径的超球包围所有样本点,并得到样本的聚类分配矩阵。

3.1.1 SVC 的数学描述

设训练样本集 $\{x_1, x_2, \dots, x_n\} \subseteq X$ (其中 $X \subseteq R^d, R^d$ 为特征空间),通过非线性变换 φ 将 x_j 映射到一个高维的特征空间中,在此空间中,寻找一个具有最小半径的超球包围所有样本点,超球的半径设为 R . 则有

$$\|\varphi(x_j) - a\|^2 \leq R^2 \quad \forall j \quad (1)$$

式中 $\|\cdot\|$ 是欧式距离, a 是超球的球心。

考虑实际情况中若干点在超球之外,引入松弛变量 ζ_j

$$\|\varphi(x_j) - a\|^2 \leq R^2 + \zeta_j \quad (2)$$

上述问题的数学描述为

$$\min R^2 + C \sum_{j=1}^N \zeta_j \quad (3)$$

$$\text{s. t.} \quad \|\varphi(x_j) - a\|^2 \leq R^2 + \zeta_j \quad (4)$$

$$\zeta_j \geq 0 \quad (5)$$

其中 C 是惩罚因子。 C 值越大,超出球的噪声点越少。为解决上述问题引入 Lagrange 乘式有

$$L = R^2 - \sum_{j=1}^N (R^2 + \zeta_j - \|\varphi(x_j) - a\|^2) \beta_j - \sum_{j=1}^N \zeta_j \mu_j + C \sum_{j=1}^N \zeta_j \quad (6)$$

式中 $\beta_j \geq 0, \mu_j \geq 0$ 为 Lagrange 乘子,惩罚因子 C 是常数, $C \sum_{j=1}^N \zeta_j$ 为惩罚项。对 Lagrange 乘式中的 a, R, ζ 分别求 1 阶偏导,并令其等于 0 得到

$$\frac{\partial L}{\partial R} = 2R - 2 \sum_{j=1}^N R \beta_j = 0 \Rightarrow \sum_{j=1}^N \beta_j = 1 \quad (7)$$

$$\frac{\partial L}{\partial a} = \sum_{j=1}^N \beta_j [\varphi(x_j) - a] = 0 \Rightarrow a = \sum_{j=1}^N \beta_j \varphi(x_j) \quad (8)$$

$$\frac{\partial L}{\partial \zeta_j} = -\beta_j - \mu_j + C = 0 \Rightarrow \beta_j = C - \mu_j \quad (9)$$

由 KKT(Karush-Kuhn-Tucker) 定理可知条件,最优解满足

$$\zeta_j \mu_j = 0 \quad (10)$$

$$[R^2 + \zeta_j - \|\varphi(x_j) - a\|^2] \beta_j = 0 \quad (11)$$

根据式(7)~式(9)化简参数只剩下 β_j ,转化为对偶形式,有

$$\max W =$$

$$\sum_{j=1}^N \varphi(x_j)^2 \beta_j - \sum_{i=1}^N \sum_{j=1}^N \beta_i \beta_j \varphi(x_i) \cdot \varphi(x_j) \quad (12)$$

满足下列条件:

$$\sum_{j=1}^N \beta_j = 1 \quad (13)$$

$$0 \leq \beta_j \leq C \quad j = 1, 2, 3, \dots, N \quad (14)$$

定义满足 Mercer 条件的核函数

$$K(x_i, x_j) = \varphi(x_i) \varphi(x_j) \quad (15)$$

式中“ $\cdot$ ”表示内积。在支持向量聚类中常用的核函数为高斯径向基核:

$$K(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / q^2) \quad (16)$$

则

$$W = \sum_{i=1}^N K(x_j, x_j) \beta_j - \sum_{i=1}^N \sum_{j=1}^N \beta_i \beta_j K(x_i, x_j) \tag{17}$$

根据式(13)、式(14)、式(16)、式(12)变为

$$\max W = 1 - \sum_{i=1}^N \sum_{j=1}^N \beta_i \beta_j K(x_i, x_j) \tag{18}$$

即

$$D(x_i) = \sqrt{\sum_{i=1}^N \sum_{j=1}^N \beta_i \beta_j K(x_i, x_j) + K(x_i, x_i) - 2 \sum_{j=1}^N K(x_j, x_i) \beta_j} \tag{21}$$

那么在特征空间中, $R = D(x_i) \mid \forall 0 < \beta_i < C$ 表示最小包含超球半径, 其中 β_i 为任一支持向量.

3.1.2 特征空间类簇的标识

SVC 算法并没有区分哪些点属于哪些类, 因此还需要进行聚类标识. 其中经典的聚类标识算法的思想为: 对于同属于一个簇的两个样本点, 那么在高维空间中, 这两个点的连接线上的所有点都不在球外, 据此构造邻接矩阵 A

$$A_{ij} = \begin{cases} 1 & \text{对于 } x_i \text{ 与 } x_j \text{ 的连接线上的点 } y \text{ 都满足 } D(y) \leq R \\ 0 & \text{其他} \end{cases} \tag{22}$$

在实际中, 考虑到计算量, 通常在连接线上取 10~20 个采样点, 只要它们都满足条件, 即可认为 $D(y) \leq R$ 成立, 然后找出由邻接矩阵 A 表示的图中的连通部分, 一个连通部分就表示一个聚类, 对于那些只有一个元素的连通部分, 标记为噪声.

