

文章编号: 1000-8055(2008)10-1933-06

聚类分析的数据挖掘方法及其在机械传动故障诊断中的应用

邵忍平, 黄欣娜, 胡军辉
(西北工业大学 机电学院, 西安 710072)

摘 要: 采用网格化处理的思想, 通过对基于密度的聚类分析方法进行改进, 提出了一种新的聚类算法. 这种算法通过对齿轮传动系统的故障信号进行测试、对故障类型进行了判定, 对不同转速下齿轮传动振动信号进行谱熵计算, 并采用网格划分方法将其表示在二维和三维空间分布平面内, 可以较好地正常、裂纹、磨损等类型的故障进行聚类 and 识别, 并通过试验验证表明能够对不同工作状态的齿轮传动信号进行可靠的聚类与区分, 聚类率为 96% 以上. 说明该方法对齿轮故障进行区分与诊断是切实可行和有效的.

关 键 词: 故障诊断; 齿轮传动系统; 网格化处理; 聚类分析; 数据挖掘

中图分类号: TH113.1

文献标识码: A

Analysis of data mining of clustering and its application to mechanical transmission fault diagnosis

SHAO Ren-ping, HUANG Xin-na, HU Jun-hui

(School of Mechatronics, Northwestern Polytechnical University, 710072 Xi'an, China)

Abstract: For mechanical fault in the gear transmission system difficult to diagnosis, an improved clustering method was presented based on the DBSCAN (Density-based spatial clustering of application with noise) by using the concept of grid processing. The faults of gear transmission system were detected and the styles of gear faults were estimated by the improved clustering method, which also calculated spectral entropy of vibration signal of gear transmission at different rotate speed, and showed the spectral entropy distribution in Two-dimension and Three-dimension domain by grid plotting. So it enables better clustering and identification of gear faults such as natural, cracked and attrited faults, etc. The method is verified experimentally to be capable of distinguishing reliably gear transmission signals at different work conditions for gear transmission system, with clustering rate up to 96%. It indicates that the method is feasible for classification and diagnosis of gear faults.

Key words: fault diagnosis; gear transmission system; grid processing; clustering; data mining

聚类分析是数据挖掘的主要方法之一, 由于其简单、有效, 已经成为数据挖掘研究领域中的一个

非常活跃的研究方向. 聚类分析就是从给定的数据集中搜索数据对象之间所存在的有价值的联

收稿日期: 2007-10-15; **修订日期:** 2008-02-13

基金项目: 国家自然科学基金(50575187); 航空科学基金(01I53073); 陕西省自然科学基金(2004E219); 西北工业大学研究生创业种子基金(Z200524)

作者简介: 邵忍平(1962-), 男, 副教授, 博士后. 主要从事机械动力学、振动噪声与控制、动态信号处理与识别、智能计算检测与诊断等方面的研究.

系,将数据对象分组,用于区分不同类的对象.其目标是使组内的对象相互之间是相似的,而不同组中的对象是不同的.目前数据挖掘方法和机械故障诊断的结合已经取得了一定的研究成果^[1-2],但是目前该技术在机械故障诊断领域的研究尚处于探讨的阶段^[3].

基于密度的方法 DBSCAN^[4] (Density-based spatial clustering of application with noise) 是利用密度连通性可以快速发现任意形状的聚类,只要 ε 邻域内的数据对象数目不少于 Minpts 就可将其作为一个簇.但由于其对用户定义的参数 ε 非常敏感,往往在以核心对象为圆心,小于 ε 的范围内数据对象的数目已经达到了 Minpts ,使得一些高密度的结果被完全包含在较低密度的簇中,从而失去有意义的信息.而且计算复杂度可能是 $O(n^2)$,这对于数据对象较多的情况是不可取的. OPTICS^[5] (Ordering pointers to identify the clustering structure) 方法利用生成一个参数化的数据集合并排序来对 DBSCAN 的缺点进行改进,也就是将核心对象的 ε 邻域内的数据对象到核心对象的距离按照从小到大的顺序插入到一个种子队列中.但是该方法需要对数据对象到核心对象之间的聚类严格按照大小排序,这是非常的繁琐,同时如果对属于某一个簇的数据,当对象数目较多时也存在计算量较大的问题.考虑到上面方法的这些缺点,本文提出采用网格化分的方法对其进行有效地改进,从而实现了数据对象的有效聚类与区分.

