文章编号: 1001-0920(2019)04-0767-08 DOI: 10.13195/j.kzyjc.2017.1299

基于信度区间的故障特征约简方法

徐晓滨[1†], 张 明[1], 文成林[1], 韩德强[2], 黄大荣[3]

(1. 杭州电子科技大学 自动化学院，杭州 310018；2. 西安交通大学 电子与信息工程学院, 西安 710049；
3. 重庆交通大学 信息科学与工程学院, 重庆 400074)

摘 要: 多源信息融合故障诊断方法可以有效提高设备故障的确诊率,但同时需要使用由不同传感器获取的多种故障特征数据. 此时若将所有特征的数据用于诊断,则计算量过大,诊断的实时性差. 对此,将证据理论与粗糙集相结合,提出基于信度区间的属性约简定理及相应的故障特征(属性)约简方法,力图利用约简后的重要特征进行快速诊断. 利用随机模糊变量和 K 均值对特征数据进行离散化处理,通过压缩二进制矩阵获取核属性,再将属性的信度区间大小作为迭代约简过程中属性的选取标准,向核属性中添加重要属性,最终获得属性约简结果. 最后进行电机转子的特征融合诊断实验,通过与经典的粗糙集简约方法对比验证所提出方法的有效性.

关键词: 故障诊断；报警监测；信度区间；属性约简；粗糙集；证据理论

中图分类号: TP391 **文献标志码:** A

Fault feature reduction based on belief interval

XU Xiao-bin[1†], *ZHANG Ming*[1], *WEN Cheng-lin*[1], *HAN De-qiang*[2], *HUANG Da-rong*[3]

(1. College of Automation，Hangzhou Dianzi University，Hangzhou 310018，China；2. College of Electronic and Information Engineering，Xi'an Jiaotong University，Xi'an 710049，China；3. College of Information Science and Engineering，Chongqing Jiaotong University，Chongqing 400074，China)

Abstract: Fault diagnosis based on multi-source information fusion can improve equipment diagnosis rate. However, such methods need to obtain a variety of fault feature data obtained from different sensors. If the data of all features are used for diagnosis, then computation burden will be too large to realize real-time diagnosis. Based on evidence theory and rough sets, a belief interval-based attribute reduction theorem and the corresponding fault feature (attribute) reduction method are presented so as to use the reduced important features to make fast diagnosis decision. In detail, the random fuzzy variable and K-means are used to discretize the data of features. The core attribute set can be achieved by compressed binary matrix. Thus, the belief interval size of attribute is taken as the criterion of attribute selection in the iterative reduction process in which the important attributes are added to the core. Finally, the reduction result is obtained. In the experiment of features fusion diagnosis of motor rotor, the effectiveness of the proposed method is illustrated by comparing with the classical rough set reduction methods.

Keywords: fault diagnosis；alarm monitoring；belief interval；attribute reduction；rough sets；evidence theory

0 引 言

故障诊断技术是工业设备稳定运行的重要保障. 随着传感器技术的发展以及计算机存储能力的提升,大量的监测数据可以通过设备上的多传感器系统获取,此时利用多源信息融合方法可以对多种故障特征的数据进行综合处理,给出更为可靠的诊断结果,并可以有效提高故障的确诊率. 但是融合过程往往计算量大,影响了诊断的实时性. 针对该问题可以考虑对故障特征(属性)进行适当的约简,利用重要的特征数据进行诊断,从而在保证较高确诊率的同时,兼顾诊断的实时性.

目前,属性简约或数据降维的主要方法是主成份分析(PCA)和粗糙集方法. PCA对相关属性进行线性变化生成新属性,根据贡献率大小对新生成的属性

收稿日期: 2017-09-30；**修回日期:** 2018-02-02.

基金项目: 国家自然科学基金项目(61433001, U1709215, 61573275, 61573076, U1509203)；浙江省科学技术厅公益技术应用研究项目；铁路轨道不平顺故障在线检测系统的研究与开发(2016C31071)；浙江省大学生科技创新活动计划项目(2017R407064).

责任编委: 阳春华.

作者简介: 徐晓滨(1980−), 男, 教授, 博士, 从事智能信息融合与复杂系统可靠性评估、故障诊断与预测等研究；张明(1991−), 男, 硕士生, 从事基于证据理论的故障诊断的研究.

†通讯作者. E-mail: xuxiaobin1980@163.com.

进行排序并选定其中的重要属性. 由于采集的属性数据常具有非线性特性,通过PCA得到的降维结果难以反映出属性中的非线性特性,并且新属性的完整性不及原始属性[1]. 粗糙集是在保持信息系统分类能力的情况下,剔除冗余属性与数据,它直接对原始数据进行分析与推理,并从中发现隐含的知识,相比于PCA更能够保持属性的完整性.

粗糙集理论能够有效地处理非精确、不确定以及不完备的数据[2-3],而属性约简是粗糙集中最重要的研究内容,其过程为:计算核属性集合,采用启发式算法向核属性集合中依次添加剩余属性,直至满足终止条件,得到属性约简结果. 其中,启发式算法通常是利用基于属性重要性和属性频率的策略得以实现的. 文献[4]提出了一种基于属性重要性的启发式约简算法,并利用差别矩阵求得核属性,将重要性最大的属性加入核属性集中,获得约简结果;文献[5-6]指出,当决策信息系统不一致时,利用文献[4]中的差别矩阵获取的核属性并不准确,并进一步采用新的差别矩阵处理系统不一致性,提出了改进型约简方法,以提高约简效率;文献[7-8]提出了另一种改进的差别矩阵,通过区分不一致对象与一致对象,分别建立差别矩阵以求得核属性,然后利用逻辑运算对分辨函数进行简化,以形成最小析取范式,获得约简结果.