由于支持向量是位于特征空间超球的表面上, 决定了特征空间超球的轮廓及大小, 而对整个样本数据点进行聚类标识需要花费大量的时间, 因此本文算法中, 我们仅对支持向量进行聚类标识, 然后进行规则提取. 经过实验表明, 可以大大降低计算量.

3.1.3 SVC 中参数的影响

惩罚因子 C 和高斯核参数 q 为对 SVC 聚类算法影响比较大的两个参数, 惩罚因子 C 控制最大允许的错误率, 高斯核参数 q 决定了样本空间中类簇的数量. 以两个仿真样本为例来说明这两个参数对 SVC 结果的影响.

对于样本 I 如图 9 所示, 当 $q=5$ 时, 在样本空间中只有一个簇; 当 q 值增大为 30 时, 样本空间中变为 3 个簇, 说明 q 的大小影响样本空间中类簇的数量. 对于样本 II, 如图 10 所示, 当 $C=1$ 时, 包括噪声在内的所有的样本都属于算法 SVC 生成的簇. 当 $C=0.1$ 时, 有 3 个样本(孤立

$$\min W' = \sum_{i=1}^N \sum_{j=1}^N \beta_i \beta_j K(x_i, x_j) \tag{19}$$

$$\begin{aligned} \text{s. t. } & 0 \leq \beta_i \leq C \\ & j = 1, 2, 3, \dots, N \end{aligned} \tag{20}$$

从而解出 β_i .

特征空间中的一个点到其最小包含超球球心的距离可以写为

点)位于簇之外, 位于簇之外的样本点越多, 得到的簇覆盖样本也更加紧密, 说明参数 C 决定了类簇的紧致性. 因此选择合适的参数可以提高规则的识别率.

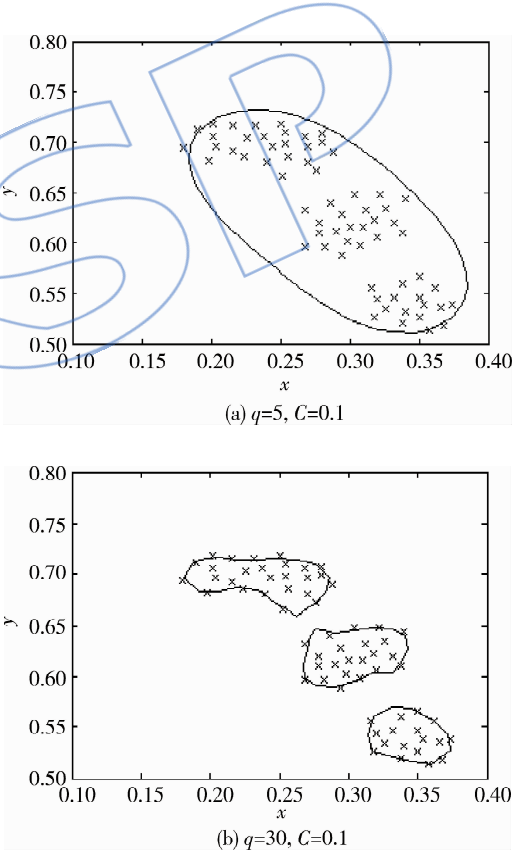


图 9 不同参数 q 时类簇的个数
Fig. 9 Effect of the scale parameter q on the number of the cluster

3.2 超矩形规则

基于 SVC 的超矩形规则提取方法的基本思想是在利用 SVC 得到覆盖样本空间的类簇基础上构造超矩形, 一个类簇决定一个超矩形 H^{j, L_j} , 由类簇中各个属性的区间 $x_1 \in [x_1^l, x_1^u] \cap \dots \cap x_i \in [x_i^l, x_i^u] \cap \dots \cap x_N \in [x_N^l, x_N^u]$ 来定义, 其中 L_j

