文章编号:1671-4598(2023)01-0051-08 DOI:10.16526/j.cnki.11—4762/tp.2023.01.008 中图分类号:TP277 文献标识码:A

一种结合谱聚类与关联规则的轴承故障诊断方法

徐秀芳，徐丹妍，徐　森，郭乃瑄，许贺洋
（盐城工学院 信息工程学院，江苏 盐城　224051）

摘要：针对现阶段机械设备轴承故障诊断方法难以挖掘隐含特征、诊断精准度低等问题，将谱聚类（spectral clustering，SC）算法与关联规则算法 Apriori 相结合，提出 SC-Apriori 算法；首先根据美国西储大学轴承数据中心网站公开发布的轴承故障数据集，选取 0 负载下的数据，计算得到滚动轴承振动信号的 9 个时域特征和 3 个频域特征；其次使用 Pearson 相关系数进行特征筛选，留下 9 个有效特征，再利用 SC-Apriori 算法挖掘出训练数据集中轴承不同特征数据之间的关联关系，并引入提升度来去除冗余的关联规则，进而构建一个规则库；再将测试数据进行处理，并与已建立的规则库进行比对，根据匹配率来判断其故障类型；在测试数据上的实验结果表明，与已有算法相比，文章设计的 SC-Apriori 算法挖掘出的规则数量大幅减少，匹配速度更快，且匹配效果更好。

关键词：轴承故障诊断；数据挖掘；关联规则；谱聚类算法；提升度

Bearing Fault Diagnosis Method Combining Spectral Clustering and Association Rules

XU Xiufang, XU Danyan, XU Sen, GUO Naixuan, XU Heyang
(School of Information Engineering, Yancheng Institute of Technology, Yancheng　224051, China)

Abstract: Aimed at the problems of difficulty in mining implicit features and low diagnostic accuracy of existing mechanical equipment bearing fault diagnosis methods, a SC-Apriori algorithm was proposed by combining the spectral clustering (SC) algorithm with the association rule (Apriori) algorithm . Firstly, based on the bearing fault dataset publicly released on the website of Bearing Data Centre of Western Reserve University, the unload data were selected to calculate and obtain nine time-domain features and three frequency-domain features of the rolling bearing vibration signal. Secondly, the Pearson correlation coefficient was used to filter the features and reserve nine effective features, and then the SC-Apriori algorithm was used to mine the association relationship between the different features of bearings in the training dataset. and the boosting was introduced to remove the redundant association rules and construct a rule base. Then the test data were processed and compared with the established rule base to judge their fault types by the matching rate. Experimental results on the test data show that compared with existing algorithms, the SC-Apriori algorithm has the advantages of mining a significantly reduced number of rules, fast matching speed and better matching effect.

Keywords: bearing fault diagnosis; data mining; association rules; spectral clustering algorithm ; lift

0　引言

“工业 4.0”[1]和《中国制造 2025》[2]将信息技术与工业技术紧密融合，推动制造业的发展，以实现智能制造。其中，对机械设备健康状况以及故障的检测已被列为智能制造中的核心技术。轴承是机械设备中关键性基础零部件，其工作状况直接影响着机械设备的工作性能[3-4]。滚动轴承的任何异常，有可能导致相关零部件的工作状况发生变化，影响整个设备的正常运转，使其整体性能下降，严重时可能导致设备损坏，甚至出现安全事故[5]。因此，必须加强轴承的定期检测、维护和保养，提前发现异常，及时诊断可能发生的早期故障，可有效避免因轴承损坏而导致的设备停工，甚至涉及生命安全的事故，减少重大经济损失，避免人员伤亡。

由于滚动轴承故障早期阶段，局部缺陷和损伤较少，故障症状不太明显，检测得到的特征信号不强，信噪比低

收稿日期:2022-09-22； **修回日期**:2022-11-03。
基金项目:国家自然科学基金项目(62076215);江苏省高等学校自然科学研究面上项目(21KJD520006);2021 年度未来网络科研基金(FNSRFP-2021-YB-46);盐城工学院研究生培养创新工程项目(SJCX21_XZ018);横向项目合同编号(2022032809);教育部产学研合作项目(202102594034)。
作者简介:徐秀芳(1973-),女,江苏建湖人,硕士,高级实验师,主要从事机器学习、智能信息处理方向的研究。
徐　森 (1983-),男,江苏滨海人,博士,教授,硕士生导师,CCF 专业会员(No.14095M),主要从事模式识别与人工智能、机器学习方向的研究。
引用格式:徐秀芳,徐丹妍,徐　森,等.一种结合谱聚类与关联规则的轴承故障诊断方法[J].计算机测量与控制,2023,31(1):51-58.

等特点，滚动轴承的早期故障诊断成为国际、国内故障诊断领域的重要研究方向和挑战[6]。

滚动轴承的故障诊断技术发展分为频谱分析诊断法、冲击脉冲诊断法、共振解调诊断法、基于微机的滚动轴承工况检测 4 个阶段。随着计算机技术的快速进步，神经网络、聚类算法、支持向量机等以微机为核心的机器学习方法被广泛用于故障诊断[7]。

关联规则可有效挖掘数据集中各项之间的隐含关系[8]，找出不同设备测量值和故障之间的内在联系。文献［9］运用关联规则，分析一次风机各测量参数间的隐含关系，形成相应的关联规则库，根据设备运行数据与规则库匹配的结果，判断是否出现故障，实现故障预警。文献［10］将 K-means 与 Apriori 算法结合，获得用水量包括生活用水、工业用水、服务业用水、生态用水、农业用水和建筑业用水与供水之间的有效强关联规则，为深圳供水波动归因分析提供更好的依据。文献［11］通过 mRMR（最小冗余最大相关）对配电网多源数据进行特征选择，将 K-means 离散化后的取值用 FP-Growth 算法挖掘关联规则，由于规则库中的条件特征是各馈线及分支线上的电气量信息，所以能对发生故障的地点做出诊断。文献［12］提出一种利用多源故障信息进行故障诊断的方法，利用 Apriori 算法获得由有向二分图和贝叶斯算法得到的可疑部件的置信度，输入诊断模型后，根据确定性来判断目标是否是故障部件，提高了有源配电网设备故障诊断的准确性，提升维护效率。