Dempster-Shafer证据理论是另一种处理不确定信息的方法,它通过定义基本信度赋值函数(证据)、置信函数以及似真函数来度量信息的不确定性,并给出证据组合规则,利用这些规则融合多个信息源提供的证据,实现融合决策,形成一套完备的不确定性推理理论和方法[9]. 粗糙集和证据理论在处理非精确和不确定信息时,其研究的机理和方法各异,但具有相容性. 粗糙集用基于知识库的一对上、下近似来描述信息,而证据理论则是用一对信任函数和似真函数对系统进行评估. 文献[10]在粗糙集理论的框架下重新诠释了信度的概念;文献[11]则结合多种粗糙集模型对证据理论中的信度概念进行了详尽说明,特别是用上近似和下近似解释了置信函数和似真函数,并进一步论证了两种理论之间的联系. 可见,将证据理论引入粗糙集的研究中,有望给出知识发现的新机制.

本文在深入分析两种理论关系的基础上,给出基于信度区间的属性约简定理. 基于该定理,提出一种基于信度区间的故障特征属性的约简方法. 首先,利用随机模糊变量(RFV)和K均值实现特征数据的离散化,并获得相应的决策信息系统;然后,构建压缩二进制矩阵求取核属性,并根据粗糙集上、下近似算子与似真、置信函数之间的转换关系,设计新的基于属性信度区间的约简算法;最后,在电机柔性转子的特征融合故障诊断实验中,与经典的粗糙集简约方法的对比表明,所提出的方法在约简效率上有所提升,而且基于约简特征的融合诊断具有较高的确诊率.

1 理论基础

1.1 Dempster-Shafer证据理论

定义辨识框架$\Theta=\{F_l|l=1,2,\cdots,L\}$,其中的元素相互独立且互斥,它的幂集为$2^{\Theta}$. 令映射$m:2^{\Theta}\to[0,1]$为一个定义在$\Theta$上的基本信度赋值(BBA)函数,它满足$m(\varnothing)=0$和$\sum\limits_{A\in 2^{\Theta}}m(A)=1$. 这里将某个信息源提供的BBA称作证据. $m(A)$表示对A为真的支持度或信度. 相应的置信函数Bel和似真函数Pl分别定义为

$$\begin{aligned}&\mathrm{Bel}(A)=\sum_{B\subseteq A}m(B),\ A\subseteq\Theta;\\&\mathrm{Pl}(A)=\sum_{B\bigcap A\neq\varnothing}m(B),\ A\subseteq\Theta.\end{aligned}\tag{1}$$

其中:$\mathrm{Bel}(A)$表示对A及其全部子集赋予的信度的总和,$\mathrm{Pl}(A)$表示A不被拒绝的信度.

定义1 信度区间. 由Bel和Pl可以定义对象A的信度区间为$[\mathrm{Bel}(A),\mathrm{Pl}(A)]$,亦称作不确定区间,区间宽度$w=\mathrm{Pl}(A)-\mathrm{Bel}(A)$. w越小,表示信度度量的不确定性越小,利用Bel/Pl进行决策时的可靠性越高;反之,则不确定性越大. 因此,当$w>0$时,表示对A有一定程度的信任,也有一定程度的不信任,信任程度随w的减小而增大.

1.2 粗 糙 集

给定决策信息系统为

$$S=\{U,C\bigcup D,V\}.$$

其中:$U=\{x_1,x_2,\cdots,x_J\}$是非空有限的样本集,$C=\{c_1,c_2,\cdots,c_K\}$是条件属性集,$D=\{D_1,D_2,\cdots,D_N\}$是决策属性集. 在故障诊断中:C中的$c_k(k=1,2,\cdots,K)$表示第k种故障特征,D中的$D_n(n=1,2,\cdots,N)$表示第n种故障,V表示c_k和D_n取值的值域.

对于任意$P\subseteq C\bigcup D$,可以定义如下不可区分关系:

$$R_P=\{\{x,y\}\in U\times U|\forall c\in P,c(x)=c(y)\},\tag{2}$$

其中R_P也称作等价关系. 利用R_P对U进行划分,则生成样本$x\subseteq U$的等价类$[x]_P$为

$$[x]_P=\{y|y\in U,(x,y)\in R_P\}. \tag{3}$$

令$\partial_P=D(y):y\in[x]_P,x\in U$表示$x$关于$P$的广义决策值. 当$P=C$,且$|\partial_C(x)|=1$时,$S$为一致决策信息系统;当$|\partial_C(x)|>1$时,$S$为不一致决策信息系统.

定义2　上下近似. 在$S=\{U,C\bigcup D,V\}$中,对于任意$X\subseteq U,P\subseteq C\bigcup D$,定义集合$X$的$P$-上下近似为

$$\overline{P}(X)=\{x\in U:[x]_P\bigcap X\neq\varnothing\},$$
$$\underline{P}(X)=\{x\in U:[x]_P\subseteq X\}. \tag{4}$$

显然$\underline{P}(X)\subseteq X\subseteq\overline{P}(X)$,当$\underline{P}(X)=\overline{P}(X)=X$时,$X$是$P$上的可定义集,否则称为$P$上的粗糙集[12].