1 改进的基于密度和网格划分的方法

网格划分就是将整个数据对象集合或部分数据作网格化处理,将数据集合按照其空间位置放在特定的网格中.如对于二维数据对象,可看作其分布在一个有边界的平面上,将该平面划分成网格,对象的位置由对象坐标唯一确定.多维数据对象集合也可这样处理.这样就把原本无序的数据转换成一种空间有序结构,搜索时只是检索当前对象的空间相邻区域就可以找出所有的邻域对象.

网格的大小直接影响到计算量,网格太大使邻域搜寻中扫描的点数过多,网格太小扫描的网格数又过多,扫描的网格的定位就变得复杂.数据对象中,只要有一个点以其为核心点, ε -邻域内的数据对象数目达到 Minpts 就可以形成一个簇,因此网格的大小可以由 ε 确定.以 2ε 作为网格的

边长进行网格划分.对于多维数据对象,可以根据其各维的 ε 邻域取 2ε 作为网格的边长,网格的数目由各维的长度与网格的边长的比值 $\text{length}/2\varepsilon$ 取整来确定.

在基于密度的聚类方法的基础上利用网格划分来对其改进优化,具体的过程可以用下面的算法描述.

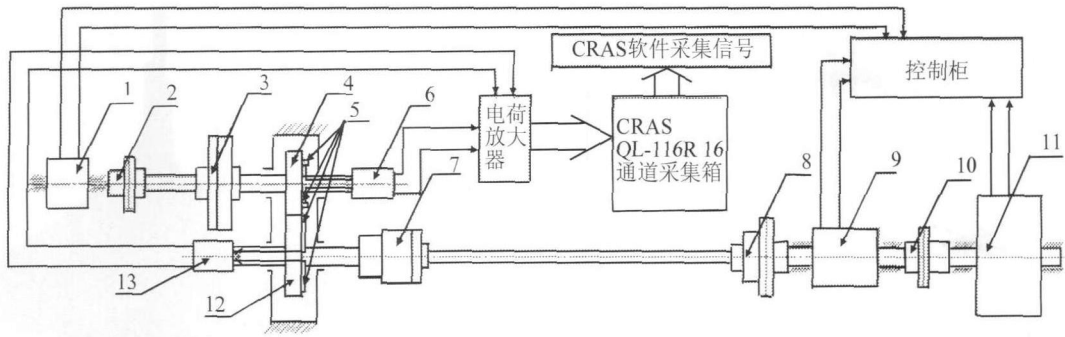
基于密度和网格划分的算法:

- 1) 确定网格大小、网格数、数据对象所处的网格的位置和 Minpts ;
- 2) 扫描整个网格,将网格中的数据点数 $\text{Pt-num} < \text{Minpts}$ 的网格中的数据点作为噪声处理,将 $\text{Pt-num} \geq \text{Minpts}$ 的网格存到一个数组 Array 中;
- 3) while Array 中有网格中的数据对象所属的簇没有确定;
- 4) 从前往后扫描 Array ,如果某一个网格 grid 未标记,也即其所属的簇还不确定,则将该网格的坐标存入堆栈 Stack ,并对其标记;
- 5) 根据存入堆栈中网格 grid 的坐标,扫描其周围的在 Array 中存在且未标记的网格,将扫描到的这些网格的坐标也存入 Stack ,并标记;
- 6) 对第 5 步中扫描到每一个网格,扫描 Array 中处于其周围且未标记的网格,将扫描到的这些网格的坐标也存入 Stack ,并标记;
- 7) 重复第 5,6 步,直到 Array 中再找不到满足要求的网格,将 Stack 中的这些网格中的数据对象作为一个簇;
- 8) end.

