面向混合属性数据集的改进半监督 FCM 聚类方法

李晓庆 [1] 唐 昊 [1] 司加胜 [1] 苗刚中 [1]

摘 要 针对混合属性数据集聚类精度低的问题, 本文提出一种基于改进距离度量的半监督模糊均值聚类 (Fuzzy C-means, FCM) 算法. 首先, 在数据集中针对类别属性进行预处理, 并设置相应的相异度阈值; 将传统聚类距离度量与改进的 Jaccard 距离度量结合, 确定混合属性数据集的距离度量函数; 最后, 将所得距离度量函数与传统半监督 FCM 算法相结合, 并在滚动轴承的不同复合故障数据的特征集中进行聚类. 实验表明, 该算法能在含无序属性的混合属性数据集的聚类中取得更好的聚类效果.

关键词 混合属性, 相异度阈值, 模糊均值聚类, Jaccard

引用格式 李晓庆, 唐昊, 司加胜, 苗刚中. 面向混合属性数据集的改进半监督 FCM 聚类方法. 自动化学报, 2018, **44**(12): 2259−2268

DOI 10.16383/j.aas.2018.c170510

An Improved Semi-supervised FCM Clustering Method for Mixed Data Sets

LI Xiao-Qing[1] TANG Hao[1] SI Jia-Sheng[1] MIAO Gang-Zhong[1]

Abstract This paper puts forward a semi-supervised fuzzy C-means (FCM) algorithm based on an improved distance measure to solve the problem of low accuracy of clustering algorithm of data sets with mixed attributes. First, the classification attributes are preprocessed in the data set, and the corresponding dissimilarity threshold is set. Then the traditional clustering distance measure is combined with the improved Jaccard distance measure to determine the distance measure function. Finally, the distance measure function is combined with the traditional semi-supervised FCM algorithm, and clustering is carried out on the characteristic data sets of different coupling fault data of rolling bearings. Simulation results show that the algorithm can achieve better clustering accuracy in mixed data sets.

Key words Mixed attributes, dissimilarity threshold, fuzzy C-means (FCM), Jaccard

Citation Li Xiao-Qing, Tang Hao, Si Jia-Sheng, Miao Gang-Zhong. An improved semi-supervised FCM clustering method for mixed data sets. *Acta Automatica Sinica*, 2018, **44**(12): 2259−2268

聚类过程主要包括数据准备、特征选取与提取、相似度计算、聚类与评估等步骤, 经典的聚类算法包含 K-means、K-modes、模糊均值聚类 (Fuzzy C-means, FCM) 算法、DBSCAN 等. 目前仍有关于经典聚类算法的衍生算法的研究, 文献 [1] 以近邻反射传播聚类算法为基础, 提出一种基于同类约束的半监督近邻反射传播聚类方法. 文献 [2] 提出 K- 近邻估计协同系数的协同模糊 C 均值算法. 然而, 这些聚类算法的距离度量函数是仅针对单属性的数据集的距离运算.

随着互联网和物联网的快速发展和广泛应用, 各种数据的数量呈现指数式增长, 可获取的数据属性也呈现出多样化. 许多学者开始致力于混合属性数据集聚类的相关研究. Huang[3] 提出一种适用于混合属性数据聚类的 K-prototypes 算法, 对于分类属性部分, 该算法采用匹配差异度来描述数据点之间相异度. 近年来, 陈晋音等[4] 提出一种面向混合属性数据的增量式聚类算法. 根据混合属性数据特征, 将特征向量集分为数值占优、分类占优和均衡型三类. 文献 [5] 对不同情况的特征选取相应的距离度量方式进行分析, 通过预设参数, 发现数据密集区域, 确定核心点, 进而利用核心点确定密度相连的对象实现聚类. 文献 [6] 提出一种基于密度的聚类中心自动确定的混合属性数据聚类算法. 以上文献在处理混合属性数据的聚类时, 并未考虑无序属性数据的聚类问题.

文献 [7] 将混合属性数据分为有序属性和无序属性两个部分, 并构造出双重近邻无向图, 但未对混合属性数据聚类时距离度量做深入研究. 文献 [8] 针对不同维度的向量间的无序属性向量集的距离度量展开研究. 文献 [9] 针对机械系统故障诊断中对先验

收稿日期 2017-09-06 录用日期 2017-12-06
Manuscript received September 6, 2017; accepted December 6, 2017
国家重点研发计划 (2017YFB0902600), 国家自然科学基金 (61573126) 资助
Supported by National Key Research and Development Program of China (2017YFB0902600) and National Natural Science Foundation of China (61573126)
本文责任编委 刘艳军
Recommended by Associate Editor LIU Yan-Jun
1. 合肥工业大学电气与自动化工程学院 合肥 230009
1. School of Electrical Engineering and Automation, Hefei University of Technology, Hefei 230009

知识利用不足和在高维特征空间中诊断难的问题, 提出一种基于成对约束和通过约束准则构造核函数的半监督谱核聚类方法. 本文基于文献 [7−9] 提出一种改进的半监督 FCM 算法, 首先对混合数据集的构成进行占优分析, 确定占优因子 α, 对 Jaccard 距离做阈值改进, 并将所获改进 Jaccard 距离作为无序属性距离度量函数, 进而将所得混合属性距离度量函数应用于半监督 FCM 聚类算法, 得到改进的半监督 FCM 聚类算法. 最后, 在滚动轴承的不同类型单故障及复合故障数据的特征集中进行算法对比验证.

1 混合属性数据集及其距离度量

数据集由多个数据组成, 每个数据对象由其属性进行描述. 数据库中的每个对象以一元组的形式呈现, 每一列代表一个属性. 数据挖掘中常用的属性类型包括: 1) 数值属性, 通常用实数值来描述, 包括离散型数值和连续型数值之分; 2) 分类 (标称) 属性, 每个不同的值代表某种类别、代码或状态, 这些值无列别顺序; 3) 二值属性, 取值只有 1 或 0 两种情况. 通常 1 表示属性值非空, 0 表示属性值为空值; 4) 序数属性, 属性取值的值域是一个有意义的序列.

以上为常规属性类型, 当数据对象包含多种属性类型时, 称为混合属性数据. 本文将混合属性分为有序属性和无序属性两类, 划分依据是此属性有无列别顺序. 常规属性中, 数值属性和序数属性属于有序属性, 分类属性属于无序属性, 若二值属性维数较多, 则只能看成有序属性, 若维数为 1, 则既能看成有序属性, 亦能看成无序属性.

对于数据集的距离度量是进行有意义的聚类分析的前提, 若存在某混合属性数据集表达式为 $\Phi = \{\boldsymbol{X_1}, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n\}$, 记混合属性特征向量 $\boldsymbol{X}_i = (x_{i1}, x_{i2}, \cdots, x_{il})$, 前 m 维属性为有序属性, 后 $l-m$ 维为无序属性. 对于上述混合属性距离度量而言, 一般将混合属性数据按照属性类型进行划分, 分别求解距离, 再进行整体距离的加权求和. 本节对有序属性和无序属性的距离度量进行简要阐述, 并对无序属性的距离度量方法加以改进, 最后给出本文提出的混合属性距离度量的完备性证明.