本文研究了一种结合谱聚类的关联规则分析方法，针对轴承相关数据可以得到有效特征及其之间的强关联规则，根据匹配度来判断故障类型，并运用几种不同关联规则挖掘算法进行对比分析，验证该方法的优越性。

1 常见的轴承故障类型

轴承一般由内圈、外圈、滚动体和保持架四部分组成，通常内圈是固定在轴上的，它与轴径配合并与轴一起做旋转运动。外圈固定在轴承座上，对轴有支撑的作用。滚动体位于外圈和内圈中间，受内圈摩擦力驱动，做滚动运动，保持架一方面用于保证滚动体之间的相对距离，另一方面有效防止滚动体滑落。根据轴承结构的不同，轴承的局部故障类型可分为 4 种：内圈故障、外圈故障、滚动体故障和保持架故障。其产生故障的机理如图 1 所示。

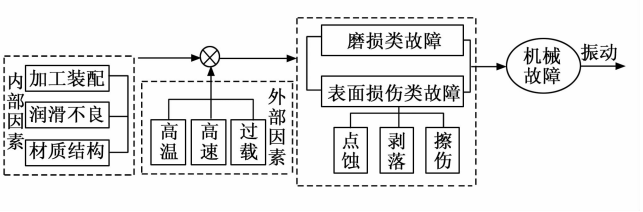

图 1 轴承故障产生机理图

轴承产生的故障有两种类型，一种是磨损类故障，另一种是表面损伤类故障。磨损类故障产生原因主要是缺少润滑油、轴承内部污染、轴电流和未对准等，包括轴承表面粗糙、不规则和大面积退化。表面损伤故障会产生特定的故障特征，一般出现在轴承局部位置，并且危害较大，一般包括点蚀、剥落、擦伤和裂纹等，主要是由于滚动体与滚道之间的接触疲劳引起的。由于磨损故障一般没有明确的故障特征信号，其危害相对较小，所以大多数研究关注危害更大的表面损伤故障[13]。

2 相关算法介绍

2.1 关联规则的 Apriori 算法

关联规则最原始的一种方法是穷举所有可能的规则，然后求出每一条规则的支持度和置信度，但这种方法开销很大。减少开销的方法是拆分支持度和置信度。由于规则支持度的大小取决于该规则先导和后继项集的支持度，一般的关联规则挖掘算法都会将其分成两个步骤：第一步，找到事务数据库中所有大于或等于预设的最小支持度阈值的频繁项集；第二步，利用频繁项集生成需要的关联规则，根据预设的最小置信度阈值进行取舍，最终生成强关联规则[14]。

Apriori 算法是一种典型的频繁项集挖掘方法，其核心思想是通过连接项集，构造出候选项和支持度，并通过剪枝产生频繁项集，从而获得最大频繁项集。在此基础上，利用产生的最大频繁项集和最小置信度阈值，得出一种较为可信的关联规则。该方法最初是用于超市销售数据库中，寻找同一用户购买不同商品之间的关联性[15]。现在关联规则挖掘已应用于医疗、金融、电商、交通等众多领域。

该算法的主要步骤：首先，采用递归法找出支持度大于预设的最小支持度的所有频繁项集；其次，利用频繁项集生成强关联规则，其置信度大于预设的最小置信度[16]。Apriori 算法的流程如图 2 所示。

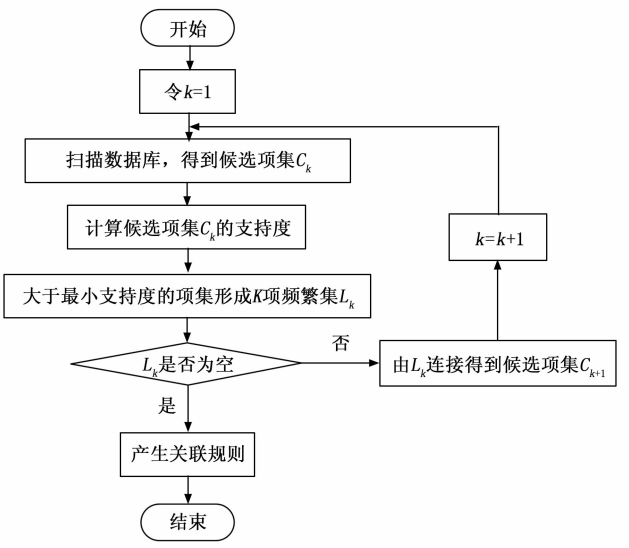

图 2 Apriori 算法流程图

假设 $I=\{I_1,I_2,\cdots,I_m\}$ 是项的集合，D 是一个事务数据库，事务 T 是 I 的非空子集，每个 T 都有一个唯一的标识符 TID 与之对应。项集是项的集合，包含 k 个项的项集

称为 k 项集。项集的出现频率是指包含项集的事务计数。如果项集 I 的支持度满足预设的最小支持度阈值，则 I 是频繁项集[17]。频繁 k 项集通常记作 L_k，是通过连接找出来的。任何频繁 k 项集都是由频繁 $k-1$ 项集组合生成的，频繁 k 项集的所有 $k-1$ 项子集一定都是频繁 $k-1$ 项集。候选项集通常记作 C_k，剪枝可以在产生候选项集的过程中减小搜索空间。

支持度表示某个项集出现的概率，是事件样本数与总样本数之间的比值，表示事件发生的概率；置信度表示关联规则的先导出现时后继也出现的概率，等价于条件概率。关联规则中的支持度和置信度的公式如下：

$$sup(X \Rightarrow Y) = p(X \cup Y) = \frac{\sigma(X \cup Y)}{N} \tag{1}$$

$$conf(X \Rightarrow Y) = p(Y \mid X) = \frac{\sigma(X \cup Y)}{\sigma(X)} \tag{2}$$