定义3　核属性. 对决策信息系统S的约简是一个NP-hard问题[13],其属性约简结果并不唯一,但S的所有约简结果的交集是确定的集合,称为核属性(集). 假设$\mathrm{RED}(C)$代表决策信息系统的所有约简结果,则信息系统S的核属性为

$$\mathrm{Core}(C)=\bigcap_{\mathrm{red}\in\mathrm{RED}(C)}\mathrm{red}. \tag{5}$$

2　基于信度区间的属性约简定理

由定义2可知,粗糙集中有一对近似算子概念,在证据理论中存在一对置信和似真函数. 由此可知两者之间存在如下联系:在$S=\{U,C\bigcup D,V\}$中,对于任意$X\subseteq U$和$P\subseteq C$,可以定义S上的一对置信和似真函数[11],即

$$\mathrm{Bel}_P(X)=\sum_{Y\in U/R_P\bigwedge Y\subseteq X}m_P(Y),$$
$$\mathrm{Pl}_P(X)=\sum_{Y\in U/R_P\bigwedge Y\bigcap X\neq\varnothing}m_P(Y). \tag{6}$$

其中

$$m_P(Y)=\begin{cases}p(Y)=|Y|/|U|,\ Y\in U/R_P;\\0.\end{cases} \tag{7}$$

这里$|Y|$和$|U|$分别代表了集合Y和集合U中元素的个数. 由此可知,集合X关于P的信度区间宽度为$w_P(X)=\mathrm{Pl_P}(X)-\mathrm{Bel}_P(X)$,由定义1和定义2可知,该式表示了$X$可被$P$定义的程度,$w_P$越小,表示定义程度越高,反之亦然. 本文基于$w_P(X)$给出的属性约简规则如定理1所示.

定理1　一致决策信息系统为$S=\{U,C\bigcup D,V\}$,针对所有$X\in U/R_D$,关于$P\subseteq C$的信度区间总宽度为$W_P=\sum_{X\in U/R_D}w_P(X)$. 对于某一给定的$P$,若$W_P=0$,且对于它的任意真子集$P'$,有$W_{P'}\neq0$成立,则$P$为$S$的属性约简结果,表示$P$对于任意$X\in U/R_D$可以完全定义.

证明　设$U/R_D=\{X_1,\cdots,X_n,\cdots,X_N\}$, X_n是由决策属性划分的样本集合,也可称为决策等价类. 对于任意$x\in X_n$,必有$[x]_P\bigcap X_n\neq\varnothing$,其中$P\subseteq C$,即有$\bigcup\{[x]_P:[x]_P\in U/R_P,[x]_P\bigcap X_n\neq\varnothing\}\supseteq X_n$,由此可得

$$\mathrm{Pl}_P(X_n)=$$
$$\sum\{m_P([x]_P):[x]_P\in U/R_P,[x]_P\bigcap X_n\neq\varnothing\}=$$
$$p(\bigcup\{[x]_P:[x]_P\in U/R_P,[x]_P\bigcap X_n\neq\varnothing\})\geqslant p(X_n).$$

定义$\xi_P(X)=\{[x]_P\in U/R_P:[x]_P\subseteq X\}$,这里$X\subseteq U$,则有

$$\mathrm{Bel}_P(X_n)=$$
$$\sum\{m_P([x]_P):[x]_P\in U/R_P,[x]_P\subseteq X_n\}=$$
$$\sum\{m_C([y]_C):[y]_C\in\xi_C([x]_P),$$
$$[x]_P\in U/R_P,[x]_P\subseteq X_n\}\leqslant$$
$$\sum\{m_C([y]_C):[y]_C\subseteq X_n,[y]_C\in U/R_C\}=$$
$$\mathrm{Bel}_c(X_n).$$

决策信息系统S为一致系统,即$\underline{C}(X_n)=X_n$,由式(6)可知$\mathrm{Bel}_C(X_n)=p(X_n)$. 综上可得$\mathrm{Pl}_P(X_n)\geqslant p(X_n)\geqslant\mathrm{Bel}_P(X_n)$,则有$W_P=\sum_{X\in U/R_D}w_P(X)=\sum_{X\in U/R_D}(\mathrm{Pl}_P(X_n)-\mathrm{Bel}_P(X_n))\geqslant0$,即$C$的任意子集$P$都满足该性质.

当P为S的属性约简(即$R_P\subseteq R_D$)时,任意$P'\subset P,R_P\subsetneq R_D$. 当$[x]_P\subseteq[x]_D$时,可得$[x]_P\bigcap X_n\neq\varnothing\Longleftrightarrow[x]_P\subseteq X_n$,并有

$$\mathrm{Pl}_P(X_n)=$$
$$\sum\{m_P([x]_P):[x]_P\in U/R_P,[x]_P\bigcap X_n\neq\varnothing\}=$$
$$\sum\{m_P([x]_P):[x]_P\in U/R_P,[x]_P\subseteq X_n\}=$$
$$\sum\{p([x]_P):[x]_P\in\xi_P(X_n)\}=p(X_n),$$
$$\mathrm{Bel}_P(X_n)=$$
$$\sum\{m_P([x]_P):[x]_P\in U/R_P,[x]_P\subseteq X_n\}=$$
$$\sum\{m_P([x]_P):[x]_P\in\xi_P(X_n)\}=$$
$$\sum\{p([x]_P):[x]_p\in\xi_p(X_n)\}=p(X_n).$$

则$W_P = \sum_{X\in U/R_D} w_P(X) = \sum_{X\in U/R_D} (\mathrm{Pl}_P(X_n) - \mathrm{Bel}_P(X_n)) = 0$,即只有当$P$为$C$的属性约简结果时,才满足该性质.