1.1 欧氏距离

本文在处理前 m 维有序属性的距离计算时, 采用欧氏距离作为距离度量函数. 在距离度量中, 闵可夫斯基距离 (Minkowski distance) 是衡量数值点之间距离的一种非常常见的方法, 计算公式为

$$\mathrm{dist}\,(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sqrt[p]{\sum_{r=1}^{m} |X_{ir} - X_{jr}|^p} \tag{1}$$

其中, 如果 $p \to \infty$ 时, 就是切比雪夫距离; $p = 1$ 时, 表示曼哈顿距离; $p = 2$ 时, 表示欧氏距离, 即

$$\mathrm{dist}\,(\boldsymbol{X}_i, \boldsymbol{X}_j) = \sqrt{\sum_{r=1}^{m} |X_{ir} - X_{jr}|^2} \tag{2}$$

可以看出, 欧氏距离是两个向量相对应维度的运算, 即欧氏距离适用于有序属性的计算.

1.2 Jaccard 距离及其改进

对于后 $l-m$ 维的无序属性, 本文采用改进的 Jaccard 距离度量方法.

传统的 Jaccard 相似度常用于二值型数据的相似度计算. 在数据挖掘中, 经常将属性值二值化, 通过计算 Jaccard 相似度, 可以简单快速地得到两个对象的相似程度. 记集合 $A = \{X_{i(m+1)}, X_{i(m+2)}, \cdots, X_{il}\}$, 集合 $B = \{X_{j(m+1)}, X_{j(m+2)}, \cdots, X_{jl}\}$, 则 A 和 B 的 Jaccard 相似系数定义为

$$D_J\,(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

相应的 Jaccard 距离定义为

$$\hat{D}_J\,(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|} \tag{4}$$

其中, Jaccard 相似系数反映了 A 和 B 集合的相交程度, 值在 $[0, 1]$ 范围之内, 若 A 和 B 不相交, 则值为 0.

广义 Jaccard 相似系数定义[10] 为

$$\vec{D}_J(A, B) = \frac{\sum\limits_{r=m+1}^{l} X_{ir} X_{jr}}{\sum\limits_{r=m+1}^{l} X_{ir}^2 + \sum\limits_{r=m+1}^{l} X_{jr}^2 - \sum\limits_{r=m+1}^{l} X_{ir} X_{jr}} \tag{5}$$

可见, 广义 Jaccard 相似系数虽然考虑向量中各维数值的大小, 但是向量属性的排序对计算结果有一定的影响. 因此, 广义 Jaccard 相似系数处理无序属性集的效果不理想.

实际生活或生产环境下, 传感设备所得数值存在一定的误差, 本文对相似系数计算做了相应改进, 引入相异度阈值系数 ε 修正属性数值的相似性判断, 则相似性判断公式为

$$1 - \varepsilon \le \frac{X_{ip}}{X_{jq}} \le 1 + \varepsilon \tag{6}$$

其中, $p = m+1, m+2, \cdots, l$; $q = m+1, m+2, \cdots, l$. 即若向量 A 和 B 中存在两个属性值 X_{ip} 和 X_{jq} 满足以上条件, 则令

$$X_{ip} = X_{jq} = \frac{X_{ip} + X_{jq}}{2} \in A \cap B$$

定义 $\|\cdot\|$ 为考虑相异度阈值下的集合长度, 则改进的 Jaccard 相似系数表达式为

$$D'_J(A,B) = \frac{\|A \cap B\|}{\|A \cup B\|} \tag{7}$$

1.3 混合属性距离度量的完备性证明

有序属性部分距离度量采用欧氏距离, 无序属性部分距离度量采用改进的 Jaccard 距离, 则混合属性的距离表达式为

$$d(\boldsymbol{x},\boldsymbol{y}) = \beta\sqrt{\sum_{i=1}^{\mathrm{m}}(x_{1i}-y_{1i})^2} + (1-\beta)\left[1 - D'_J(x_2,y_2)\right] \tag{8}$$

其中, $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 均为前 m 个有序属性以及 $l-m$ 个无序属性组成的混合属性向量, $\boldsymbol{x}_1$ 和 $\boldsymbol{y}_1$ 分别为 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 的前 m 个有序属性组成的向量, x_2 和 y_2 分别为 $\boldsymbol{x}$ 与 $\boldsymbol{y}$ 的 $l-m$ 个无序属性组成的集合, 为了均衡非占优属性对数据对象整体相似性的影响, 引入占优因子 α[3], 并针对本文算例取值 0.6, 若 $m/l > \alpha$, 则特征向量集是数值占优数据集, 则令 $\beta = 0.4$, 若 $(l-m)/l > \alpha$, 则特征向量集是分类占优数据集, 则令 $\beta = 0.6$, 若以上两个条件均不满足, 则特征向量集是均衡型混合属性数据集, 令 $\beta = 0.5$.

距离定义需满足同一性、非负性、对称性和三角不等性, 为了使证明过程更加清晰, 记有 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 和 $\boldsymbol{z}$ 三个向量, $\boldsymbol{x}_A$ 和 $\boldsymbol{x}_B$ 和 $\boldsymbol{x}_C$ 为有序向量部分, 维数为 m, A 和 B 和 C 为无序属性部分构成的集合, 维数为 l.

定理 1. 若 $\|A \cap B\| = |M| = k$, $0 \le k \le l$, 且有 $\|M \cap C\| = p$, $0 \le p \le k$, 则 $\|A \cap C\| + \|B \cap C\| \le p + l$ (M 为集合 A 和 B 考虑相异度阈值情况下, 求交集所得的集合).

证明. 若 $\|A \cap B\| = k$, 则 $\|A \cup B\| = 2l - k$. 向量 A 中已有 p 个元素属于 C, $k-p$ 个元素不属于 C, 及 $l-k$ 个元素可能属于 C. 同理, B 的情况亦然. 易证, A 和 B 中相异元素属于 C 的个数最大值为 $l-p$, 即

$$\|A \cap C\| + \|B \cap C\| \le 2p + l - p = p + l$$

□

推论 1. 本文所提混合属性距离满足三角不等性.