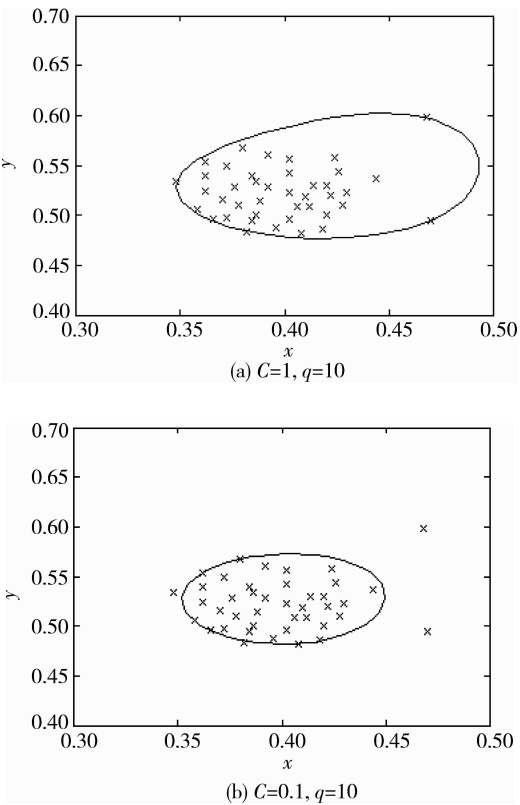


图 10 不同参数 C 时类簇的紧致性
Fig. 10 Effect of the parameter C on cluster compactness

代表类标号, x_i^l, x_i^u 分别代表类簇中样本第 i 个属性的取值区间. 将超矩形 H^{j, L_j} 投影到坐标轴上, 即得到如下的 if-then 规则 R^{j, L_j} :

$$R^{j, L_j}: \text{if } x_1 \in [x_1^l, x_1^u] \cap \cdots \cap x_i \in [x_i^l, x_i^u] \cap \cdots \cap x_N \in [x_N^l, x_N^u]$$
$$\text{then class } L_j \quad (23)$$

在统计分析中, 用于评价关联规则有效性的两个指标为支持度 (support) 和置信度 (confidence). 对于超矩形 H^{j, L_j} 规则 R^{j, L_j} , 我们定义如下的支持度和置信度:

$$\text{supp.}(R^{j, L_j}) = \frac{N_{H^{j, L_j}, L_j}}{N_{L_j}} \quad (24)$$

$$\text{conf.}(R^{j, L_j}) = \frac{N_{H^{j, L_j}, L_j}}{N_{H^{j, L_j}}} \quad (25)$$

式中 N_{H^{j, L_j}, L_j} 表示被超矩形 H^{j, L_j} 覆盖并且类标号为 L_j 的样本数; $N_{H^{j, L_j}}$ 表示被超矩形 H^{j, L_j} 覆盖的样本数; N_{L_j} 表示类标号为 L_j 的样本数.

为了提取符合实际需求的规则, 我们取同时满足最小支持度阈值 (minimum support threshold, MST) 和最小置信度阈值 (minimum confidence threshold, MCT) 的规则, 在实际应用中, MCT 和 MST 的值通常预先给定.

3.3 基于规则的样本识别方法

3.3.1 距离识别法

距离识别法即根据一个样本到其最近的超矩形规则的距离来确定其类标号. 空间中的样本 $X=(x_1, \cdots, x_N)$ 到超矩形 H^{j, L_j} 的距离可以定义为

$$D(X, H^{j, L_j}) = \sqrt{\sum_{i=1}^N \{w_{f_i} \times [d_i(X, H^{j, L_j})]^2\}} \quad (26)$$

其中 w_{f_i} 是第 i 维属性的权

$$d_i(x, H^{j, L_j}) = \begin{cases} x_i - H_{\text{upper}, i}^{j, L_j} & \text{if } x_i > H_{\text{upper}, i}^{j, L_j} \\ H_{\text{lower}, i}^{j, L_j} - x_i & \text{if } x_i < H_{\text{lower}, i}^{j, L_j} \\ 0 & \text{otherwise} \end{cases} \quad (27)$$

其中 $H_{\text{lower}, i}^{j, L_j}$ 和 $H_{\text{upper}, i}^{j, L_j}$ 为超矩形第 i 个属性的范围.

3.3.2 范围识别法

范围识别法即根据规则的范围直接确定样本的类标号. 对于每一个超矩形 H^{j, L_j} , 其规则范围为 $[x_{j1}^l, x_{j1}^u] \cap \cdots \cap [x_{ji}^l, x_{ji}^u] \cap \cdots \cap [x_{jN}^l, x_{jN}^u]$, x_{ji}^l, x_{ji}^u 分别代表第 j 类样本 x_j 第 i 个分量 x_{ij} 的取值区间. 对于空间中的样本 $X=(x_1, \cdots, x_N)$, 若 $x_1 \in [x_{j1}^l, x_{j1}^u] \cap \cdots \cap x_i \in [x_{ji}^l, x_{ji}^u] \cap \cdots \cap x_N \in [x_{jN}^l, x_{jN}^u]$, 则该样本属于第 j 类.

3.4 规则的简化

利用 SVC 得到规则以后, 需要采用一些方法如规则合并、维数约简、区间延伸等来进一步简化规则, 使得其变得更加简洁, 更容易理解.

3.4.1 规则的合并

规则合并即按照最近邻的策略, 将距离比较近、支持度相对较小并且属于同一类的超矩形规则合并起来. 在规则合并过程中, 对于同属于一类的两个不同超矩形, 重叠程度越大表示两个超矩形距离越近, 这里重叠程度根据两个超矩形重叠部分的面积来衡量. 其中, 最小置信度阈值用于判断规则合并的有效性.