当S为不一致决策信息系统时,可以通过文献[14]的方法将广义决策属性作为决策属性,即$S' = \{U, C\bigcup \partial_C, V\}$,此时可得一致决策信息系统,即定理1适用于$S'$. □

3 基于信度区间的故障特征约简方法

基于信度区间的故障特征约简方法分为3个步骤:

1) 基于K均值和RFV的故障特征数据离散化. 因为传感器采集的数据一般都是连续型实值,而粗糙集一般处理离散型属性,因此要对数据进行离散化处理. K均值算法具有简单快速等特点,通过该算法可以初步划分故障特征数据. 另外,在设备运转过程中,测量环境的影响以及测量系统本身的系统误差(如传感器的精度或A/D转换器的量化误差)使得这些监测数据总是带有不确定性. 不确定性影响因素主要来自两个方面:一是传感器工作时受到的随机噪声干扰;二是传感器自身及其后续信号调理电路的系统性误差. 因此,在对故障特征数据进行离散化时,应该充分考虑“随机噪声干扰”和“系统性误差”这两个不确定性因素的区别,而RFV可以合理地表示测量不确定性中不同的影响因素(随机性和系统性误差),较为全面和自然地描述测量所具有的不确定性,通过RFV可以获取数据的离散信度表示,从而达到对数据的精细化离散处理.

2) 基于压缩二进制矩阵的核属性求取. 通常核属性是由差别矩阵求得,但当样本数量较多时,差别矩阵会占用大量的存储空间,同时,若存在重复样本,则会进行重复运算. 因此,本文采用压缩二进制矩阵,从而有效地减少矩阵的存储空间和计算量. 另外,在压缩二进制矩阵的构造过程中,通过将一致和不一致对象加以区分,并根据对应的条件构造矩阵中的元素,可以同时处理一致/不一致决策表,得到相应的核属性集.

3) 基于信度区间的属性简约算法. 将信度区间的宽度作为向核属性集中添加剩余属性的标准,信度区间宽度越小,表示对所有$X_n \in U/R_D$的可定义程度越高,当核属性集的信度区间左右边界相等时,表示其能完全定义U/R_D中的集合,此时的核属性集为决策信息系统的属性约简结果.

3.1 基于K均值和RFV的故障特征数据离散化

在故障类型D_n发生时,采集到故障特征c_k的Q个测量值$\{a_1, a_2, \cdots, a_Q\}$,则故障特征$c_k$共有$J = N \times Q$个测量数据. 通过分析数据的分布,选择数据簇个数$M$,然后使用$K$均值算法将数据划分为$M$个数据簇. 每个数据簇代表一个离散值,此时并没有考虑数据簇之间边界的模糊性,因此这里引入RFV来表示这种模糊性,它是一种特殊的II型模糊变量[15],由内外两个I型隶属度函数构成. 其中内部的隶属度函数表示系统性误差引起的不确定性,一般是依据传感器厂商提供的精度标准确定;外部隶属度函数表示监测数据的随机不确定性. 用RFV来建模不确定性信息时,能够合理地表示不同因素(随机性和系统性误差)对测量不确定性的贡献,相较于I型隶属度函数,其对不确定信息的刻画更为精细,能够捕捉到更多有用的信息[16].

具体离散过程如下: 1) 首先针对c_k采集的J个测量数据使用K均值算法,将其划分为M个数据簇$\mathrm{Cluster}_{k,t}(t = 1, 2, \cdots, M)$; 2) 利用$\mathrm{Cluster}_{k,t}$构造相应的RFV模型$\mathrm{RFV}_{k,t}$; 3) 将$c_k$的测量数据与$\mathrm{RFV}_{k,t}$匹配生成对应的BBA,并利用最大信度决策规则确定离散值. 具体构造RFV模型和匹配生成BBA的方法见文献[15,17]. 例如,根据故障特征c_k的测量数据划分的数据簇为$\mathrm{Cluster}_{k,1}, \mathrm{Cluster}_{k,2}, \mathrm{Cluster}_{k,3}$,对应的RFV模型分别为$\mathrm{RFV}_{k,1}, \mathrm{RFV}_{k,2}, \mathrm{RFV}_{k,3}$,若$c_k = 0.1136$,则将其与构造的RFV模型进行匹配所生成的BBA为(0.24, 0.58, 0.12, 0.06),由最大信度决策规则可以判定其对应的离散值为2. 将所有样本经过上述处理,即可得到如表1所示的决策信息系统.

表1 经离散化处理后的决策信息系统

U	c_1	$\cdots$	c_k	$\cdots$	c_K	D
x_1	$t_{1,1}$	$\cdots$	$t_{1,k}$	$\cdots$	$t_{1,K}$	D_1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_j	$t_{j,1}$	$\cdots$	$t_{j,k}$	$\cdots$	$t_{j,K}$	D_n
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_J	$t_{J,1}$	$\cdots$	$t_{J,k}$	$\cdots$	$t_{J,K}$	D_N

在表1中:$t_{j,k}$代表第j个样本中故障特征c_k的离散值,并且$t_{j,k} \in \{1, 2, \cdots, M\}$.

3.2 基于压缩二进制矩阵的核属性求取

在介绍压缩二进制矩阵(CBM)之前,首先给出二进制矩阵的定义[18].