证明. 需证 $d(\boldsymbol{x},\boldsymbol{z}) + d(\boldsymbol{y},\boldsymbol{z}) \ge d(\boldsymbol{x},\boldsymbol{y})$, 即

$$\sqrt{\sum_{i=1}^{m}(x_{Ai}-x_{Ci})^2} + 1 - D'_J(A,C) + \sqrt{\sum_{i=1}^{m}(x_{Bi}-x_{Ci})^2} + 1 - D'_J(B,C) \ge \sqrt{\sum_{i=1}^{m}(x_{Ai}-x_{Bi})^2} + 1 - D'_J(A,B)$$

将欧氏距离统一放置等式左侧, 即

$$\sqrt{\sum_{i=1}^{m}(x_{Ai}-x_{Ci})^2} + \sqrt{\sum_{i=1}^{m}(x_{Bi}-x_{Ci})^2} - \sqrt{\sum_{i=1}^{m}(x_{Ai}-x_{Bi})^2} + 1 - D'_J(A,C) - D'_J(B,C) \ge -D'_J(A,B)$$

不等式左侧 $\ge 1 - D'_J(A,C) - D'_J(B,C)$, 根据定理 1, 有

$$\begin{aligned} 1 - D'_J(A,C) - D'_J(B,C) &= 1 - \left(\frac{\|A \cap C\|}{2l - \|A \cap C\|} + \frac{\|B \cap C\|}{2l - \|B \cap C\|}\right) \ge \\ &1 - \left(\frac{\|A \cap C\|}{2l - \|A \cap C\|} + \frac{l + p - \|A \cap C\|}{l - p + \|A \cap C\|}\right) \ge \\ &1 - \left(\frac{\|A \cap C\|}{2l - \|A \cap C\|} + \frac{l + k - \|A \cap C\|}{l - k + \|A \cap C\|}\right) \ge \\ &1 - \left(\frac{\|A \cap C\|}{2l - \|A \cap C\|} + \frac{2l - \|A \cap C\|}{\|A \cap C\|}\right) \ge 0 \end{aligned} \tag{9}$$

□

推论 2. 混合属性距离度量满足距离度量准则.

证明.

1)

$$d(x_A,x_A) = \sqrt{\sum_{i=1}^{m}(x_{Ai}-x_{Ai})^2} + 1 - D'_J(A,A) = 0$$

满足到自己距离为零;

2) $D'_J(A,B) \in [0,1]$, 可知,

$$d(x,y)=\sqrt{\sum_{i=1}^{m}(x_{Ai}-x_{Bi})^2}+[1-D'_J(A,B)]\geq 0$$

满足非负性;

3) $d(x,y)=d(y,x)$, 满足对称性;

4) 由推论 1 可知, 满足三角不等性. 故混合属性距离度量满足距离度量准则. □

2 改进半监督 FCM 算法

2.1 FCM 算法

FCM 算法是根据不同样本点对聚类中心的隶属度不同来划分聚类的算法, 它的隶属度取值由 K-means 聚类算法的 $\{0,1\}$, 拓展至 $[0,1]$, 即每个样本的类别隶属度为一个实数区间, 相较而言, 更具灵活性.

记 $\boldsymbol{X}_i$ $(i=1,2,\cdots,n)$ 中每一个向量 $\boldsymbol{X}_i$ 均有 l 维属性. 根据选定的相似性度量函数, 划分为 c 个聚类中心称为簇 V_k, 其中 $k=1,2,\cdots,c$. 那么 n 个样本分别属于 c 个类别的隶属度矩阵记为 $U=[u_{ik}]_{c\times n}$ (模糊划分矩阵), 其中 u_{ik} $(1\leq i\leq n,\ 1\leq k\leq c)$ 表示第 i 个样本 $\boldsymbol{X}_i$ 属于第 k 个类别的隶属度, 应满足以下约束条件:

$$u_{ik}\in[0,1],\qquad 1\leq i\leq n,\ 1\leq k\leq c \tag{10}$$

$$\sum_{k=1}^{c}u_{ik}=1,\qquad 1\leq i\leq n \tag{11}$$

FCM 算法的目标函数定义为

$$J(U,V)=\sum_{k=1}^{c}\sum_{i=1}^{n}(u_{ik})^2d^2(\boldsymbol{X}_i,\boldsymbol{v}_k) \tag{12}$$

聚类中心的迭代公式为

$$\boldsymbol{v}_k=\frac{\sum\limits_{i=1}^{n}(\boldsymbol{X}_iu_{ik}^2)}{\sum\limits_{i=1}^{n}u_{ik}^2} \tag{13}$$

以下为传统 FCM 的算法描述.

算法 1. FCM 算法

输入. 待聚类样本.

输出. 聚类中心及隶属度矩阵.

步骤 1. 给定需要划分的聚类中心数目 c 及相关参数;

步骤 2. 初始化隶属度矩阵 U;

步骤 3. 根据公式计算 c 个聚类中心;

步骤 4. 计算出各个样本点到聚类中心的距离矩阵, 并得到新的隶属度矩阵 (若分母为 0, 则令 $u_{ik}=1$);

步骤 5. 计算出目标函数值 J. 如果小于给定的阈值 δ 或与上次循环产生的目标函数值之差小于阈值 δ 则算法停止. 否则, 返回步骤 2.

2.2 半监督 FCM 算法的改进

多数情况下的聚类集成算法建立在非监督方式之上, 由于缺乏对先验知识的利用, 致使聚类集成的准确性、鲁棒性和稳定性有所降低.

半监督模糊聚类突破了有监督和无监督模糊聚类中只考虑一种样本类型的局限, 整体考虑数据集中的所有样本, 提高了未知样本的使用率, 从而改善了聚类效果. 它的核心思想是利用监督数据, 得到初始的聚类划分, 然后利用得到的初始的聚类划分对未标记的数据进行约束指导[11].

本文将改进距离度量公式与半监督模糊聚类算法结合, 得到改进的半监督 FCM 算法目标函数.

$$J(U,W)=\sum_{k=1}^{C}\sum_{i=1}^{N}(u_{ik})^2\left\{\sum_{r=1}^{\mathrm{m}}(x_{ir}-v_{kr})^2+[1-D'_J(x''_i,v''_k)]\right\} \tag{14}$$

记 $\boldsymbol{x}'_i$ 表示向量 $\boldsymbol{X}_i$ 的有序属性向量部分, x''_i 表示向量 $\boldsymbol{X}_i$ 的无序属性构成的集合. 聚类中心向量 $\boldsymbol{v}_k$ 为有序属性 $\boldsymbol{v}'_k$ 和无序属性 $\boldsymbol{v}''_k$ 组成的混合属性向量, $\boldsymbol{v}''_k$ 迭代运算中以集合形式 v''_k 存在.