其中：$X \Rightarrow Y$ 是关联规则的一种蕴含式，X 是关联规则的先导项集，Y 是关联规则的后继项集，$X \cup Y$ 表示 X、Y 的并集，N 表示总事务个数，σ 表示计数。

当 $sup(X \Rightarrow Y) \geqslant min\_sup$ 且 $conf(X \Rightarrow Y) \geqslant min\_conf$ 时，则提取出的备选关联规则为强关联规则。其中 min_sup 是最小支持度，是自定义用来衡量支持度的一个阈值，表示项集的最低重要度；min_conf 是最小置信度，是自定义用来衡量置信度的一个阈值，表示项集的最低可靠性。

挖掘关联规则就是在已知的事务数据库 D 中，找到符合最小支持度与最小置信度阈值的关联规则。比如轴承在运行状态中发生了故障，记为 Y，故障现象可能是由于时域、频域特征值随时间变化逐渐偏离正常值引起的。故障征兆可以记为：X_1，X_2，…，X_M，规则（X_1，X_2，…，$X_{M-1} \Rightarrow X_M$）表示征兆之间的关联。通过规则匹配可以推断出故障 Y 的规则。

2.2　改进的 SC-Apriori 算法

关联规则是目前数据挖掘中应用最为广泛的一种研究方法。通常采集到的轴承振动数据属于数值型数据，为找出此类数据之间的关联，需要进行离散化处理。聚类算法可以有效地对连续型数据进行分类，保证不同类间数据相似性低，类内数据相似度高[18]。

K-means 算法是一种经典的基于梯度的聚类方法[19]，首先把 N 个对象划分为 k 个簇，然后用簇中对象的平均值代表每个簇的质心，并进行迭代直到簇内的对象不再改变，使得簇内对象具有较高的相似性，簇间具有较低的相似性[20]。

谱聚类（spectral clustering，SC）[21-23]是一种以图论为基础的聚类算法，其核心思想是把所有数据都看成是空间上的点，以全连接法用边将它们连起来；两个相隔较远的点，其边权重值较小，两个邻近的点，其边权重值较高，将各数据点组成的图进行划分，使得各子图间的边权重之和尽量小，各子图中的边权重之和尽量大，以实现聚类。其具体步骤为：首先构建 $n * n$ 的邻接矩阵 $\boldsymbol{A}$（$\boldsymbol{A}$ 的对角元素设为 0）；再构建拉普拉斯矩阵 $\boldsymbol{L} = D^{-1/2}AD^{-1/2}$；然后根据 $\boldsymbol{L}$ 的前 K 个最大特征值对应的特征向量 p_1，…，p_k，构建矩阵 $\boldsymbol{X} = [p_1, \cdots, p_k]$，对 $\boldsymbol{X}$ 的行向量进行规范化处理，使向量的欧式范数为 1，得到矩阵 $\boldsymbol{Y} = [y_1, \cdots, y_k]$，设 Y 的每一行为一个 K 维向量，得到数据集 $Z = [z_1, \cdots, z_n]$；最后使用 K-means 算法对 Z 进行聚类，生成 K 个簇[24]。与 K-means 算法相比，谱聚类方法只需将待聚类的不同点之间的相似性矩阵用于聚类，即可展现较好的聚类效果。

为了解决支持度和置信度无法过滤掉一些无用的强关联规则，导致产生过多的规则，使得匹配时间过长的问题，本文引入提升度，来优化强关联规则的判断框架。获取一条真正有效的强关联规则的评价标准与支持度、可信度和提升度均相关，如式（3）所示：

$$X \Rightarrow Y[Support, Confidence, Lift] \tag{3}$$

对于一条规则 $X \Rightarrow Y$，提升度表示 X 条件下，同时 Y 也出现的概率，与 Y 总体出现的概率之比。其计算方式如式（4）所示：

$$lift(X \Rightarrow Y) = \frac{p(Y \mid X)}{P(Y)} = \frac{conf(X \Rightarrow Y)}{sup(Y)} \tag{4}$$

如果 $lift(X \Rightarrow Y) < 1$，则 X 和 Y 是负相关，代表一个出现可能导致另一个不出现；如果 $lift(X \Rightarrow Y) > 1$，则 X 和 Y 是正相关的，代表一个出现，另一个也会同时出现；如果 $lift(X \Rightarrow Y) = 1$，则 X 和 Y 是独立的，它们之间没有相关性。提升度越高，说明关联度越强，提升度越低，说明关联度越小。

2.3　SC-Apriori 算法流程

结合谱聚类的关联规则分析方法，可以得到有效特征及其之间的强关联规则。其主要步骤如下：

1）将提取的特征值用谱聚类离散化。

2）扫描事务数据库 D，令 $k=1$，进行计数，产生候选 1 项集，表示为 C_1。

3）根据最小支持度，由 C_1 产生频繁 1 项集表示为 L_1。

4）若 $k>1$，重复 5）、6）和 7）步骤。

5）由 L_k 执行连接和剪枝操作，产生候选 $k+1$ 项集 C_{k+1}。

6）根据最小支持度，由 C_{k+1} 产生频繁 $k+1$ 项集 L_{k+1}。

7）若频繁项集 $L \neq \phi$，则 $k=k+1$，跳往步骤 5）；否则跳往步骤 8）。

8）根据最小置信度和最小提升度，由频繁项集产生强关联规则。

根据挖掘出的关联规则，就可以进行故障诊断。

3　基于 SC-Apriori 算法的轴承故障诊断

3.1　实施方案及流程

结合谱聚类与关联规则的轴承故障诊断实施方案及流程如图 3 所示。首先将收集到的轴承故障数据，经过一系列的预处理分成训练数据和测试数据。将训练数据输入 SC-Apriori 模型进行训练，输出符合要求的关联规则，并形成