定义4 二进制矩阵. 对于$S = \{U, C\bigcup D, V\}$,

定义如下二进制矩阵

$$\mathrm{BM}=[\mathrm{exampair}, r_{ij}].$$

其中: exampair 表示 U 中 x_i 和 x_j 的样本对; r_{ij} 中的元素 $r(i,j,c_k)$ 表示在属性 c_k 下, x_i 与 x_j 的差别. 令 $U_1=\bigcup\limits_{X\in U/R_D}$, $U_2=U-U_1$, 则有

$$r(i,j,c_k)=\begin{cases}1,\ (c_k(x_i)\neq c_k(x_j))\bigwedge\\ \quad(((x_i,x_j\in U_1)\bigwedge(D(x_i)\neq D(x_j)))\\ \quad\bigvee((x_i\in U_1)\bigwedge(x_j\in U_2)));\\ 0,\ \text{otherwise.}\end{cases}\tag{8}$$

式(8)中, 若 U_1 和 U_2 中存在条件属性相同的样本, 则矩阵会重复存储对应的元素, 造成内存的浪费.

定义5　过渡型二进制矩阵. 对于 $S=\{U,C\bigcup D,V\}$, 定义二进制矩阵 $\mathrm{BM}^*=[\mathrm{exampair}^*, r_{ij}^*]$, 其中 $\mathrm{exampair}^*$ 是论域 $U^*=\mathrm{delrep}(U)$ 的一组样本, 则有

$$r^*(i,j,c_k)=\begin{cases}1,\ (c_k(x_i)\neq c_k(x_j))\bigwedge\\ \quad(((x_i,x_j\in U_1^*)\bigwedge(D(x_i)\neq D(x_j)))\\ \quad\bigvee((x_i\in U_1^*)\bigwedge(x_j\in U_2^*)));\\ 0,\ \text{otherwise.}\end{cases}\tag{9}$$

其中: $U_1^*=\mathrm{delrep}(U_1)$, $U_2^*=\mathrm{delrep}(U_2)$, delrep 表示将集合中的重复样本删除.

当论域 U 中的样本过多时, 过渡型二进制矩阵依然需要很大的存储空间, 因此可以使用压缩二进制矩阵进一步减小矩阵所需的存储空间[19].

定义6　压缩二进制矩阵(CBM). 对于 $S=\{U, C\bigcup D,V\}$, 定义 $\mathrm{CBM}=[\mathrm{exampair}^*, \mathrm{totlone}_{ij}, \mathrm{ms}_{ij}]$. 其中: $\mathrm{ms}_{ij}=\sum\limits_{k=1}^{m}2^{m-k}r^*(i,j,c_k)$ 表示样本对的二进制编码, $\mathrm{totalonee}_{ij}=\sum\limits_{k=1}^{m}r^*(i,j,c_k)$ 表示编码中“1”出现的个数.

在CBM中找到满足 $\mathrm{totalone}_{i,j}=1$ 的行, 并找到这些行中 $\mathrm{ms}_{ij}=1$ 的对应的条件属性, 这些属性构成了 S 的核属性集Core. 压缩二进制矩阵可以将传统二进制矩阵空间复杂度由 $O(|\mathrm{exampair}|\times|C|)$ 降为 $O(|\mathrm{exampair}^*|\times 3)$, 其中 $|\mathrm{exampair}|$ 和 $|\mathrm{exampair}^*|$ 分别表示删除重复样本前后所能形成的样本对个数, 而在实际的仿真编程中, 可只保留 $\mathrm{totalone}_{i,j}=1$ 的那些行, 因此可以极大地减少二进制矩阵的存储空间.

3.3　基于属性信度区间的约简方法

本节根据定理1提出基于信度区间的属性约简方法. 具体的, 根据3.2节方法求取核属性集后, 将信度区间作为条件属性的选取标准, 向核属性集添加剩余属性, 直至满足终止条件, 并获得约简结果. 该算法描述如下.

输入: 决策信息系统 $S=\{U,C\bigcup D,V\}$;

输出: 属性约简结果red.

Step 1: 初始化 $\mathrm{red}=\varnothing$, $\mathrm{Core}=\varnothing$;

Step 2: 计算压缩二进制矩阵CBM, 将 $\mathrm{ms}_{ij}=1$ 且 $\mathrm{totalone}_{i,j}=1$ 对应的条件属性加入Core;

Step 3: do{

$C=C-\mathrm{Core}$

for each $c_j\in C$

$P=\mathrm{Core}\bigcup c_j$

$W_P=\sum\limits_{X\in U/R_D}(\mathrm{Pl}_P(X)-\mathrm{Bel}_P(X))$

end for

$c_j=\arg\min(W_P)$

$\mathrm{Core}=\mathrm{Core}\bigcup c_j$

}while$(\min(W_{\mathrm{Core}\bigcup c_j})\neq 0)$;

Step 4: 令 $\mathrm{red}=\mathrm{Core}$, 并输出red.