聚类中心有序部分迭代公式为

$$\boldsymbol{v}'_k=\frac{\sum\limits_{i=1}^{n}(\boldsymbol{x}'_iu_{ik}^2)}{\sum\limits_{i=1}^{n}u_{ik}^2} \tag{15}$$

v''_k 的运算较为复杂. 假设某次迭代后, 属于第 k 个聚类中心的特征向量有 s 个, 且分别为 $x_{z_1},x_{z_2},\cdots,x_{z_s}$, 式中 $z_1,z_2,\cdots,z_s$ 代表向量的真实下标, 记 $S=\sum_{d=1}^{s}\{\bigcup_{r=m+1}^{l}x_{z_d,r}\}$, 在本节中, $\bigcup$ 表示向量所有维度或集合所有元素全部放在一个向量或集合中, 而并非求并集, 对向量集 S 中的元素进行降序排列, 并取分段中位数组合成新的无序属性部分的聚类中心 v''_k. 其中分段中位数求取公式如下:

当 s 为奇数时, 令

$$z=\frac{1+s}{2}$$

则

$$v''_k = \bigcup_{r=1}^{l-m} S_{z+(r-1)\times s} \tag{16}$$

当 s 为偶数时, 令

$$z = \frac{s}{2}$$

则

$$v''_k = \bigcup_{r=1}^{l-m} \frac{S_{z+(r-1)\times s} + S_{z+1+(r-1)\times s}}{2} \tag{17}$$

定义 $R(\cdot)$ 为将集合转换成一维行向量的运算, 则 $\boldsymbol{v}''_k = R(v''_k)$, 由于无序属性部分顺序无关, 故 $\boldsymbol{v}''_k$ 的形式并不唯一, 取其中一种形式, 与有序属性部分聚类中心联合, 最终求得 $\boldsymbol{v}_k = [\boldsymbol{v}'_k\ \boldsymbol{v}''_k]$. 即改进的 FCM 算法中的聚类中心每次更新是由有序部分更新结果与无序部分更新结果共同构成.

以下为改进半监督 FCM 的算法描述.

算法 2. 改进的半监督 FCM 算法

输入. 标记样本和未标记样本.

输出. 聚类中心及未标记样本的隶属度矩阵.

步骤 1. 将标记样本和未标记样本进行筛选及降维预处理;

步骤 2. 利用 FCM 算法对标记样本进行预聚类;

步骤 3. 利用步骤 2 所得聚类中心对未标记样本做如下操作: 采用改进距离度量函数计算未标记样本与聚类中心的距离, 选择最靠近第 i 个聚类中心的未标记样本并贴上标签 i, 加入到标记样本中, 并从未标记样本中删除;

步骤 4. 计算各个样本点到聚类中心的距离矩阵, 并得到新的隶属度矩阵 (若分母为 0, 则令 $u_{ik} = 1$);

步骤 5. 对最新获得的标记样本进行重聚类处理, 计算目标函数值 J. 迭代至 J 小于给定的阈值 δ 或与上次循环产生的目标函数值之差小于阈值 δ 则算法停止.

3　仿真与分析

3.1　训练数据及验证数据的获取

本文所提算法主要针对包含有序和无序属性的混合属性数据集的聚类方法, 为验证聚类算法的聚类精度, 选用滚动轴承多种工况下的振动信号进行预处理和时频分析[12], 并提取相应特征值构成训练数据和测试数据.

在轴承运行过程中, 当内滚道发生剥落、裂纹、点蚀等损伤时, 会产生一定频率的冲击振动, 轴承外圈亦是同理, 当滚动体产生损伤时, 缺陷部位通过内圈或外圈滚道表面时, 也会产生一定频率的冲击振动, 现实中的滚动轴承的振动信号, 主要通过安放在轴承座上的传感器测取设备获得, 测得的信号是包含若干成分的混合. 损伤故障大致可以分为两类: 1) 可以从转速和轴承的几何尺寸求得的通过频率, 又称为故障特征频率. 2) 由于损伤冲击作用诱发的轴承系统的高频固有振动成分. 若不考虑机械系统的非线性因素, 近似构造出包含轴系和轴承的复合振动信号数学模型如下[13]:

$$\begin{cases} x(t) = x_1(t) + x_2(t) + n(t) \\ x_1(t) = \sum\limits_i a_i \cos 2\pi f_i t + \sum\limits_j b_j \cos 2\pi f_j t \\ x_2(t) = \sum\limits_k A_k[1 + b_{k,j}(t)] \cos 2\pi f_{k,gz} t \end{cases} \tag{18}$$

其中, $x(t)$ 为加速度传感器采集的轴承座综合振动信号; $x_1(t)$ 为与轴转频和轴承各元件通过频率相关的低频振动信号; a_i 为与轴转频相关的第 i 个低频振动信号分量的幅值; f_i 为频率; b_j 为滚动轴承故障隐患所引起的第 j 个低频振动信号分量的幅值; f_j 为滚动轴承元件的故障通过频率; $x_2(t)$ 为以固有频率为载波频率, 以滚动轴承通过频率为调制频率的调制信号; $b_{k,j}(t)$ 为滚动轴承第 k 个调制信号, 其调制频率为滚动轴承的各元件的通过频率; $f_{k,gz}$ 为载波频率, 是各零部件的固有频率; $n(t)$ 为 $x(t)$ 中的噪声分量.

由某故障轴承的结构参数计算得到转速为 1 800 r/min 下的故障特征频率, 可知,

$$f_r = \frac{n}{60} = 30\,\text{Hz}$$

相应地, 各故障特征频率如表 1 所示.

表 1　轴承各部件故障特征频率 (Hz)

Table 1　Characteristic frequency of rolling bearings (Hz)

内圈	外圈	保持架	滚动体
163.2	107.4	11.9	141.2

将以上四种故障频率分别作为单故障振动信号的频率, 忽略机械系统的非线性因素, 近似构造出包含轴系和轴承的复合振动信号.

对复合振动信号进行特征提取, 并构造混合属性向量, 特征向量中有序属性部分包含最大值、最小值、峭度值、均值标准差 5 个指标, 无序属性部分的构建主要是通过对复合振动信号进行经验模态分解 (Empirical mode decomposition, EMD)[14], 得

到若干本征模函数 (Intrinsic mode function, IMF) 分量, 再进行希尔伯特变换, 进而求得特征频率值而获得. 对于构造的外圈故障和滚动体故障复合振动信号进行 EMD 分解, 最终得到 8 组本征模函数分量及对应频谱图, 如图 1 所示.

3.2 测试实验 1

实验部分选取五种故障 (各取 50 组), 进行聚类处理及分析. 五种故障包括内圈故障、外圈故障、滚动体故障三个单故障及内外圈、滚动体外圈两种复合故障. 聚类结果采用聚类精度均值来衡量, 即每个簇中占比最高的对象所占的比例的平均值.

轴承的混合属性特征向量中有序属性与无序属性数值差异性较大, 图 2 (a) 和图 2 (b) 分别为未标准化数据及标准化数据的预聚类结果.

从图 2 可以看出, 未标准化数据对预聚类的正确率影响较明显, 标准化数据预聚类正确率更高. 预聚类所得聚类中心对最终聚类结果正确率有直接影响, 故本文预聚类前对于原始数据做标准化的预处理.