2 系统试验测试与动态信号提取

建立齿轮系统试验测试装置如图 1 所示.图 1 中由电动机 1 驱动整个系统运转,经过联轴器 2 和 3 后,将动力输给减速器齿轮 4 和 12,经过减速器输出后通过齿式联轴器 7 和 8,联轴器 10 将动力传递给磁粉加载器 11.图 1 中 11 是为了给系统加载而设置的,为了测定系统的载荷和转速,设置了转矩转速仪器 9.传输环 6 和 13 被设置在轴端,将随齿轮一起转动的加速度传感器 5 测得的信号准确的输送出来,用于信号分析.试验采用先进的 CRAS 信号采集软件采集信号和提取信号特征^[6],用于聚类 and 识别故障.

实验分别对无故障齿轮、齿根裂纹故障齿轮、分度圆裂纹故障齿轮和齿面磨损齿轮四种类型进



1 直流电机; 2 电机联轴器; 3 变速箱联轴器; 4 主动齿轮; 5 加速度传感器; 6 传输环; 7、8 齿式联轴器; 9 转矩转速传感器; 10 联轴器; 11 磁粉加载器; 12 被动齿轮; 13 传输环.

图 1 齿轮试验系统测试与信号提取原理图

Fig. 1 Principle diagram of test and signal extraction for gear experiment system

行减速器系统振动响应检测. 减速器的小齿轮作为信号提取检测对象, 采用加速度传感器拾取信号, 加速度传感器安装在故障齿附近. 分别选取 300, 900, 1 200, 1 500 r/min 四种转速工况, 在系统运转的情况下进行测试, 并用上述基于密度和网格划分的方法进行分析处理^[7], 进而研究故障的区分和聚类的特征, 以及故障出现时系统所表现出的动态特征.

3 谱熵计算和聚类分析与模拟

如图 1 所示的齿轮传动系统试验, 试验中直接从驱动齿轮上采集不同转速下 (300, 900, 1 200, 1 500 r/min) 四种类型工作状态的齿轮 (正常齿轮, 齿根裂纹和齿面磨损) 加速度振动信号, 每周测取 $N = 1\,024$ 个数据点. 对时域信号经过离散余弦变换后得到的频域信号, 按照频域信号幅值从小到大分段, 将分布在每段上的个数与 N 的比值作为谱熵计算的概率, 以此计算每个周期信号的谱熵值^[8]. 同时计算频谱熵的重心位置, 得到谱熵值和谱熵重心组成二维谱熵数据对象集合. 二维谱熵除了能够反映频谱分布特征和能量集中程度外, 还增加了频线熵的重心位置维指标, 用以反映谱的能量集中位置.

谱熵及其谱熵重心的计算式为

$$H(X) = - \sum_i^q P(a_i) \log_2 P(a_i) \quad (1)$$

$$C_G = \frac{\sum_k^N \frac{k}{N} X(k)}{\sum_k^N \frac{k}{N} p_k} = \sum_k^N \frac{k}{N} p_k \quad (2)$$

式中 $H(X)$ 表示谱熵, $P(a)$ 表示取不同值出现的

概率, C_G 表示谱熵重心, N 表示数据点数, p_k 表示概率, $X(k)$ 表示信号 $x(n)$ 离散余弦变换的频域信号幅值.

通过模拟与分析计算得到各种转速下不同类型的齿轮信号的二维谱熵分布如图 2~图 5 所示, 图中横坐标表示谱熵值, 纵坐标表示谱熵的重心. 从图中可以清楚地看出能够对齿轮传动系统的动态信号进行归类和分类, 这样用于故障诊断, 具有明显的效果.