计算属性约简结果的迭代过程中, 将 $W_{\mathrm{Core}\bigcup c_j}$ 取值最小的 c_j 添加至Core中, 就能使 $\mathrm{Core}\bigcup c_j$ 对于任意 $X\in U/R_D$ 的可定义程度最高, 通过该策略能够快速找到属性约简结果, 避免盲目地向Core中添加属性. 该算法计算压缩二进制矩阵的时间复杂度为 $O(|C|\times|U|^2)$. 在Step 3中, 首先计算近似分类的时间复杂度 $O(|\mathrm{Core}|\times|U|^2)$, 剩余 $(|C|-|\mathrm{Core}|)$ 个属性, 以最坏的情况考虑, 该步骤的代价为 $O(|C|^2\times|U|^2)$, 因此整个算法的时间复杂度为 $O((|C|+1)\times|C|\times|U|^2)=O(|C|^2\times|U|^2)$. 由于压缩二进制矩阵占据的内存很少, 整个算法依然具有良好的效率. 同时, 该算法从证据理论的角度, 通过属性集在决策属性对论域划分上的信度区间表示该属性集对决策表自身分类能力拟合的不确定程度, 从而进行属性约简, 为粗糙集知识发现提供了一种新机制.

4　电机柔性转子故障诊断实验及对比分析

4.1　实验设置

这里以电机柔性转子故障诊断为例验证所提出方法的效果. 实验设备为ZHS-2型多功能柔性转子试验台, 将振动位移传感器分别安置在转子支撑座的水平和垂直方向来采集转子振动信号, 经过HG-8902采集箱将信号传输至计算机, 然后通过Labview环境

下的HG-8902数据分析软件得到转子振动加速度频谱以及时域振动位移平均幅值,并将其作为故障特征信号[9,16].

在试验台上设置6种故障模式:正常运行D_1、转子不平衡D_2、转子不对中D_3、基座松动D_4、连接器松动D_5、齿轮缺齿D_6.通过5个传感器对电机转子运行情况进行监测,由大量实验数据的分析可知,以上故障模式都会引起一定频率成分的振动幅值的增加或减少.因此,这里选取对故障较为敏感的1X、2X和3X倍频的幅值作为故障特征变量.转子转速为1 500 r/m,则1X = 25 Hz, 2X = 50 Hz, 3X = 75 Hz.那么,5个传感器共计有15个故障特征(条件属性),分别记为$c_1, c_2, \cdots, c_{15}$,根据振动传感器出厂标定可知,系统误差为$\Delta = \pm\varepsilon\% = \pm 1\%$.

在每种故障模式下,采集各故障特征$c_1, c_2, \cdots, c_{15}$下的210次测量结果.随机选取其中100个数据作为训练样本并用于属性约简,属性约简的PC机硬件环境如下:处理器为酷睿i3双核1.9 GHz,内存为4 GB,硬盘为500 GB(5 400转/分),运行软件为Matlab 2010a;剩余的110个作为检验分类性能的测试样本.根据3.1节方法将训练样本中的数据进行离散化处理,得到如表1所示的600×16的矩阵型决策信息系统S.

4.2 本文方法与传统粗糙集方法的对比与分析

为了说明本文所提出的离散化方法的有效性,这里与具有代表性的离散化方法进行对比:方法1为传统的基于K均值离散化方法(非监督离散化方法);方法2则利用信息熵计算断点来划分数据[20](有监督离散化方法).同时,采用本文的约简算法得到属性约简结果,结果如表2所示.

表2 离散化方法对比

方法	约简后的属性	属性个数	确诊率/%
方法1	$c_1, c_2, c_6, c_7, c_8, c_{10}, c_{13}, c_{14}$	8	90.2
方法2	$c_1, c_2, c_3, c_6, c_8, c_{10}, c_{13}$	7	87.9
本文方法	$c_1, c_2, c_6, c_{13}, c_{14}$	5	89.8

由于3种离散化方法得到的决策表是不一样的,根据其计算的核属性以及属性组合划分的等价类也是不同的,而在迭代过程中计算相应属性集的信度区间依赖于该等价类,这使得3种方法的约简结果不同.根据方法1和方法2获得的决策表,约简后的属性个数都要比本文方法多,单纯比较约简后的属性个数并不能完全说明离散化方法的优劣性,在这里给出相应的故障诊断准确率,诊断结果都是由文献[17]的证据融合故障诊断方法给出的.方法1约简后的属性都包含了本文方法约简后的属性,用于诊断的信息量也比较丰富,因此其诊断准确率比本文方法高,但并没有相差太多,同时会造成诊断的实时性效果较差;方法2约简后的属性虽然也比本文约简后的属性多,但却没有特征c_{14},因而导致确诊率低,就故障诊断实验中的确诊率而言,特征c_{14}是比较重要的属性;本文方法得到的属性约简结果个数最少,同时保持了较高的诊断准确率,在实时性上要比方法1和方法2好,准确率与方法1相差无几,比方法2高.综合上述分析,本文提出的离散化方法是有效的并且适用于故障诊断领域的特征数据离散.

针对本文离散化方法所得的决策信息系统S,利用3.2节压缩二进制定义求得核属性集$\text{Core} = \{c_1, c_{14}\}$,则可得剩余属性集$C = \{c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}, c_{11}, c_{12}, c_{13}, c_{15}\}$.