图 3 (a) 为 FCM 重聚类结果, 相同分组用实线相连, 纵坐标为数据点实际组别, 聚类实验结果用实线相连. 可以看出传统半监督 FCM 聚类算法单故障聚类结果较理想, 聚类不纯度较低, 但耦合故障聚类的实验结果与实际组别交叉严重, 聚类结果不理想. 图 3 (b) 为改进 FCM 重聚类结果图, 与传统半监督 FCM 聚类结果相比, 耦合故障的聚类精度明显提高, 详细结果如表 2 所示. 图 4 (a) 和图 4 (b) 为两种聚类算法聚类结果的柱状统计图 (柱状图坐标分别为: x: 实验结果组别号, y: 实际组别号, z: 统计数).

表 2 聚类精度对比表

Table 2 Comparison table of clustering accuracy

	单故障	耦合故障
传统 FCM 聚类精度	0.98	0.65
改进 FCM 聚类精度	1.00	0.87

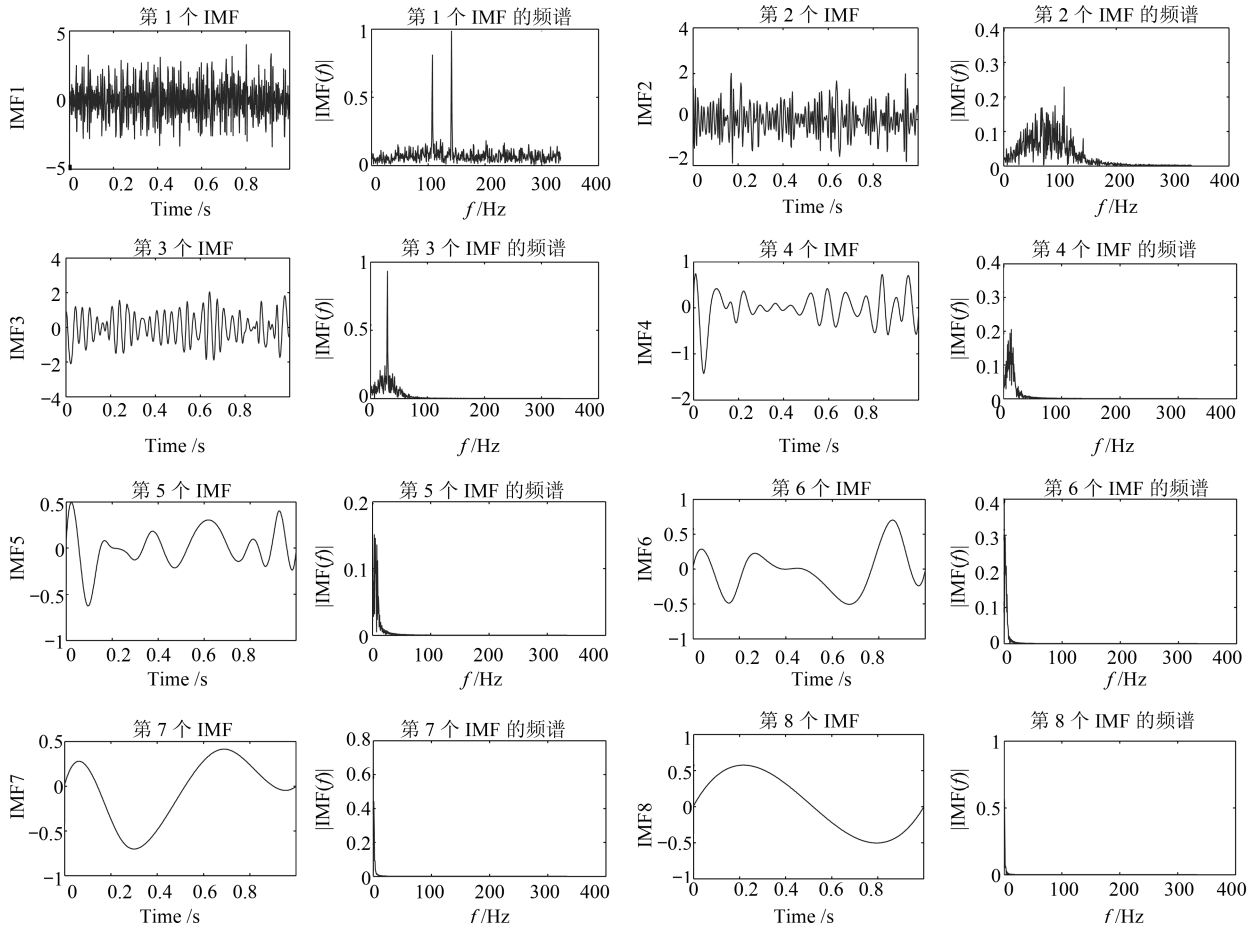
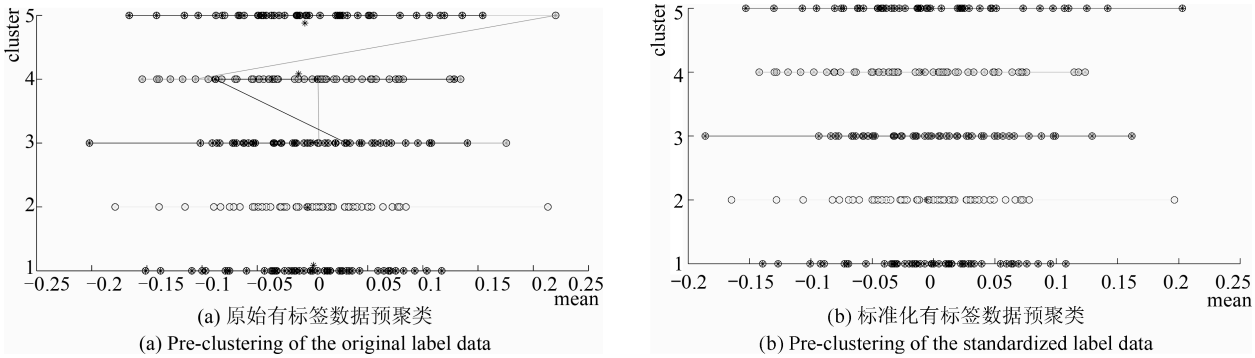