上述的二维谱熵分布结果的网格化处理后的柱状图显示如图 6~图 9 所示, 图中底面表示二维谱熵所确定的平面, 柱的高度代表位于该网格中的数据对象的个数. 可以清楚地看出不同齿轮状态类型的分类情况和各类簇中具有的数据对象的多少.

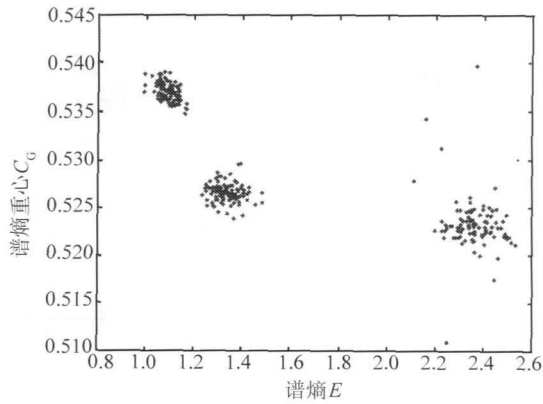


图 2 300 r/min 时二维谱熵分布
Fig. 2 Two-dimension spectral entropy distribution (300 r/min)

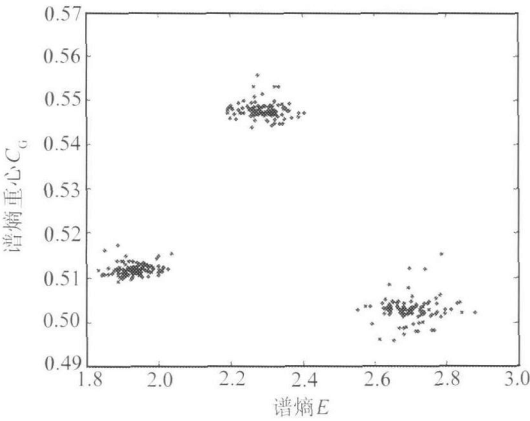


图 3 900 r/min 时二维谱熵分布
Fig. 3 Two-dimension spectral entropy distribution (900 r/min)

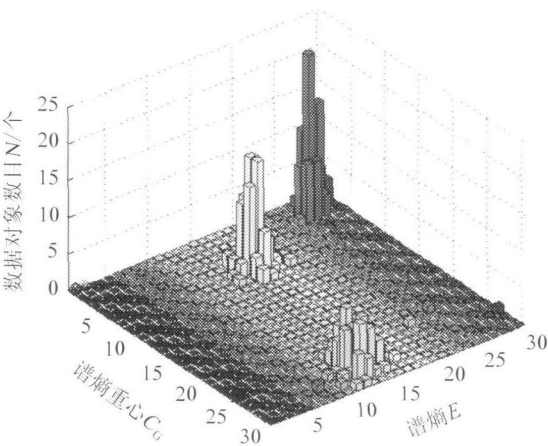


图 6 300 r/min 时二维谱熵柱状图
Fig. 6 Two-dimension spectral entropy histogram (300 r/min)

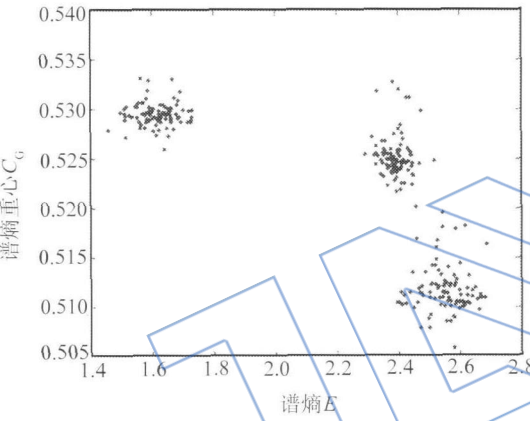


图 4 1 200 r/min 时二维谱熵分布
Fig. 4 Two-dimension spectral entropy distribution (1 200 r/min)