分别计算C中各属性的信度区间宽度,即

$$W_{c_2 \cup \text{Core}} = 0.761\,7,\ W_{c_3 \cup \text{Core}} = 1.018\,3,$$

$$W_{c_4 \cup \text{Core}} = 1.065,\ W_{c_5 \cup \text{Core}} = 1.17,$$

$$W_{c_6 \cup \text{Core}} = 0.536\,7,\ W_{c_7 \cup \text{Core}} = 1.35,$$

$$W_{c_8 \cup \text{Core}} = 1.695,\ W_{c_9 \cup \text{Core}} = 1.103\,4,$$

$$W_{c_{10} \cup \text{Core}} = 0.996\,6,\ W_{c_{11} \cup \text{Core}} = 1.298\,3,$$

$$W_{c_{12} \cup \text{Core}} = 1.601\,7,\ W_{c_{13} \cup \text{Core}} = 1.008\,3,$$

$$W_{c_{15} \cup \text{Core}} = 1.813\,3.$$

由3.3节给出的算法可知,应将c_6添加至Core,此时$\text{Core} = \{c_1, c_6, c_{14}\}$,$C = \{c_2, c_3, c_4, c_5, c_7, c_8, c_9, c_{10}, c_{11}, c_{12}, c_{13}, c_{15}\}$,继续按照上述算法步骤进行迭代运算直至$W(\text{Core}) = 0$,停止运算,最终计算得到的属性约简集为$\text{red} = \text{Core} = \{c_1, c_2, c_6, c_{13}, c_{14}\}$,那么将原有的15个属性约简至5个.

为了进一步说明本文算法的有效性,将所提算法与以下3种经典的粗糙集方法进行对比:方法A为基于属性频率的约简算法[21];方法B为基于属性重要度的约简算法[22];方法C为基于区分度和区分率的约简算法[23].这3种方法分别将属性频率、单属性的近似精度以及区分度和区分率作为迭代过程中属性选取的标准.

表3显示了对应方法的约简结果,运行时间以及诊断效果.

表 3　粗糙集方法的性能比较

方法	约简后的属性	属性个数	约简耗时 / 诊断耗时 / s	确诊率 / %
不约简	/	/	/ 34.43	90.8
方法A	$c_1, c_2, c_6, c_7, c_9, c_{10}, c_{12}, c_{13}, c_{14}, c_{15}$	10	22.15 / 22.86	89.8
方法B	$c_1, c_2, c_6, c_{13}, c_{14}$	5	27.55 / 11.07	89.8
方法C	$c_1, c_2, c_3, c_6, c_{10}, c_{13}, c_{14}$	7	113.56 / 16.43	90.1
本文方法	$c_1, c_2, c_6, c_{13}, c_{14}$	5	23.62 / 11.07	89.8

由表3的统计结果可以看出,利用约简后的属性进行诊断的确诊率与全部属性下的确诊率相差无几,这说明确实存在一些不重要的特征属性,对它们的约简是有必要的. 方法A虽然算法效率较高(约简耗时相对较小),但约简效果欠佳,特征属性保留过多,影响诊断效率(诊断耗时相对较长),而本文方法所用属性远远少于方法A,但两者的确诊率一样,这说明本文方法的诊断效率将远远优于方法A;方法B虽然与本文方法约简结果一致,但在构造传统的差别矩阵上,需要大量的存储空间,故约简效率不高;方法C采用传统的二进制矩阵,因为样本规模和属性数量较大,同时存在重复样本,导致重复计算,所以约简效率最差,约简结果不理想;在计算压缩二进制矩阵的过程中,本文方法去除了重复样本,压缩了二进制矩阵占用的存储空间,故减少了计算量,并且本文方法是通过属性信度衡量属性的价值的,其要比属性频率和区分率更加可靠,这是因为计算属性信度时,考虑了条件属性与决策属性之间的关系,同时该方法也为粗糙集知识发现提供了一种新机制,故本文方法的约简效率普遍优于其他方法. 在故障诊断过程中,与使用全属性进行融合诊断的耗时相比,属性约简能够大大节省后续诊断所需时间,保证了诊断的实时性. 特别是在面对更大规模数据的情况下,约简后的诊断时效优势会更加明显.

最后,20次交叉实验表明,本文方法属性约简后故障特征的个数都为5 ～ 7个,故障诊断的确诊率在87.4％ ～ 89.8％之间,而其他方法约简后的特征个数普遍比本方法多(5 ～ 11个),但确诊率并没有得到大幅提升(86.9％ ～ 90.1％),这进一步说明了本文方法的有效性和稳定性.

5　结　论

针对故障诊断中故障特征(条件属性)数量过多会影响诊断时效性的问题,本文根据粗糙集上下近似与证据理论信度测度之间的联系,提出了基于信度区间的属性约简定理,即属性约简结果能对所有决策等价类完全定义,且该结果的信度区间宽度为0,并在该定理的基础上给出了一种基于属性信度区间的故障特征约简方法来满足诊断的实时性需求. 利用K均值和RFV对故障特征数据进行精细化离散处理,获得了决策信息系统;再构建压缩二进制矩阵快速得到核属性集;计算剩余属性的信度区间,并将具有最小信度区间的属性添加至核属性集,直至满足终止条件,得到属性约简结果. 电机柔性转子故障诊断实验表明,所提出的属性约简方法效果较好,不仅能提高属性约简过程的效率,而且能根据约简后的属性进行故障诊断,大幅减少诊断时间,保证故障诊断的实时性.

参考文献(References)

[1] Shao R, Hu W, Wang Y, et al. The fault feature extraction and classification of gear using principal component analysis and kernel principal component analysis based on the wavelet packet transform[J]. Measurement, 2014, 54(6): 118-132.

[2] Pawlak Z. Rough sets[J]. Int J of Parallel Programming, 1982, 38(5): 88-95.

[3] Chen D, Li W, Zhang X, et al. Evidence-theory-based numerical algorithms of attribute reduction with neighborhood-covering rough sets[J]. Int J of Approximate Reasoning, 2014, 55(3): 908-923.

[4] Hu X, Cercone N. Learning in relational databases: A rough set approach[J]. Computational Intelligence, 1995, 11(2): 323-338.