图 1 复合振动信号 EMD 分解

Fig. 1 The EMD decomposition of complex vibration signals

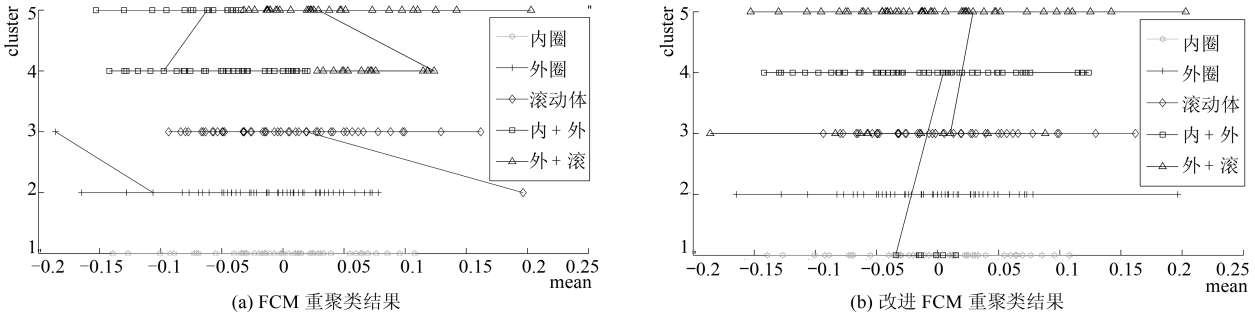


(a) 原始有标签数据预聚类
(a) Pre-clustering of the original label data

(b) 标准化有标签数据预聚类
(b) Pre-clustering of the standardized label data

图 2　有标签数据预聚类

Fig. 2　Pre-clustering of the label data

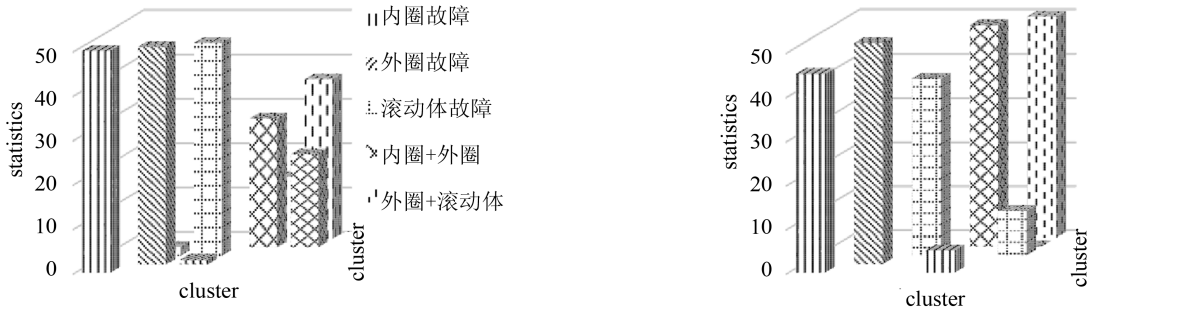


(a) FCM 重聚类结果
(a) Result of re-clustering result by FCM algorithm

(b) 改进 FCM 重聚类结果
(b) Result of re-clustering result by improved FCM algorithm

图 3　重聚类结果

Fig. 3　Re-clustering result



(a) FCM 重聚类结果柱状统计图
(a) Bar chart of re-clustering result by FCM algorithm

(b) 改进 FCM 重聚类结果柱状统计图
(b) Bar chart of re-clustering result by improved FCM algorithm

图 4　重聚类结果柱状统计图

Fig. 4　Bar chart of re-clustering result

经计算可得, 欧氏距离作为距离度量函数所得试验结果的聚类精度为 0.848, 改进的混合属性距离度量函数所得试验结果的聚类精度为 0.94.

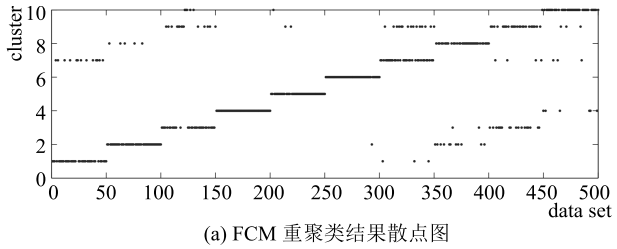
表 2 为 FCM 聚类算法改进前及改进后在单故障及复合故障聚类中的精度对比. 从表 2 可以看出, 在本实验部分, 复合故障之间的干扰对传统 FCM 聚类精度有较大影响, 改进的混合属性距离作为距离度量函数在耦合故障诊断方面具有显著优势.

3.3　测试实验 2

实验选取 4 组单故障及 6 组耦合故障的特征数据集 (每组 50 个向量) 进行聚类处理, 此时故障类型较多, 复合故障之间干扰较强, 传统 FCM 的聚类精度急剧下降, 实验结果部分添加了混合属性聚类的 K-prototypes 方法作为对比.

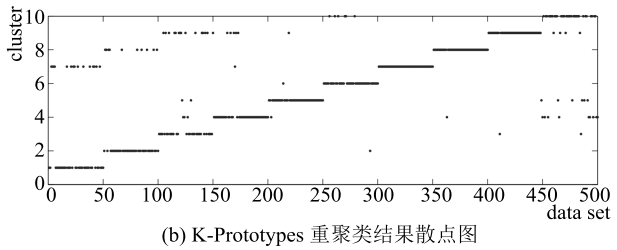
重聚类结果散点图如图 5 所示, 图 5 (a) 为传统半监督 FCM 聚类的结果, 图 5 (b) 为 K-prototypes

聚类的结果, 图 5 (c) 为改进半监督 FCM 聚类的结果, 纵坐标代表类别, 试验数据共有 10 种故障, 每种故障 50 组数据, 并分别加上类别属性编号 1～10, 故图中横坐标 1～50, 51～100, ···, 451～500 的实际类别应该依次对应 1～10 类, 图中的散点分布为聚类方法所得的结果, 图 5 (a) 图中横坐标 1～50 的区间, 有若干点纵坐标为 7, 横坐标 300～350 的区间, 有若干点纵坐标为 9, 这些都是实际结果与实验结果不相符的情况. 三种聚类算法的柱状统计图如图 6 所示.



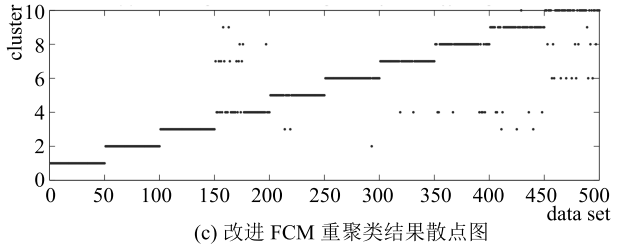
(a) FCM 重聚类结果散点图

(a) Scatter diagram of re-clustering result by FCM algorithm



(b) K-Prototypes 重聚类结果散点图

(b) Scatter diagram of re-clustering result by K-Prototypes algorithm



(c) 改进 FCM 重聚类结果散点图

(c) Scatter diagram of re-clustering result by improved FCM algorithm

图 5 重聚类结果散点图

Fig. 5 Scatter diagram of re-clustering result

由正确率柱状图对比可知, 当故障类型较多时, 改进 FCM 重聚类的聚类效果最好, K-prototypes 次之, 传统 FCM 重聚类的聚类效果较差, 三种算法的聚类精度如表 3 所示.