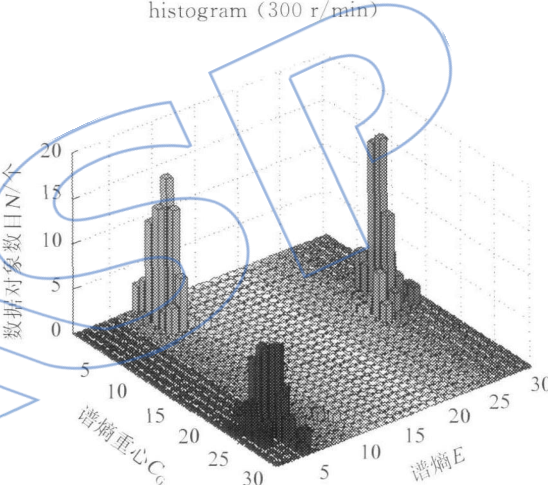


图 7 900 r/min 时二维谱熵柱状图
Fig. 7 Two-dimension spectral entropy histogram (900 r/min)

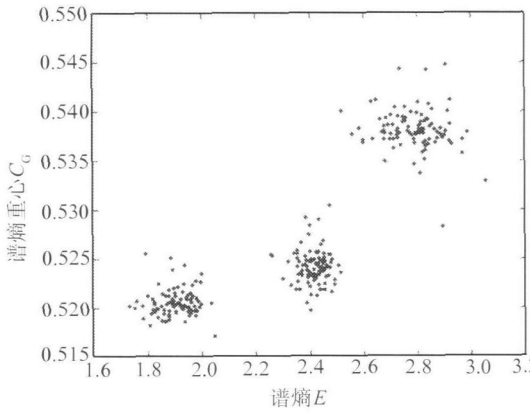


图 5 1 500 r/min 时二维谱熵分布
Fig. 5 Two-dimension spectral entropy distribution (1 500 r/min)

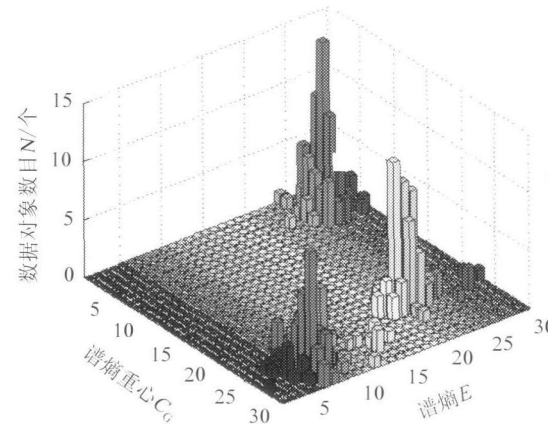


图 8 1 200 r/min 时二维谱熵柱状图
Fig. 8 Two-dimension spectral entropy histogram (1 200 r/min)

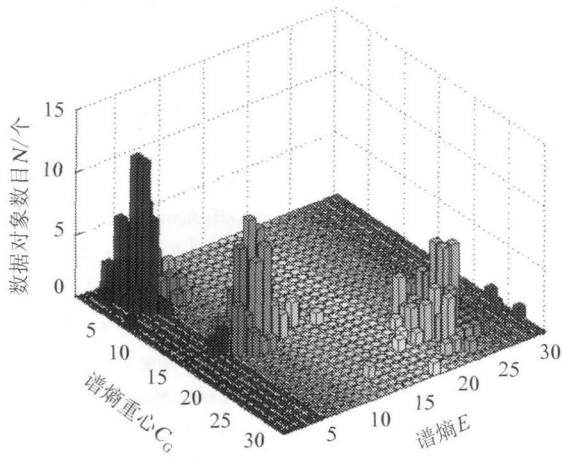


图 9 1 500 r/min 时二维谱熵柱状图

Fig. 9 Two-dimension spectral entropy histogram (1 500 r/min)

采用基于密度和网格划分算法得到的聚类二维谱熵结果如图 10~图 13 所示,图中每种符号标识为一个簇,其中的叉号表示其所显示的数据对象被作为噪声点处理.从图可见该方法能够对齿轮不同的故障进行很好地聚类,对于故障诊断具有理想的效果.