以图 11 中的超矩形合并为例, 首先找到距离超矩形 $H^{1,1}$ 最近的超矩形 $H^{2,1}$, 经过合并得到一个较大的超矩形 $H^{a,1}$, 然后 $H^{a,1}$ 与距离其最近的超矩形 $H^{3,1}$ 合并, 得到最后的超矩形 $H^{b,1}$. 在合并过程中, 计算每个新产生的超矩形的置信度和支持度, 如果小于预先指定的 MCT 和 MST, 则取消这次合并.

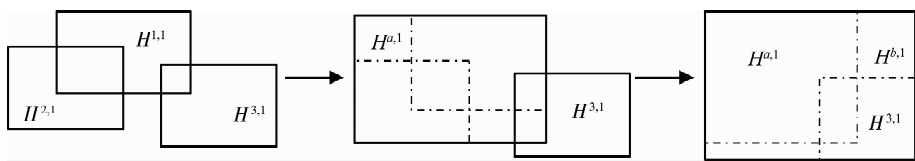


图 11 超矩形规则合并过程

Fig. 11 Hyper rectangle combination process

3.4.2 规则的约简

为了使规则更容易理解和解释,更具有代表性,需要对规则进行约简,约简过程包括区间延伸和维数约简两个步骤. 区间延伸即将生成的规则属性值的闭区间转换为开区间; 维数约简则是将规则中影响识别率比较小的属性剔除. 在约简过程中,同样采用 MCT 和 MST 进行判断,如果小于预先指定的 MCT 和 MST,则取消这次约简.

以图 12 的规则约简为例,首先对超矩形 $H^{1,1}$ 进行区间延伸,经过区间延伸之后,超矩形中属性 x_2 的闭区间转换为开区间,对应的规则变为

$$R^{1,1}: x_1 \in [x_{1l}, x_{1u}] \cap x_2 >$$
$$x_{2l} \cap x_3 \in [x_{3l}, x_{3u}] \quad (28)$$

如果将属性 x_2 从规则 $R^{1,1}$ 中除掉后,规则的置信度和支持度仍然大于指定的 MCT 和 MST,那么规则 $R^{1,1}$ 经过维数约简后可以重新写成

$$R^{1,1}: x_1 \in [x_{1l}, x_{1u}] \cap x_3 \in [x_{3l}, x_{3u}] \quad (29)$$

如果计算新生成规则的置信度和支持度,小于给定的 MCT 和 MST,则取消这一过程.

3.5 规则提取步骤

基于 SVC 的规则提取过程主要包括:数据预处理、支持向量聚类、超矩形规则提取、规则简化等步骤,整个规则提取流程如图 13 所示.