故障关联规则库。

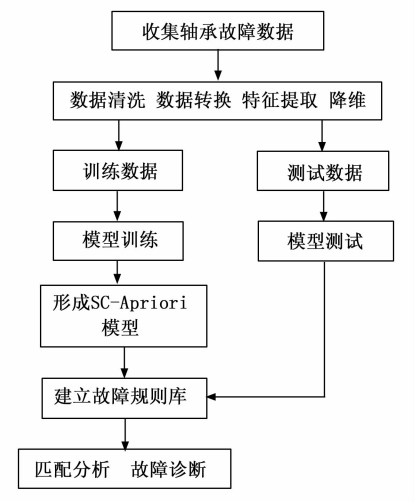

图 3 实施方案及流程

用测试数据进行模型测试，将生成的关联规则与规则库进行匹配分析，并通过匹配度来确定具体的故障类型。

具体步骤为：

1）首先，对原始数据进行信号零均值化处理；其次，根据时域、频域公式提取对应的特征值，并使用相关系数进行筛选；最后，将筛选后得到的 9 个特征值（方差、均方根、峰值、峰值因子、峭度系数、波形因子、裕度因子、均方频率、重心频率）归一化处理。

2）将得到的有效特征数据，运用谱聚类算法进行离散化处理，一一映射到各个区间，形成待挖掘数据库。

3）通过选取合适的参数（最小支持度 min _ sup、最小置信度 min _ conf 和最小提升度 min _ lift），并把 SC-Apriori 算法挖掘出的规则组成故障规则库。

4）选取待检测的样本数据进行预处理，将得到的各个特征数据，按照已经划分好的区间进行标记，运用 SC-Apriori 算法找出待测样本数据相应的关联规则，并通过匹配率[25]来确定具体的故障类型。关联规则匹配率是在不同阈值下挖掘出的关联规则与标准关联规则的相同率[24]。

3.2 特征提取

对轴承数据进行分析处理，可以有效地减少误 差，获取更多的故障信息。本文选取 4 个时域有量纲特征、5 个时域无量纲指标、3 个频域指标进行分析，其公式按上述顺序排列如表 1 所示。

由于各个特征值的幅值大小不一，不便于比较同一特征值的不同样本之间的差异，所以要将特征数据进行归一化处理。归一化处理是为了消除量纲，使得指标之间具有可比性；将数据限制到一定区间，使得运算更为便捷。原始数据经过数据标准化处理后，各指标处于同一数量级，适合进行综合对比评价。归一化的公式如下所示。

$$x = \frac{x - min}{max - min} \tag{5}$$

其中：max 为样本数据的最大值，min 为样本数据的最小值。

表 1 时域、频域特征公式表

序号	名称	公式
1	均值	$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i$
2	方差	$\sigma^2 = \sum_{i=1}^{N}(x_i - \overline{X})$
3	均方根值	$X_{\text{rms}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$
4	峰值	$X_{\text{peak}} = \frac{1}{2}(\max(x_i) + \min(x_i))$
5	峰值因子	$C = \frac{X_{\text{peak}}}{X_{\text{rms}}}$
6	峭度系数	$K = \sum_{i=1}^{N} \frac{x_i^4}{NX_{\text{rms}}}$
7	波形因子	$S = \frac{NX_{\text{rms}}}{\sum_{i=1}^{N} \lvert x_i \rvert}$
8	脉冲因子	$I = \frac{NX_{\text{peak}}}{\sum_{i=1}^{N} \lvert x_i \rvert}$
9	裕度因子	$CL = \frac{X_{\text{peak}}}{(\frac{1}{N}\sum_{i=1}^{N} \sqrt{\lvert x_i \rvert})^2}$
10	重心频率	$FC = \frac{\sum_{i=1}^{N} f_i p_i}{\sum_{i=1}^{N} p_i}$
11	频率方差	$VF = \frac{\sum_{i=1}^{N} (f_i - fc)^2 p_i}{\sum_{i=1}^{N} p_i}$
12	均方频率	$MSF = \frac{\sum_{i=1}^{N} f_i^2 p_i}{\sum_{i=1}^{N} p_i}$

其中：N 为样本数量，X 为数据集，x 为数据集 X 中的数据，max 表示最大值，min 表示最小值，f 是频率，p 是信号的功率谱，fc 是重心频率。

然后用相关系数进行相关性分析，进行特征筛选。利用 Pearson 相关系数[26]计算每两个故障特征之间的相关性，设定相关性阈值，选择相关性高于阈值的故障特征。

Pearson 相关系数的计算公式为 $r_{X,Y} = \frac{Cov(X,Y)}{S_X S_Y}$，其中 $Cov(X,Y)$ 是样本协方差，S_X 是 X 的样本标准差，S_Y 是 Y 的样本标准差，$\overline{X}$ 是样本均值，X、Y 表示两个项集，n 表示样本总数，公式如（6）～（8）所示。

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1} \tag{6}$$

$$S_X = \sqrt{\frac{\sum_{i=1}^{n}(X_i - X)^2}{n-1}} \tag{7}$$

$$S_Y=\sqrt{\frac{\sum_{i=1}^{n}(Y_i-Y)^2}{n-1}} \tag{8}$$

3.3 区间划分方法

Apriori 算法是一种先验算法，只能识别热编码，因此，有必要进行数据离散化，成为 Apriori 算法可以识别的数据结构。数据离散化是将数据的值范围划分为离散区间。采用谱聚类方法，可以将连续的参数离散到若干区间，达到数据离散化的目的。例如，对方差数据进行离散化，设定划分为 3 个簇。其划分过程依次为：把方差的所有数据值作为目标样本集输入，生成相似矩阵 $\boldsymbol{S}$；根据 $\boldsymbol{S}$ 构建邻接矩阵 $\boldsymbol{A}$ 和对角度矩阵 $\boldsymbol{D}$，得到对应的拉普拉斯矩阵 $\boldsymbol{L}$，并将其标准化；计算最小的特征值所对应的特征向量；将特征向量标准化形成特征矩阵 $\boldsymbol{F}$，用 K-means 进行聚类，得到簇划分。