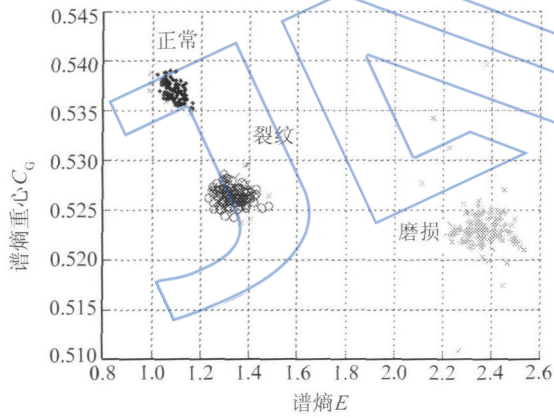


图 10 300 r/min 时二维谱熵聚类结果

Fig. 10 Clustering results of two-dimension spectral entropy (300 r/min)

另外从聚类结果还可以看出,该方法可以处理任意形状的数据集,同时对一些异常数据的处理也相当得好,可以采用适当的处理将这些异常数据屏蔽掉.但是还可以发现,有些所处的位置相对较独立但又由于达到所要求的 $Minpts$,也被划归为一个独立的簇.产生这种情况的原因是用于聚类的数据对象相对较少,而运用基于密度和网格划分的算法进行聚类时所定义的 $Minpts$ 也相对较小.如果定义 $Minpts$ 较大,则将会有较多的数据点被作为噪声处理.可以预见如果数据对象

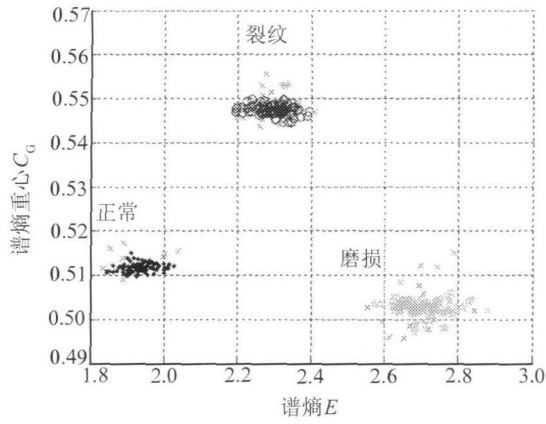


图 11 900 r/min 时二维谱熵聚类结果

Fig. 11 Clustering results of two-dimension spectral entropy (900 r/min)

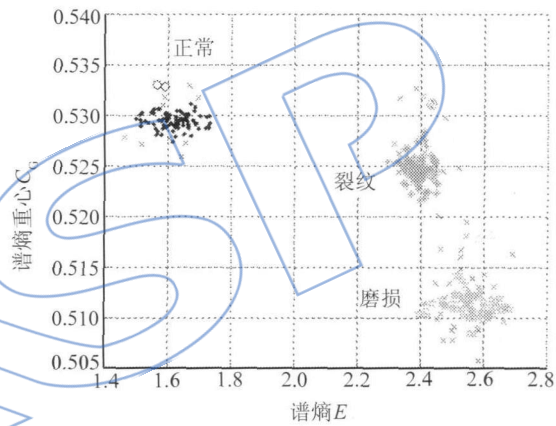


图 12 1 200 r/min 时二维谱熵聚类结果

Fig. 12 Clustering results of two-dimension spectral entropy (1 200 r/min)