表 3 三种算法聚类精度对比表

Table 3 Comparison table of clustering accuracy by three algorithms

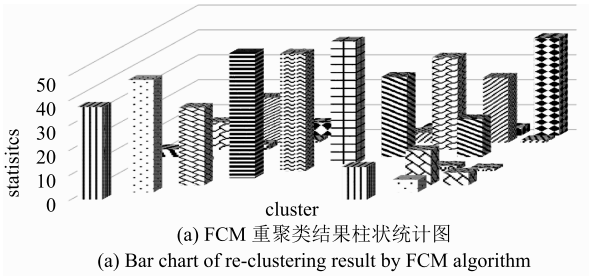
	传统 FCM	K-prototypes	改进 FCM
聚类精度	0.786	0.842	0.902

表 4 为在改进 FCM 中不同相异度阈值 ε 下的聚类精度对比表.

表 4 不同 ε 值下聚类精度对比表

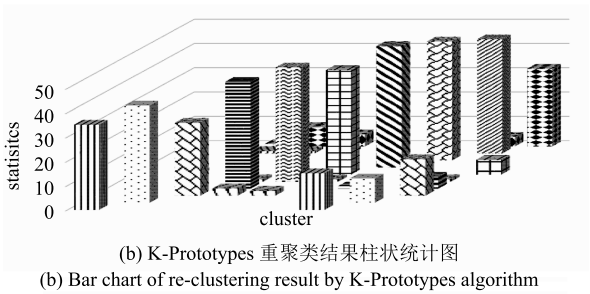
Table 4 Comparison table of clustering accuracy by different ε

ε	0.09	0.10	0.11	0.12	0.13	0.14
聚类精度	0.796	0.868	0.898	0.902	0.88	0.822



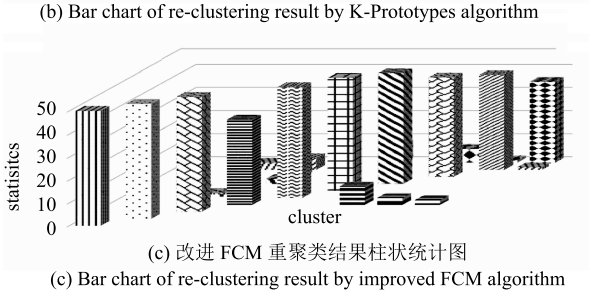
(a) FCM 重聚类结果柱状统计图

(a) Bar chart of re-clustering result by FCM algorithm



(b) K-Prototypes 重聚类结果柱状统计图

(b) Bar chart of re-clustering result by K-Prototypes algorithm



(c) 改进 FCM 重聚类结果柱状统计图

(c) Bar chart of re-clustering result by improved FCM algorithm

图 6 重聚类结果柱状统计图

Fig. 6 Bar chart of re-clustering result

考虑到噪音对低频信号有较大干扰, 对 4, 7, 8, 9 故障聚类结果进行分析, 并对无序属性部分距离度量计算时的相异度阈值 ε 采用自适应阈值调整, 自适应阈值调整公式如下:

$$\varepsilon = \varepsilon_0 + \frac{f_{\max} - f}{f_{\max}} \times \gamma$$

由表 4 可知, $\varepsilon = 0.12$ 时, 聚类效果最好, 因此基准值 ε_0 取 0.12, 在特定区间内, 相异度阈值越高则低频信号聚类精度越高, 超过一定区间则会导致高频信号的错归类, 进而影响聚类精度. 根据式 (6), 结合本文实验算例, 可知最易错归类的相异度阈值为 0.125, 故乘数因子 γ 取值 0.005. 式中 $f_{\max}$ 取

值 163.2, f 为计算 Jaccard 距离的两个数的平均值. 根据以上参数设置, 得到最终结果如图 7 和图 8 所示.

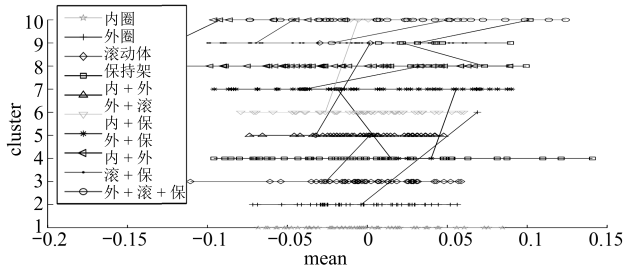


图 7　改进 FCM 自适应阈值调整后重聚类结果

Fig. 7　Re-clustering result by improved FCM algorithm after adaptive threshold

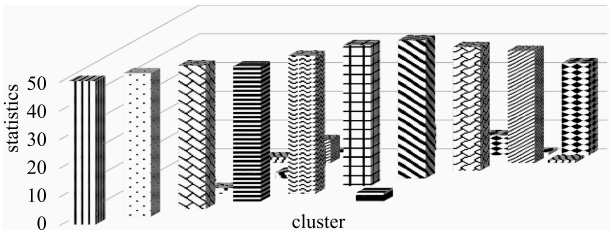


图 8　改进 FCM 自适应阈值调整后重聚类结果柱状统计图

Fig. 8　Bar chart of re-clustering result by improved FCM algorithm after adaptive threshold

将图 8 与图 6 (c) 对比, 可知混合属性距离度量公式进行自适应阈值调整后, 聚类精度进一步提升, 由聚类精度计算公式求得, 聚类精度提升至 0.912.

4　结束语

本文提出一种基于改进 Jaccard 距离的混合属性距离度量方法, 并运用于半监督 FCM 聚类算法中, 得到改进的半监督 FCM 算法, 将在数值属性数据集的聚类方法扩展到了混合属性数据集的聚类问题中. 通过对聚类算法的聚类精度这一指标值进行比较, 证明了改进的半监督 FCM 算法在聚类效果方面有了显著提升, 并得到如下结论.

1) 传统半监督 FCM 算法将样本不同特征量赋予相同的权重, 忽略了不同属性特征量本身的相异性, K-prototypes 算法作为混合属性聚类算法, 对分类属性采用匹配差异度的距离度量方法, 但是和广义的 Jaccard 距离有相同的弊端, 即向量维度对计算结果有很大影响, 处理含无序属性的混合属性数据集时, 精度较低. 改进半监督 FCM 聚类在处理含无序属性的混合属性数据集的聚类问题时, 采用欧氏距离与改进的 Jaccard 相结合的距离度量方式, 聚类精度明显优于传统的半监督 FCM 聚类和 K-prototypes 聚类.

2) 当聚类中心较多时 (对应试验中故障类型较多), 对于改进半监督 FCM, 相异度阈值 ε 可采用自适应阈值调整, 即对于无序属性部分自适应改变 ε 的值, 聚类精度得到提高.

半监督聚类的标记样本数据必须满足每个簇都至少有一个样本被标记出, 且初始样本数据对聚类结果影响较大. 换而言之, 半监督聚类算法是建立在对标记样本完全信任的基础上的. 因此, 如何提高算法对于不均衡数据集的聚类精度问题需要进一步研究. 另外, 将轨迹坐标值作为无序属性分量, 并将本文提出算法与时间翘曲距离结合, 对轴心轨迹进行相似性判断并聚类, 也是下一步工作的重点.