首先，将同一参数的数据值聚成不同的类，从每个类中取出最小值和最大值，最小值设为区间的左端，最大值设为区间的右端。然后，将该参数的所有取值划分到相应区间，并应用 Apriori 算法对离散化后的数据进行相关处理以获得关联规则。

3.4 故障诊断

应用 SC-Apriori 算法得到的关联规则通常无法直接进行故障诊断，还需要通过计算匹配率来进行评估。在进行故障诊断时，将测试数据集挖掘出的关联规则与规则库中的各种故障类型的规则相匹配，根据不同故障类型的规则，得到匹配的规则数量并计算其匹配率，对比得出故障诊断的结论。

4 实验及结果分析

4.1 数据来源

采用凯斯西储大学（case western reserve university，CWRU）轴承数据中心的轴承故障数据集[27]，实验中试验台的实际结构，由一个 2 马力的电动机、一个扭矩传感器/译码器和一个测功器组成。实验中使用加速度传感器采集振动信号，通过使用磁性底座将传感器放在电机壳上。轴承信号采集试验台如图 4 所示。

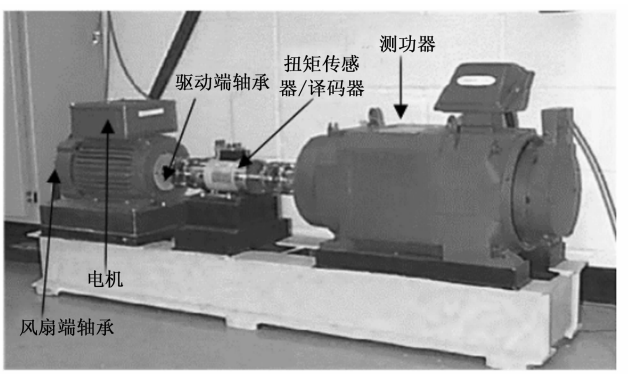

图 4　CWRU 轴承信号采集试验台

实验为人工采用电火花在轴承的内圈、外圈和滚动体上模拟加工出尺寸为 0.007 英寸、0.014 英寸、0.021 英寸的单点故障，用来表示不同的故障损坏程度，实验共有 9 种不同的损伤状态和 1 种健康状态[28]。

实验所需轴承参数，如表 2、表 3 所示。

表 2　实验所需参数

参数	数值
主轴转速	1 797 r/min
采样频率	12 kHz
采样长度	20 480 点
轴承型号	6 205-2RS JEM SKF

表 3　实验轴承参数

轴承状态	故障直径/mm	电机载荷/HP
正常	无	0
内圈故障	0.007/0.014/0.021	0
外圈故障	0.007/0.014/0.021	0
滚动体故障	0.007/0.014/0.021	0

实验中，样本长度设定为 1 024，并按照 3：1 划分训练集与测试集，数据集组成如表 4 所示。

表 4　实验轴承数据集

标签	轴承状态	故障直径/英直径寸	训练集数量	测试集数量
ir007	内圈轻微故障	0.007	88	30
ir014	内圈中度故障	0.014	88	30
ir021	外圈重度故障	0.021	89	30
or007	外圈轻微故障	0.007	89	30
or014	外圈中度故障	0.014	88	30
or021	内圈重度故障	0.021	89	30
b007	滚动体轻微故障	0.007	89	30
b014	滚动体中度故障	0.014	88	30
b021	滚动体重度故障	0.021	89	30

4.2 轴承特征提取

信号处理方面，选取零负载、电机转速近似为 1 797 圈/每分钟的故障数据，提取时域特征和频域特征。将每一类的故障数据以 1 024 个数据为一组进行分组，其中 75%的数据用于训练，其余的数据用于测试。从图 5 可以看出，零均值处理后可以消除频率在 0 处出现的大谱峰，去除其对周围小峰值产生的影响，便于频域分析。

正常轴承和发生不同故障轴承的时域图如图 6 所示，对比分析轴承在正常、内圈故障、外圈故障和滚动体故障不同状态下，其振动信号变化显著，波形分布和幅度随故障位置及故障尺寸的变化而变化，正常轴承的波动幅值在 −0.2 到 0.2 之间波动，故障轴承的波动幅值明显大于正常轴承的波动幅值，其中轴承内圈轻微故障的波动幅值在 −1 到 2 之间，轴承滚动体轻微故障的波动幅值在 −0.5 到 0.5 之间，轴承外圈轻微故障的波动幅值在 −2 到 2 之间，发生不同故障的轴承也能从时域图中区分出来。