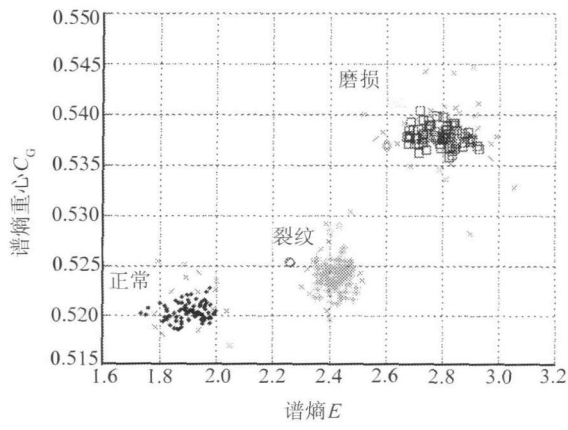


图 13 1 500 r/min 时二维谱熵聚类结果

Fig. 13 Clustering results of two-dimension spectral entropy (1 500 r/min)

集合相对较大,所得到的聚类结果将会得到改进,效果将会更好.根据熵具有等概率的场熵最大、确定场(无随机性)的熵最小、随机场分辨率越模糊

信息量越小的性质,我们可以推断,正常齿轮、裂纹齿轮和磨损齿轮的谱熵将会越来越大,就可以对数据对象进行正确的判断。

4 结 论

聚类分析作为一种实际中应用非常活跃的方法,发展非常迅速。本文在基于密度的聚类算法的基础上,对其进行网格化处理,改进后的算法可以进行简单、有效的邻域查询和类簇分析。该方法的处理时间与数据对象的数目无关,但与数据对象每一维空间所划分的单元数相关。由于改进的基于密度和网格划分的聚类方法处理时间很短,通过试验验证,效果也相当得好。因此,可以预见在同样情况下该方法可对任意形状的数据对象进行正确的聚类,同时计算速度更快,花费的时间更短。

参考文献:

- [1] SUN Weixiang, CHEN Jin, LI Jiaqing. Decision tree and PCA-based fault diagnosis of rotating machinery[J]. Mechanical Systems and Signal Processing, 2007, 21(3): 1300-1317.
- [2] Zhang S, Mathew J, Ma L, et al. Best basis-based intelligent machine fault diagnosis[J]. Mechanical Systems and Signal Processing, 2005, 19(2): 357-370.
- [3] 杨文献,姜节胜. 基于数据挖掘的柴油机气门故障诊断技术研究[J]. 机械工程学报, 2004, 40(10): 25-29.
YANG Wenxian, JIANG Jiesheng. Data mining based technique for diagnosing the faults presenting on engine valves[J]. Chinese Journal of Mechanical Engineering, 2004, 40(10): 25-29. (in Chinese)
- [4] Ng R T, Han J. Efficient and effective clustering methods for spatial data mining [C]// Proc. of the 20th VLDB Conf. Santiago: Morgan Kaufmann, 1994: 144-155.
- [5] Ankerst M, Breunig M, Kriegel H-P, et al. OPTICS: Ordering points to identify the clustering structure [C]// Proc. 1999ACM-SIGMOD Int. Conf. Management of Data Philadelphia, PA: [s. n], 1999, 6: 49-60.
- [6] 邵忍平,沈允文,郭万林. 基于时频分布的齿轮系统辨识研究[J]. 中国机械工程, 2001, 12(7): 738-740.
SHAO Renping, SHEN Yunwen, GUO Wanlin. A new method for identifying the gear transmission system by the time-frequency distribution [J]. China Mechanical Engineering, 2001, 12(7): 738-740. (in Chinese)
- [7] 王宏禹. 非平稳随机信号分析与处理[M]. 北京:国防工业出版社, 1999: 68-82.
- [8] 胡军辉,邵忍平,曾泽君. 基于谱熵的齿轮故障诊断方法研究[J]. 机械传动, 2007, 31(5): 84-87.
HU Junhui, SHAO Renping, ZENG Zejun. Method of gear's fault diagnosis based on spectral entropy[J]. Journal of Mechanical Transmission, 2007, 31(5): 84-87. (in Chinese)