References

1 Xu Ming-Liang, Wang Shi-Tong, Hang Wen-Long. A semi-supervised affinity propagation clustering method with homogeneity constraint. *Acta Automatica Sinica*, 2016, **42**(2): 255−269
(徐明亮, 王士同, 杭文龙. 一种基于同类约束的半监督近邻反射传播聚类方法. 自动化学报, 2016, **42**(2): 255−269)

2 Zhao Hui-Zhen, Liu Fu-Xian, Li Long-Yue. Novel collaboration fuzzy C-means algorithm with K-nearest neighbor method determined Collaboration Coefficient. *Computer Engineering and Applications*, 2016, **52**(19): 19−24
(赵慧珍, 刘付显, 李龙跃. K- 近邻估计协同系数的协同模糊 C 均值算法. 计算机工程与应用, 2016, **52**(19): 19−24)

3 Huang Z X. Clustering large data sets with mixed numeric and categorical values. In: Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore, Singapore: PAKDD, 1997. 21−34

4 Chen Jin-Yin, He Hui-Hao. Density-based clustering algorithm for numerical and categorical data with mixed distance measure methods. *Control Theory and Applications*, 2015, **32**(8): 993−1002
(陈晋音, 何辉豪. 基于密度和混合距离度量方法的混合属性数据聚类研究. 控制理论与应用, 2015, **32**(8): 993−1002)

5 Huang De-Cai, Li Xiao-Chang. Incremental relative density-based clustering algorithm for mixture data sets. *Control and Decision*, 2013, **28**(6): 815−822
(黄德才, 李晓畅. 基于相对密度的混合属性数据增量聚类算法. 控制与决策, 2013, **28**(6): 815−822)

6 Chen Jin-Yin, He Hui-Hao. Research on density-based clustering algorithm for mixed data with determine cluster centers automatically. *Acta Automatica Sinica*, 2015, **41**(10): 1798−1813
(陈晋音, 何辉豪. 基于密度的聚类中心自动确定的混合属性数据聚类算法研究. 自动化学报, 2015, **41**(10): 1798−1813)

7 Chen Xin-Quan. Dual clustering method of mixed data set. *Computer Engineering and Science*, 2013, **35**(2): 127−132
(陈新泉. 面向混合属性数据集的双重聚类方法. 计算机工程与科学, 2013, **35**(2): 127−132)

8 Gardner A, Kanno J, Duncan C A, Selmic R. Measuring distance between unordered sets of different sizes. In: Proceedings of the 2014 IEEE Conference on Computer Vision and

Pattern Recognition (CVPR). Columbus, OH, USA: IEEE, 2014. 137−143

9 Li Cheng-Liang, Ma Yun, Zhang Rui, Wei Wei. Rotor system fault diagnosis based on semi-supervised spectrum kernel clustering. *Journal of Vibration, Measurement and Diagnosis*, 2016, **36**(3): 562−567
(李城梁, 马芸, 张锐, 魏伟. 基于半监督谱核聚类的转子系统故障诊断. 振动、测试与诊断, 2016, **36**(3): 562−567)

10 Ji Wei-Hua, Lv Guo-Fang. Conflicting evidence combination method based on generalized Jaccard coefficient. *Control Engineering of China*, 2015, **22**(1): 98−101
(嵇威华, 吕国芳. 基于广义 Jaccard 系数处理冲突证据方法. 控制工程, 2015, **22**(1): 98−101)

11 Zhou Chen-Xi, Liang Xun, Qi Jin-Shan. A semi-supervised agglomerative hierarchical clustering method based on dynamically updating constraints. *Acta Automatica Sinica*, 2015, **41**(7): 1253−1263
(周晨曦, 梁循, 齐金山. 基于约束动态更新的半监督层次聚类算法. 自动化学报, 2015, **41**(7): 1253−1263)

12 Yuan Jie, Wang Fu-Li, Wang Shu, Zhao Lu-Ping. A fault diagnosis approach by D-S fusion theory and hybrid expert knowledge system. *Acta Automatica Sinica*, 2017, **43**(9): 1580−1587
(袁杰, 王福利, 王姝, 赵露平. 基于 D-S 融合的混合专家知识系统故障诊断方法. 自动化学报, 2017, **43**(9): 1580−1587)

13 Zhang Chao, Chen Jian-Jun, Guo Xun. Complex fault diagnosis for rotor systems using the second generation wavelet and extremum field mean mode decomposition. *Journal of Vibration, Measurement and Diagnosis*, 2011, **31**(1): 98−103
(张超, 陈建军, 郭迅. 基于第 2 代小波和 EMMD 的转子系统复合故障诊断. 振动、测试与诊断, 2011, **31**(1): 98−103)

14 Hao H, Wang H L, Rehman N U. A joint framework for multivariate signal denoising using multivariate empirical mode decomposition. *Signal Processing*, 2017, **135**: 263−273

李晓庆 合肥工业大学电气与自动化工程学院博士研究生. 2013 年获得合肥工业大学学士学位. 主要研究方向为故障预测及健康管理.
E-mail: lixiaoqing@mail.hfut.edu.cn
(**LI Xiao-Qing** Ph. D. candidate at the School of Electrical Engineering and Automation, Hefei University of Technology. She received her bachelor degree from Hefei University of Technology in 2013. Her research interest covers prognostic and health management.)

唐 昊 合肥工业大学电气与自动化工程学院教授. 2002 年获得中国科学技术大学博士学位. 主要研究方向为离散事件动态系统, 随机决策与优化理论, 智能优化与控制方法. 本文通信作者.
E-mail: htang@hfut.edu.cn
(**TANG Hao** Professor at the School of Electrical Engineering and Automation, Hefei University of Technology. He received his Ph. D. degree from University of Science and Technology of China in 2002. His research interest covers discrete event dynamic system, stochastic decision and optimization theory, intelligent optimization and control method. Corresponding author of this paper.)

司加胜 合肥工业大学智能制造技术研究院硕士研究生. 2015 年获得东北大学学士学位. 主要研究方向为故障预测与健康管理.
E-mail: jasenchn@hotmail.com
(**SI Jia-Sheng** Master student at the Intelligent Manufacturing Institute, Hefei University of Technology. He received his bachelor degree from Northeastern University in 2015. His research interest covers prognostic and health management.)

苗刚中 合肥工业大学电气与自动化工程学院副教授. 1991 年获合肥工业大学工程硕士学位. 主要研究方向为电工与电子技术, 物联网相关技术, 数据挖掘, 移动手机软件开发.
E-mail: miaogzh@126.com
(**MIAO Gang-Zhong** Associate professor at the School of Electrical Engineering and Automation, Hefei University of Technology. He received his master degree from Hefei University of Technology in 1991. His research interest covers electrical and electronic, the internet of things, data mining, and software development about mobile phone.)