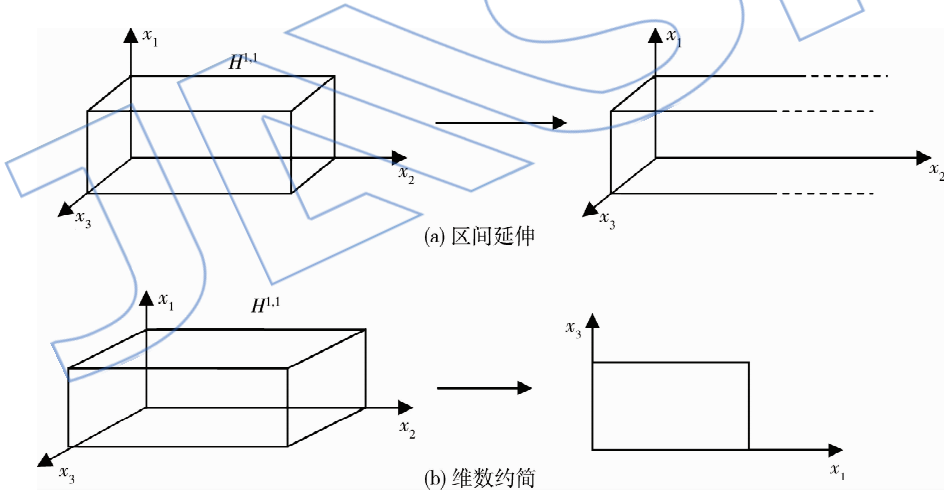


图 12 超矩形规则的约简

Fig. 12 Hyperrectangle rules reduction

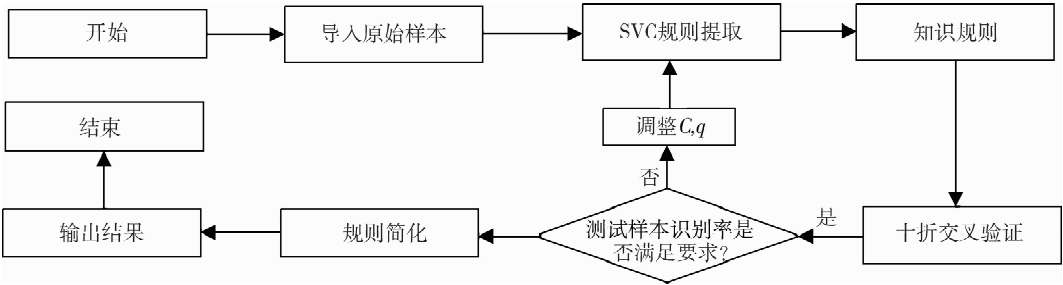


图 13 规则提取流程

Fig. 13 Rules extraction

4 基于 UCI 数据的验证分析

为了验证本文方法的有效性,本文利用国际上权威的 UCI(University of California Irvine)机器学习数据库中的 6 个数据集进行训练测试. UCI 各个数据集及其特征如表 2 所示. 将本文算法与国际上著名的开源数据挖掘软件 Weka 提供的基于 C4. 5 决策树理论方法及基于 BayesNet 的规则提取方法进行十折交叉验证测试. 比较结果如表 3 所示. 从表 3 可以看出,对比 ACT, Iris 及 Hepatitis 数据,本文方法取得最好的结果,表明本文方法的优越性.

Iris 数据是一个经典的模式分类的数据集,本文用该数据进行 SVC 规则提取方法研究. 特征选取结果如表 2 所示,MCT 和 MST 分别设为 0. 8 和 0. 1. 表 4、表 5、表 6 给出了 SVC 算法中参数 C 和 q 取不同值时得到的超矩形规则. 从表中可以看

出,当参数 $q=10$ 时,得到 3 条具有较大支持度的规则. 当 q 值增大到 20 时,规则数变大,而支持度则变小. 当参数 C 从 0. 5 减小到 0. 1 时,得到的规则的支持度也变小了,从表 6 中可以看出,3 条规则的支持度分别减少为 0. 71,0. 64 和 0. 68.

表 2 各数据集及其属性特征选取
Table 2 Training set and its characteristics

数据集	训练集大小	测试集大小	特征数量
ACT	126	14	6
Ecoli	302	34	7
Iris	135	15	4
Glass	193	21	9
Hepatitis	139	16	19
Wine	160	18	13

表 3 各算法在 UCI 数据集上的识别率比较
Table 3 Recognition rate on different training sets with different methods

	ACT	Ecoli	Iris	Glass	Hepatitis	Wine
SVC	0. 8929	0. 7581	0. 98	0. 5111	0. 8667	0. 9375
C4. 5	0. 8857	0. 8423	0. 96	0. 6589	0. 8129	0. 9382
BayesNet	0. 8571	0. 8125	0. 9267	0. 7477	0. 8323	0. 9888

表 4 Iris 数据集 SVC 规则提取结果 ($C=0. 5,q=10$)
Table 4 Results of Iris data set from SVC rule ($C=0. 5,q=10$)

超矩形规则	支持度和 置信度	距离 识别法	范围 识别法
if $x_1 \in [4. 4, 5. 7]$ $x_2 \in [2. 9, 4. 4]$ $x_3 \in [1. 2, 1. 9]$ $x_4 \in [0. 1, 0. 5]$ then class 1	[0. 91, 0. 97]		
if $x_1 \in [5. 2, 6. 9]$ $x_2 \in [2. 3, 3. 4]$ $x_3 \in [3. 5, 5. 0]$ $x_4 \in [1. 1, 1. 8]$ then class 2	[0. 76, 0. 94]	0. 95	0. 88
if $x_1 \in [5. 6, 7. 4]$ $x_2 \in [2. 5, 3. 4]$ $x_3 \in [4. 8, 6. 3]$ $x_4 \in [1. 5, 2. 5]$ then class 3	[0. 78, 0. 91]		

表 5 Iris 数据集 SVC 规则提取结果 ($C=0. 5,q=20$)
Table 5 Results of Iris data set from SVC rule ($C=0. 5,q=20$)

超矩形规则	支持度和 置信度	距离 识别法	范围 识别法
if $x_1 \in [4. 6, 5. 4]$ $x_2 \in [3. 0, 4. 4]$ $x_3 \in [1. 3, 1. 9]$ $x_4 \in [0. 1, 0. 6]$ then class 1	[0. 91, 1]		
if $x_1 \in [5. 4, 7. 0]$ $x_2 \in [2. 3, 3. 3]$ $x_3 \in [3. 7, 4. 9]$ $x_4 \in [1. 1, 1. 7]$ then class 2	[0. 78, 0. 97]		
if $x_1 \in [6. 2, 7. 2]$ $x_2 \in [2. 7, 3. 4]$ $x_3 \in [4. 8, 5. 3]$ $x_4 \in [1. 5, 2]$ then class 3	[0. 31, 0. 82]	0. 98	0. 85
if $x_1 \in [6. 0, 6. 3]$ $x_2 \in [2. 5, 3. 0]$ $x_3 \in [5. 3, 6. 1]$ $x_4 \in [2. 1, 2. 5]$ then class 3	[0. 2, 1]		

对利用距离判别法得到的识别率最高的规则(如表 5)进行进一步的简化. 表 7 为表 5 进行规则合并后的结果, 支持度和置信度都变大了, 测试集的识别率仍为 0.98. 然后对规则进行约简, 包括区间延伸和维数归约, 得到简洁、具有较强解释性的

规则, 直接根据所得范围判断测试样本的类别, 得到的识别率为 0.9867, 如表 8 所示, 即根据最后所得的规则, 可以直接判断所要求样本的类别, 表明所得规则具有很强的解释性.