然后，按照表 1 所示时域、频域特征提取公式，将 12 个特征提取出来，并进行归一化处理。

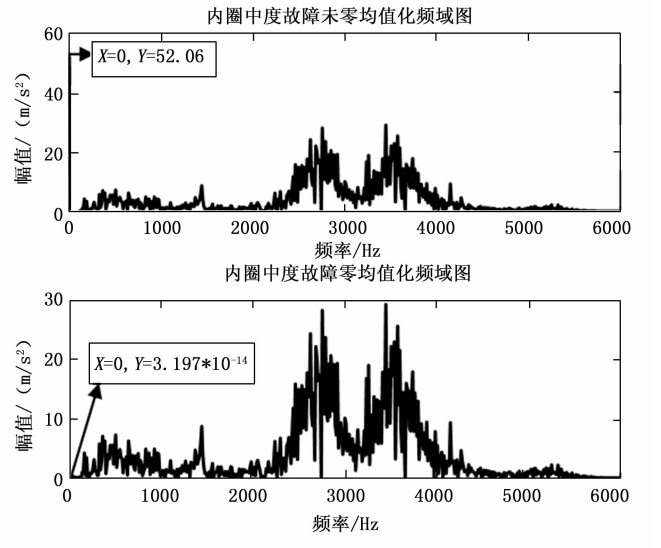

图 5 零均值化处理图

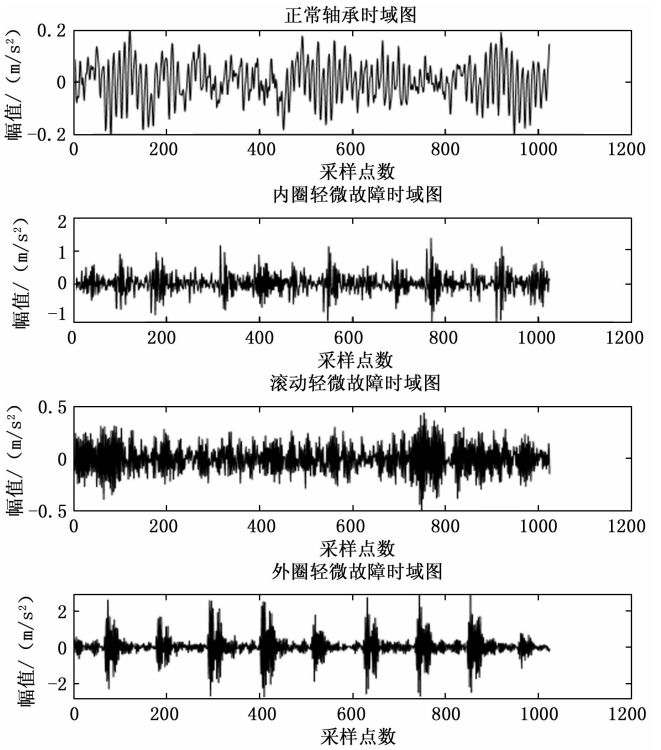

图 6 不同状态下的轴承时域对比图

一般情况下，数值在｜0～0.3｜代表特征之间是弱相关的，数值在｜0.3～0.6｜代表特征之间是相关的，数值在｜0.6～1｜代表特征之间是强相关的。去除相关系数低的均值、脉冲因子和频率方差后，给保留下来相关系数值均大于 0.3 的 9 个特征编号。编号依次为：方差：10；均方根值：20；峰值：30；峰值因子：40；峭度系数：50；波形因子：60；裕度因子：70；均方根值：80；重心频率：90。

4.3 区间划分过程

对 9 个特征值进行划分，为便于表示，将每个参数的区间按照 1 到 3 从小到大编号，部分离散化区间如表 5 所示。

表 5 部分特征值离散化区间

编号	区间 1	区间 2	区间 3
10	[0,0.054]	[0.054,0.45]	[0.45,1]
20	[0,0.124]	[0.124,0.39]	[0.39,1]
30	[0,0.096]	[0.096,0.343]	[0.343,1]
40	[0,0.161]	[0.161,0.437]	[0.437,1]
50	[0,0.062]	[0.063,0.283]	[0.283,1]
……	……	……	……

当一个特征数据落入某一区间时，即用该特征的编号加上区间编号表示。离散化前的部分数据如表 6 所示。例如，现将方差的 3 个区间按 1～3 进行编号，将方差中的数据按区间划分大小映射进相应的区间中，表 5 的第二行数据为方差的具体数值，该行第一列的数据值为 0.07，其值落在表 4 中方差（特征编号 10）对应的区间 2，所以离散化后的编号为 102。

表 6 离散化前的部分数据

编号	10	20	30	40	50	……
数据	0.07	0.20	0.15	0.35	0.21	……
	0.13	0.31	0.27	0.51	0.29	……
	0.02	0.09	0.05	0.18	0.05	……
	0.07	0.21	0.11	0.21	0.07	……

用这种方式对所有数据编号就得到了离散化后的数据，部分离散化后的数据如表 7 所示。

表 7 离散化后的部分数据

编号	10	20	30	40	50	……
数据	102	202	302	402	502	……
	102	202	302	403	503	……
	101	201	301	402	501	……
	102	202	302	403	502	……

4.4 故障诊断结果分析

为使挖掘出的规则能准确表达各个特征频率之间的关系，选取最小支持度 min _ sup、最小置信度 min _ conf 的参数值就显得尤为关键。如果最小支持度阈值设置得太高，虽然可以减少数据挖掘过程中频繁项的计算时间，但很容易导致隐藏在数据中的一些重要频繁项集被过滤掉。由于置信度需要在支持度之后进行计算，所以最小支持度阈值应尽可能小。如果最小置信度阈值设置得太低，可能会生成大量弱关联度，导致计算负载过高，大大增加了数据挖掘的时间。由于提升度的计算需要一定数量的关联规则，因此不必将最小置信度阈值设置得很高，从而保证生成更多的规则。

图 7 为使用轴承故障数据集，将改进后的 SC-Apriori 算法在不同支持度和置信度下进行对比，分析产生规则数随支持度、置信度变化的情况。

从图 7 可以看出，在置信度不变的情况下，随着最小

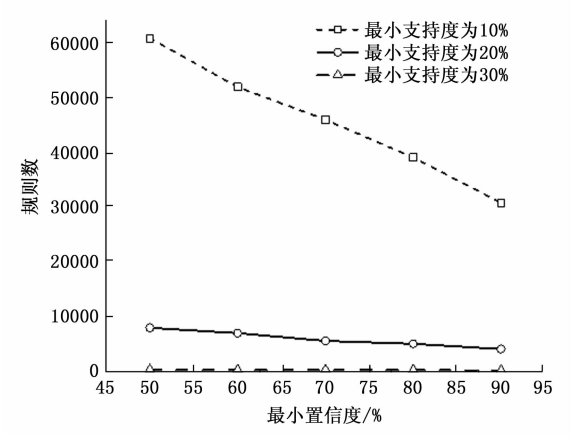

图 7　不同置信度下的支持度曲线

支持度的增大，挖掘出的规则数量在不断减少；在支持度不变的情况下，随着最小置信度的增大，挖掘出的规则数也在不断减少。规则数量太多，其本身数据挖掘就耗费了过多的时间，得出的规则可信度也不高，在故障匹配时，花费时间会很长；规则数量太少时，会导致故障匹配失败，匹配率过低，所以要选取合适的参数值。例如当最小支持度为 10%时，产生的规则数量太多，而最小支持度为 30%时，产生的规则数量又过少，为此，本文选用 $min\_sup=20\%$、$min\_conf=80\%$。