[5] 叶东毅. Jelonek 属性约简算法的一个改进[J]. 电子学报, 2000, 28(12): 81-82.
(Ye D Y. An improvement to Jelonek's attribute reduction algorithm[J]. Acta Electronica Sinica, 2000, 28(12): 81-82.)

[6] 叶东毅, 陈昭炯. 一个新的差别矩阵及其求核方法[J]. 电子学报, 2002, 30(7): 1086-1088.
(Ye D Y, Chen Z J. A new discernibility matrix and the computation of a core[J]. Acta Electronica Sinica, 2002, 30(7): 1086-1088.)

[7] 张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001: 36-38.
(Zhang W X, Wu W Z, Liang J Y, et al. Rough sets theory

and method[M]. Beijing: Science Press, 2011: 36-38.)

[8] 刘文军, 谷云东, 冯艳宾, 等. 基于可辨识矩阵和逻辑运算的属性约简算法的改进[J]. 模式识别与人工智能, 2004, 17(1): 119-123.
(Liu W J, Gu Y D, Feng Y B, et al. Improvement of attribute reduction algorithm based on discernibility matrix and logic operation[J]. Pattern Recognition and Artificial Intelligence, 2004, 17(1): 119-123.)

[9] 徐晓滨, 文成林, 孙新亚, 等. 设备故障诊断中的证据融合与决策方法[M]. 北京: 科学出版社, 2017: 1-30.
(Xu X B, Wen C L, Sun X Y, et al. Evidence fusion and decision making methods in equipment fault diagnosis[M]. Beijing: Science Press, 2017: 1-30.)

[10] Skowron A. The relationship between the rough set theory and evidence theory[J]. Bulletin of Polish Academy of Science: Mathematics, 1989, 37(1): 87-90.

[11] Wu W Z, Leung Y, Zhang W X. Connections between rough set theory and Dempster-Shafer theory of evidence[J]. Int J of General Systems, 2002, 31(4): 405-430.

[12] 张清华, 王国胤, 肖雨. 粗糙集的近似集[J]. 软件学报, 2012, 23(7): 1745-1759.
(Zhang Q H, Wang G Y, Xiao Y. Approximate sets of rough sets[J]. J of Software, 2012, 23(7): 1745-1759.)

[13] Ye D, Chen Z, Ma S. A novel and better fitness evaluation for rough set based minimum attribute reduction problem[J]. Information Sciences, 2013, 222(3): 413-423.

[14] Wu W Z. Attribute reduction based on evidence theory in incomplete decision systems[J]. Information Sciences, 2008, 178(5): 1355-1371.

[15] Xu X B, Zhou Z, Wen C L. Data fusion algorithm of fault diagnosis considering sensor measurement uncertainty[J]. Int J on Smart Sensing & Intelligent Systems, 2013, 6(1): 171-190.

[16] 徐晓滨, 文成林, 王迎昌. 基于模糊故障特征信息的随机集度量信息融合诊断方法[J]. 电子与信息学报, 2009(7): 1635-1640.
(Xu X B, Wen C L, Wang Y C. Information fusion algorithm of fault diagnosis based on random set metrics of fuzzy feature[J]. J of Electronics and Information Technology, 2009, 31(7): 1635-1640.)

[17] 侯平智, 张明, 徐晓滨, 等. 基于K近邻证据融合的故障诊断方法[J]. 控制与决策, 2017, 32(10): 1766-1744.
(Hou P Z, Zhang M, Xu X B, et al. Fault diagnosis based on KNN evidence fusion[J]. Control and Decision, 2017, 32(10): 1766-1744.)

[18] 杨明. 一种基于改进差别矩阵的核增量式更新算法[J]. 计算机学报, 2006, 29(3): 407-413.
(Yang M. An incremental updating algorithm of the computation of a core based on the improved discernibility matrix[J]. Chinese J of Computers, 2006, 29(3): 407-413.)

[19] 丁棉卫, 张腾飞, 马福民. 基于二进制区分矩阵的不完备系统增量式属性约简算法[J]. 计算机工程, 2017, 44(1): 244-250.
(Ding M W, Zhang T F, Ma F M. Increment attribute reduction algorithm based on binary discernibility matrix[J]. Computer Engineering, 2017, 44(1): 244-250.)

[20] 谢宏, 程浩忠, 牛东晓. 基于信息熵的粗糙集连续属性离散化算法[J]. 计算机学报, 2005, 28(9): 1570-1574.
(Xie H, Cheng H Z, Niu D X. Discretization of continuous attributes in rough set theory based on information entropy[J]. Chinese J of Computers, 2005, 28(9): 1570-1574.)

[21] 葛浩, 李龙澍, 杨传健. 新的可分辨矩阵及其约简方法[J]. 控制与决策, 2010, 25(12): 1891-1895.
(Ge H, Li L S, Yang C J. New discernibility matrix and attribute reduction method[J]. Control and Decision, 2010, 25(12): 1891-1895.)

[22] Liu Y, Jiang Y, Yang J. Feature reduction with inconsistency[J]. Int J of Cognitive Informatics and Natural Intelligence(IJCINI), 2010, 4(2): 77-87.

[23] Yang P, Li J, Huang Y. An attribute reduction algorithm by rough set based on binary discernibility matrix[C]. The 5th Int Conf on Fuzzy Systems and Knowledge Discovery. Jinan: IEEE, 2008: 276-280.

(责任编辑：闫 妍)