表 6 Iris 数据集 SVC 规则提取结果 ($C=0.1, q=10$)
Table 6 Results of Iris data set from SVC rule ($C=0.1, q=10$)

超矩形规则	支持度和 置信度	距离 判别法	范围 判别法
if $x_1 \in [4.6, 5.5]$ $x_2 \in [3.0, 4.2]$ $x_3 \in [1.3, 1.7]$ $x_4 \in [0.1, 0.5]$ then class 1	[0.71, 0.99]		
if $x_1 \in [5.4, 6.9]$ $x_2 \in [2.4, 3.4]$ $x_3 \in [3.5, 4.9]$ $x_4 \in [1, 1.6]$ then class 2	[0.64, 1]	0.97	0.83
if $x_1 \in [5.6, 6.9]$ $x_2 \in [2.5, 3.4]$ $x_3 \in [4.8, 6.0]$ $x_4 \in [1.5, 2.5]$ then class 3	[0.68, 1]		

表 7 规则合并结果
Table 7 Rules combination results

超矩形规则	支持度和 置信度	距离 识别法	范围 识别法
if $x_1 \in [4.6, 5.4]$ $x_2 \in [3.0, 4.4]$ $x_3 \in [1.3, 1.9]$ $x_4 \in [0.1, 0.6]$ then class 1	[0.91, 1.0]		
if $x_1 \in [5.4, 7.0]$ $x_2 \in [2.3, 3.3]$ $x_3 \in [3.7, 4.9]$ $x_4 \in [1, 1.7]$ then class 2	[0.82, 0.97]	0.98	0.88
if $x_1 \in [6.0, 7.2]$ $x_2 \in [2.5, 3.4]$ $x_3 \in [4.8, 6.1]$ $x_4 \in [1.5, 2.5]$ then class 3	[0.78, 0.97]		

表 8 规则约简结果
Table 8 Rules reduction results

超矩形规则	支持度和 置信度	范围 识别法
if $x_4 \leq 0.6$ then class 1	[0.91, 1]	
if $x_4 \in [1, 1.6]$ then class 2	[0.8, 0.97]	0.99
if $x_3 \geq 4.8$ then class 3	[0.78, 0.97]	

5 航空发动机碰摩部位识别知识
规则提取

利用本文基于 SVM 的数据挖掘方法, 从大量的碰摩部位实验数据中提取转静碰摩部位诊断的知识规则. 本文实验数据来自带机匣的航空发动机转子实验器的碰摩实验, 碰摩位置分别为涡轮机匣垂直上、水平右、垂直下、水平左 4 个方向. 利用机匣 4 个部位的应变测试, 进行了 3 次独立实验, 获取了 4 个碰摩部位和 4 个测点的大量实验数据. 将 3 次实验数据混合在一起, 数据分布图如图 14 所示.

任意选取其中三分之一数据进行模型训练,

将剩余样本作为未知样本进行测试. 利用十折交叉验证法得到 SVC 算法中的惩罚因子 C 和高斯核参数 q , 优化得到的 C 和 q 分别为 0.1 和 0.15, MCT 和 MST 分别设为 0.9 和 0.1. 重复 10 次随机选取三分之一样本进行训练, 剩余三分之二进行测试. 表 9 给出了规则对测试样本的平均识别率, 在该表中列出了距离识别法和范围识别法得到的识别率. 在这两种情况下, 均达到了 90% 以上, 表明该算法提取的规则具有很好的质量. 由于篇幅限制仅列出其中一次测试得到的规则, 如表 10 所示.

为了使规则更加简洁, 易于解释, 采用区间延伸、维数约简对超矩形规则进行进一步简化. 表 11 为进行规则约简后, 对测试集进行测试得到的识别率. 从表中可以看出: 经过规则约简后, 范围识别法的识别率从 0.9348 增长为 0.9939, 识别率大大增加. 表 12 列出了约简之后得到的规则, 简洁并且具有很强的解释性. 因此, 得到该约简后的规则后, 可以直接判断测试样本类别, 而不用再求样本到其最近的超矩形规则的距离来确定其类标号.

从表 12 可以看出规则中仅用到了特征 3 和特

征 4,说明利用特征 3 和特征 4 即可实现碰摩故障部位的识别。从图 14 的数据特征分布图中可以看出,特征 3 和特征 4 对样本的分类能力最强,由此可见,本文所提出的规则反映了数据的本质特征。