分别使用 $min\_lift>1$ 和 $min\_lift\geqslant 1$ 在 $min\_sup=20\%$、$min_conf=80\%$的情况下进行测试，图 6 规则数对比表中，ir007 表示内圈轻微故障、ir014 表示内圈中度故障、ir021 表示内圈重度故障、or007 表示外圈轻微故障、or014 表示外圈中度故障、or021 表示外圈重度故障。由图 8 可见，$min\_lift>1$ 时，ir007（内圈轻微故障）产生的规则数为 0，无法进行关联规则匹配，所以无法选用 $min\_lift>1$。故障类型为 ir014（内圈中度故障）时，原规则数为 31270，$min\_lift>1$ 时的规则数为 12 258，$min\_lift\geqslant 1$ 时的规则数为 14 664，引入提升度后的关联规则数量大大减少。所以选用 $min\_lift\geqslant 1$，此时提升度仍然可以过滤掉一部分规则。

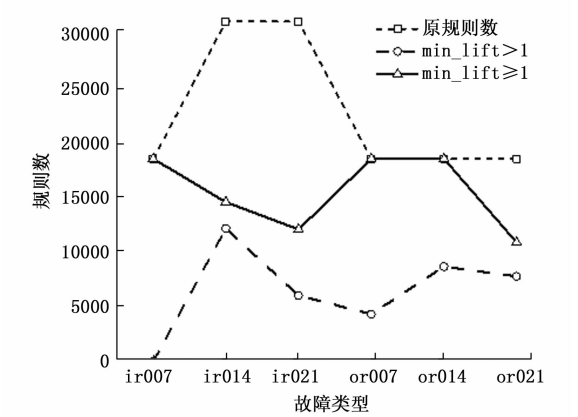

图 8　规则数对比表

在 $min\_sup=20\%$、$min_conf=80\%$和 $min\_lift\geqslant 1$ 的条件下，进行关联规则挖掘，得出的部分关联规则如表 8 所示。

表 8　部分关联规则

关联规则	支持度	置信度	提升度
103⇒203	0.625	1	1.023 3
[101,203]⇒702	0.352 3	1	1.035 3
[403,601,801]⇒[702,903]	0.909 1	0.975 6	1.059 9
[103,203]⇒[403,601,702]	0.602 3	0.963 6	1.009 5

表中第一条规则“103⇒203”表示，方差取值落在第三个区间 103 时，均值方根的取值会落在第三个区间 203，并且这条规则的支持度是 0.625、置信度是 1、提升度是 1.023 3。表中第二条规则“［101，203］⇒702”表示，方差取值落在第一个区间 101 且均值方根取值落在第三个区间时，裕度因子的取值会落在第二个区间 702，并且这条规则的支持度是 0.352 3、置信度是 1、提升度是 1.035 3。

匹配率的表达式如下：

$$m=\frac{k_1}{k_2}*100\% \tag{9}$$

其中：m 表示匹配率，指所建立规则库对被测试样本数据的基本适用性，k_1 是与被测试样本数据与原数据库中相应故障相匹配的关联规则数量，k_2 是与被测样本数据相匹配的关联规则总数。

为进一步验证 SC-Apriori 算法的效果，现将 SC-Apriori 算法、文献［10］使用的 Apriori 算法、文献［11］使用的 FP-Growth 算法，在相同的支持度、置信度等条件下进行仿真分析，比较 3 种算法对应故障类型匹配率和运行时间，如表 9、表 10 所示。

表 9　3 种故障诊断方法的匹配率

故障类型	故障匹配率/%		
	文献[10]	文献[11]	本文方法
内圈轻微故障	78.69	78.69	94.50
内圈中度故障	93.60	93.74	96.79
内圈重度故障	71.97	71.97	78.08
外圈轻微故障	73.19	73.19	95.11
外圈中度故障	54.81	54.81	70.44
外圈重度故障	86.15	86.16	86.92

表 10　3 种故障诊断方法的时间

故障类型	时间/min		
	文献[10]	文献[11]	本文方法
内圈轻微故障	364.66	468.99	207.22
内圈中度故障	249.91	241.07	132.56
内圈重度故障	264.60	262.97	151.85
外圈轻微故障	272.46	236.76	196.11
外圈中度故障	378.93	452.92	162.52
外圈重度故障	242.80	239.78	197.53

由表 9 可见，使用文献［10］和文献［11］得出的故障匹配率差别不大。针对轴承内圈轻微故障，运用文献［10］

和文献［11］的算法，得到的故障匹配率均为 78.69%，运用本文的 SC-Apriori 算法得到的匹配率为 94.5%，匹配率提升了 15.81%；针对轴承外圈轻微故障，运用文献［10］和文献［11］的算法，得到的匹配率均为 73.19%，运用本文的 SC-Apriori 算法得到的匹配率为 95.11%，匹配率提升了 21.92%。根据表 9 中 3 种故障诊断算法效果对比分析可见，使用 SC-Apriori 算法的故障匹配率明显要优于 Apriori 算法和 FP-Growth 算法，算法的优越性得到了验证。

表 10 是对 3 种算法进行故障诊断分析占用运行时间的对比。由于文献［10］和文献［11］在与本文相同条件下，挖掘出的规则数更多，所以在规则匹配时，花费的时间更多；本文使用的 SC-Apriori 算法，在经过提升度的筛选后，规则数量大大减少，得到故障诊断结果所需要的运行时间明显减少。在相同的条件下，本文使用的 SC-Apriori 算法可以更快的得出故障诊断结果。