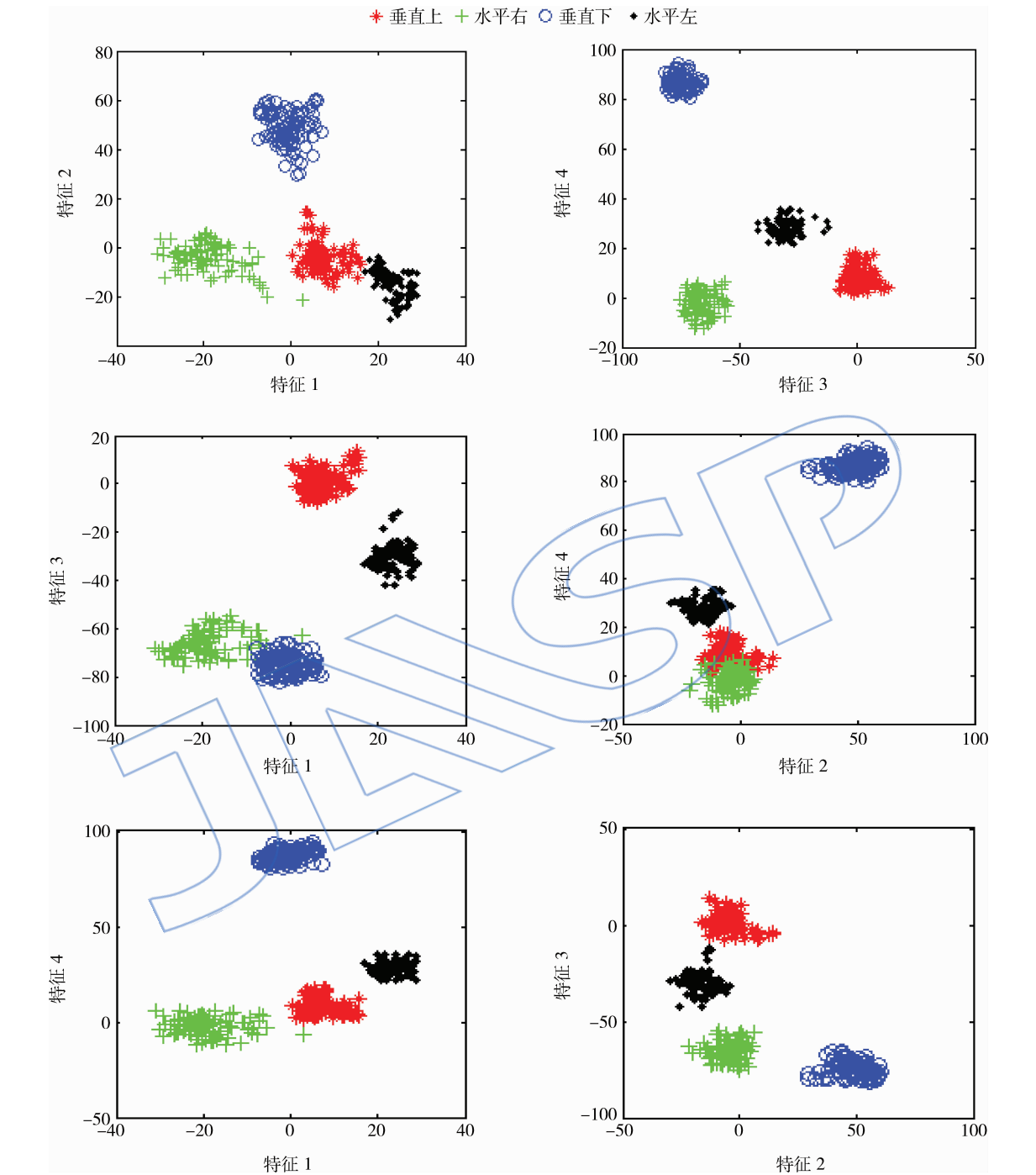


图 14 实验数据特征分布图
Fig. 14 Experimental data feature profile

表 9 10 次测试样本碰摩部位识别率
Table 9 Recognition rate of rubbing position for 10 times samples

	1	2	3	4	5	6	7	8	9	10	平均
距离识别法	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
范围识别法	0.92	0.93	0.93	0.95	0.92	0.92	0.96	0.92	0.94	0.96	0.935

表 10 规则提取结果
Table 10 Rules reduction results

超矩形规则	支持度和置信度
if $x_1 \in [0.45, 15.8495]$ $x_2 \in [-15.789, 14.4552]$ $x_3 \in [-7.6973, 13.4359]$ $x_4 \in [1.3080, 17.8201]$ then class 1	[1.00, 1.00]
if $x_1 \in [-30.7812, 2.7875]$ $x_2 \in [-21.4372, 5.2064]$ $x_3 \in [-75.4992, -54.849]$ $x_4 \in [-12.1510, 6.1471]$ then class 2	[1.00, 1.00]
if $x_1 \in [-7.3645, 7.1107]$ $x_2 \in [29.6294, 59.6560]$ $x_3 \in [-82.027, -65.8358]$ $x_4 \in [80.4158, 93.0757]$ then class 3	[1.00, 1.00]
if $x_1 \in [17.1285, 28.989]$ $x_2 \in [-27.4223, -4.9217]$ $x_3 \in [-42.0648, -12.0789]$ $x_4 \in [22.0242, 35.3310]$ then class 4	[1.00, 1.00]

表 11 规则约简后 10 次测试样本碰摩部位识别率
Table 11 Recognition rate of rubbing position after rule reduction for 10 times samples

	1	2	3	4	5	6	7	8	9	10	平均
距离识别法	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
范围识别法	0.99	0.98	0.99	1.00	1.00	1.00	0.99	0.99	0.99	1.00	0.993

表 12 约简后规则列表
Table 12 Rules after reduction

超矩形规则	支持度和置信度
if $x_3 \geq -7.6973$ then class 1	[1.00, 1.00]
if $x_3 \leq -54.849 \cap x_4 \leq 6.1471$ then class 2	[1.00, 1.00]
if $x_4 \geq 80.4158$ then class 3	[1.00, 1.00]
if $x_4 \in [21.5173, 35.5964]$ then class 4	[1.00, 1.00]

6 结 论

1) 利用带机匣的航空发动机转子实验器,进行了 4 个不同部位的碰摩实验,利用机匣 4 个部位的应变测试,获取了 4 个碰摩部位和 4 个测点的大量实验数据,提取了碰摩故障特征,形成了大量的碰摩样本。

2) 提出了一种基于 SVC 的知识规则提取新方法,利用 SVC 算法得到特征选取后样本的聚类分配矩阵,最后根据聚类分配矩阵构建超矩形规则。为了使规则更加简洁,易于解释,采用规则合并、区间延伸、维数约简等方法对超矩形规则进行进一步简化。

3) 用 UCI 机器学习数据,将本文所取算法与开源数据挖掘软件 Weka 提供的基于 C4.5 决策树理论的规则提取方法和基于 BayesNet 的规

则提取方法进行对比验证,结果表明本文方法具有较好的泛化性能。

4) 将所提出的基于 SVM 的数据挖掘方法应用于航空发动机碰摩部位识别和诊断知识获取研究中,从大量的碰摩部位实验数据中提取出了转静碰摩部位诊断的知识规则,并进行了相应解释和验证,规则验证结果表明规则具有很高的精度。因此该方法能较好地满足工程需要,可以作为航空发动机碰摩部位识别和诊断知识规则提取的有效工具,将有效地提升专家系统的智能水平和知识获取能力,对于有效地实现发动机状态监测和视情维修具有重要意义。

参考文献:

[1] 王德友. 发动机转静碰摩振动特性的提取与理论研究[D]. 北京:北京航空航天大学,1995.
WANG Deyou. Research and extract of vibration features

- on rubbing between rotor and stator for aero-engine[D]. Beijing: Beijing University of Aeronautics and Astronautics, 1995. (in Chinese)
- [2] Wang Q, Chu F. Experimental determination of the rubbing location by means of acoustic emission and wavelet transform[J]. Journal of Sound and Vibration, 2001, 248(1): 91-103.
- [3] Han Q K, Yu T, Li H, et al. Hybrid model based identification of local rubbing fault in rotor systems[J]. Key Engineering Materials, 2005, 293-294: 355-364.
- [4] Bachschmid N, Pennacchi P, Vania A. Identification of multiple faults in rotor systems[J]. Journal of Sound and Vibration, 2002, 254(2): 327-366.
- [5] Chu F, Lu W. Determination of the rubbing location in a multi-disk rotor system by means of dynamic stiffness identification[J]. Journal of Sound and Vibration, 2001, 248(2): 235-246.
- [6] HAN Jiawei, Kamber M. Data mining: concepts and techniques[M]. San Francisco, US: Morgan Kaufmann Publishers, 2001.
- [7] Mannila H. Data mining: machine learning, statistics, and databases[C]//Proceedings of Eight International Conference on Scientific and Statistical Database Systems. Los Alamitos, US: IEEE Computer Society Press, 1996: 2-9.
- [8] FU Chaoyang, ZHENG Jiashen, ZHAO Jingmao, et al. Application of grey relational analysis for corrosion failure of oil tubes[J]. Corrosion Science, 2001, 43(5): 881-889.
- [9] Fu L. Knowledge discovery based on neural networks[J]. Communications of the ACM, 1999, 42(11): 47-50.
- [10] Ziarko W. Discovery through rough set theory[J]. Communications of the ACM, 1999, 42(11): 55-57.
- [11] Quinlan J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.
- [12] 张英. 基于支持向量机的过程工业数据挖掘技术研究[D]. 杭州: 浙江大学, 2005.
- ZHANG Ying. A study on process industrial data mining on support vector machine[D]. Hangzhou: Zhejiang University, 2005. (in Chinese)
- [13] Vapnik V. Statistical learning theory[M]. New York: John Wiley & Sons Inc, 1998.
- [14] Barakat N, Bradley A P. Rule-extraction from support vector machines: a review[J]. Neurocomputing, 2010, 74(1/2/3): 178-190.
- [15] Zhang Y, Su H, Jia T, et al. Rule extraction from trained support vector machines[C]//Proceedings of Ninth Pacific-Asia Conference, Advances in Knowledge Discovery and Data Mining. Berlin, Germany: Springer Berlin Heidelberg, 2005: 61-70.
- [16] Ben-Hui A, Hom D, Sidgelman H T. Support vector clustering[J]. Journal of Machine Learning Research, 2002, 2: 125-137.