5 结束语

本文研究了结合谱聚类与关联规则的 SC-Apriori 算法在轴承故障诊断中的应用，将谱聚类与 Apriori 算法相结合，提高了分析方法的可靠性，通过引入提升度，提出的 SC-Apriori 算法可以更好地反映轴承故障特征的关联关系，减少相对无用的规则数量，缩短规则匹配时间，实验结果最高能达到 96.79%的匹配率。通过和其他关联规则算法对比，验证了 SC-Apriori 算法在故障匹配率的优越性。《中国制造 2025》的制高点、突破口和主攻方向是制造业数字化、网络化、智能化[29]，因此，轴承故障智能诊断具有良好的发展前景。

参考文献：

［1］谢 峰，吕玉琳，雷小宝，等. 面向机械制造业的工业 4.0 技术实验平台的开发［J］. 实验技术与管理，2020，37（6）：207－210.

［2］国务院. 国务院关于印发《中国制造 2025》的通知［EB/OL］. http://www.gov.cn/zhengce/content/2015-05/19/content_9784.htm.

［3］周 济，李培根，周艳红，等. 走向新一代智能制造［J］. Engineering，2018，4（1）：28－47.

［4］刘兴建，原振文. Spark 平台环境下基于 Aco-kmeans 算法的滚轴故障检测算法研究［J］. 计算机应用与软件，2021，38（1）：256－261.

［5］郭德峰，吴鲁滨. 循环双谱的简化及在滚动轴承故障诊断的应用［J］. 机电产品开发与创新，2019，32（3）：76－78，100.

［6］张言伟. 圆柱滚子轴承速度型振动信号分析系统［D］. 洛阳：河南科技大学，2017.

［7］翟嘉琪，杨希祥，程玉强，等. 机器学习在故障检测与诊断领域应用综述［J］. 计算机测量与控制，2021，29（3）：1－9.

［8］陈 佩，李凤华，李子孚，等. 基于规则关联的安全数据采集策略生成［J］. 网络与信息安全学报，2021，7（5）：132－148.

［9］高 瑜，仝卫国. 基于关联规则的一次风机故障预警方法研究［J］. 电力科学与工程，2016，32（10）：42－46.

［10］LIU X，SANG X F，CHANG J X，et al. The water supply association analysis method in Shenzhen based on kmeans clustering discretization and apriori algorithm.［J］. PloS one，2021，16（8）：e0255684.

［11］刘思怡，苏 运，张 焰. 基于 FP-Growth 算法的 10 kV 配电网分支线断线故障诊断与定位方法［J］. 电网技术，2019，43（12）：4575－4582.

［12］SHI F F，GENG Y J，LI Y D，et al. A Method of Fault Diagnosis for Secondary System Based on Multi-Source Fault Information［C］// 2020 IEEE 4th Conference on Energy Internet and Energy System Integration（EI2）. IEEE，2020.

［13］安 晶，周临震，安 鹏. 基于人工智能的故障诊断方法［M］. 北京：高等教育出版社，2022.

［14］李华群. 基于改进 Apriori 算法在图书馆数据挖掘中应用分析［J］. 内蒙古科技与经济，2021（24）：66－68，73.

［15］张良均，谭立云，刘名军，等. PYTHON 数据分析与挖掘实战（第 2 版）［M］. 北京：机械工业出版社，2020.

［16］沈慧娟，曹晓丽. 基于频集的 Apriori 关联规则算法的应用研究［J］. 物联网技术，2020，10（10）：57－61.

［17］赵仁朋. 基于位组合的大豆启动子频繁项集挖掘算法的研究［D］. 哈尔滨：黑龙江大学，2019.

［18］邵长龙，孙统风，丁世飞. 基于信息熵加权的聚类集成算法［J］. 南京大学学报（自然科学），2021，57（2）：189－196.

［19］徐秀芳，徐 森，花小朋，等. 一种基于 t-分布随机近邻嵌入的文本聚类方法［J］. 南京大学学报（自然科学），2019，55（2）：264－271.

［20］郭永坤，章新友，刘莉萍，等. 优化初始聚类中心的 K-means 聚类算法［J］. 计算机工程与应用，2020，56（15）：172－178.

［21］ROY ARNAB KUMAR，BASU TANMAY. Postimpact similarity：a similarity measure for effective grouping of unlabelled text using spectral clustering［J］. Knowledge and Information Systems，2022，64（3）：723－742.

［22］陈 迪，刘惊雷. 基于乘法更新规则的 K-means 与谱聚类的联合学习［J］. 南京大学学报（自然科学），2021，57（2）：177－188.

［23］徐 森，皋 军，花小朋，等. 一种改进的自适应聚类集成选择方法［J］. 自动化学报，2018，44（11）：2103－2112.

［24］徐 森. 面向海量文本的聚类集成技术研究［M］. 成都：四川大学出版社，2022.

［25］陈碧云，丁 晋，陈绍南. 基于关联规则挖掘的电力生产安全事故事件关键诱因筛选［J］. 电力自动化设备，2018，38（4）：68－74.

［26］肖 杨，李 亚，王海瑞，等. 基于皮尔逊相关系数的滚动轴承混合域特征选择方法［J］. 化工自动化及仪表，2022，49（3）：308－315.

［27］CASE WESTERN RESERVE UNIVERSITY. Bearing Data Center［EB/OL］.［2020－06－01］. http://csegroups.case.edu/bearingdatacenter/home.

［28］胡向东，梁 川. 基于 SE-ResNeXt 的滚动轴承故障诊断方法［J］. 计算机测量与控制，2021，29（7）：46－51.

［29］李传鑫，刘增力. 时频分析与 VGG19 迁移学习的轴承故障检测［J］. 电子测量技术，2021，44（5）：161－165.