信息与控制　2020 年　第 49 卷　第 4 期：455 ~ 463

DOI：10.13976/j.cnki.xk.2020.9359　　文章编号：1002 - 0411(2020) - 04 - 0455 - 09

基于 D-S 证据理论的不完整数据混合分类算法

段中兴[1,2]，毕瀚元[1]，张作伟[3]

1. 西安建筑科技大学信息与控制工程学院，陕西 西安　710055；2. 西部绿色建筑国家重点实验室，陕西 西安　710055；
3. 西北工业大学自动化学院，陕西 西安　710072

基金项目：国家自然科学基金资助项目(51678470)

通信作者：段中兴，zhx_duan@163.com　收稿/录用/修回：2019 - 07 - 01/2019 - 10 - 12/2020 - 05 - 30

摘要

针对传统不完整数据插补聚类算法未考虑插补值对类中心的影响以及不完整样本建模带来的不确定性等问题，提出了一种基于 D-S 证据理论的不完整数据混合分类算法. 首先，利用经典软聚类算法对数据集中的完整样本进行聚类并选择训练样本，再根据剩余样本已知属性构建若干训练集，并利用基础分类器分类；然后在 D-S 证据理论下，将属于若干个类别概率相近的样本划分到相应复合类以降低误分类率；最后，对处于复合类中的不完整样本，分别在构成其复合类的单类中进行 K 近邻插补并分类，将若干个分类结果自适应融合以决定这些样本的最终类别. 模拟数据集和 UCI 数据集验证表明，算法能够合理地表征由缺失值引起的不确定性，降低了误分率.

关键词

不完整数据
聚类
D-S 证据理论
不确定性
多源信息融合

中图法分类号：TP181

文献标识码：A

A D-S Evidence Reasoning Based Hybrid Classification Algorithm for Incomplete Data

DUAN Zhongxing[1,2], BI Hanyuan[1], ZHANG Zuowei[3]

1. *School of Information & Control Engineering*, *Xi'an University of Architecture and Technology*, *Xi'an* 710055, *China*;
2. *State Key Laboratory of Green Building in Western China*, *Xi'an* 710055, *China*;
3. *School of Automation*, *Northwestern Polytechnical University*, *Xi'an* 710072, *China*.

Abstract

To address the problems of the traditional incomplete data imputation clustering algorithm, which does not consider the influence of imputation on the class center and the uncertainty caused by incomplete sample modeling, a hybrid classification algorithm for incomplete data based on the D-S evidence theory (HCA) is proposed. First, the classical soft clustering algorithm is used to cluster the complete samples in the dataset and select the training samples. Then, several training sets are constructed on the basis of the known attributes of the remaining samples, and the basic classifiers are used to classify them. Under the D-S evidence theory, samples belonging to several classes with similar probability are divided into corresponding metaclasses to reduce the misclassification rate. Finally, the incomplete samples in the metaclasses are classified after imputing by K-nearest neighbor to their hard-to-distinguish classes, and several classification results are adaptively fused to determine the final class of these samples. The validation of the simulated datasets and UCI standard datasets show that the algorithm can reasonably represent the uncertainty caused by missing values and reduce the error rate.

Keywords

incomplete data;
clustering;
D-S evidence theory;
uncertainty;
multi-source information fusion

0　引言

在实际应用中，由于多种原因导致数据集出现了大量的不完整样本(incomplete data)[1]使得人们无法直接使用一些已经取得广泛应用的经典聚类算法[2-3]. Pedro 等[4]根据数据的缺失机制，将缺失数据分为三种：完全随机缺失(missing completely at random, MCAR)、随机缺失(missing at random, MAR)和非随机缺失(not missing at random,

NMAR). 其中, 完全随机缺失是指缺失值既与自身值无关也与其它任何值无关, 例如传感器设备失灵等; 随机缺失是指缺失值与自身值无关, 但可以根据数据库中其它样本进行预测, 例如调查问卷信息不完整; 非随机缺失是指缺失值与自身值有关, 例如传感器无法测量超出量程以外的值而导致的数据缺失. 对于完全随机缺失和随机缺失两种缺失机制, 在处理缺失数据时可以忽略对数据缺失原因的分析, 而处理非随机缺失时要分析数据缺失的原因后对缺失数据进行处理. 本文对完全随机缺失和随机缺失两种情况进行了研究. 根据这两种缺失机制, 人们提出了许多缺失数据(不完整样本)预处理方法[4]使经典的聚类方法得到更加广泛的应用.

预处理方法中最简单是删除法[5], 直接删除带有缺失值的样本或者删除带有缺失值的属性, 但这会导致许多有效信息被浪费且仅适用于不完整样本在数据集中占比较小的情况(一般情况下小于5%). 目前插补策略是对不完整样本缺失值进行预处理的主流方法, 主要分为基于统计分析的插补方法和基于机器学习的插补方法. 在基于统计分析的插补方法中, 主要有均值插补(MI)[6]、回归插补[7]、冷热平台插补[8]和多重插补[9]. 均值插补[6]利用所有样本与缺失值同一属性的均值来代替缺失值或者利用和不完整样本属于同一类样本中与缺失值同一属性的均值来代替缺失值; 回归插补[7]适用于缺失值的属性与完整样本中的属性值特别相关的情况; 热平台插补[8]是利用与含有缺失值样本最相似的一个完整样本来对缺失值进行估计, 它的缺点是只基于数据集中的单个完整样本对缺失的插补而忽略了属性的全局性, 与热平台插补类似, 冷平台插补使用的数据源必须与当前数据集不同. 以上三种方法都只为缺失值提供一个确定的值, 这样不能反映预测值的不确定性. 在多重插补中[9], 缺失值通过一个合适的模型被估计 M 次得到 M 个估计值, 然后将多次填补后的多个数据集进行分析合并得到最终的估计值. 基于机器学习的插补方法主要有 K-NN 插补(KNNI)[10], 利用与带有缺失值的样本最相似的 K 个完整样本来对缺失值进行估计; FCM 插补(FCMI)[11], 根据 FCM 产生的聚类中心, 通过含有缺失值的样本与中心之间的距离对缺失值进行估计; Samad 和 Harp[12]将自组织映射应用到缺失数据插补中, 可以将高维的输入数据映射到低维空间, 并通过映射节点的向量加权运算得到缺失值的估计值; Huang 等[13]在医学领域利用多层感知器对缺失数据进行插补, 使用一个基本的多层感知器利用完整样本训练一个回归模型, 每个不完整的属性都可以通过剩余完整样本的属性学习得到; 在文[14]中 Kim 等利用循环神经网络对缺失数据进行插补, 使用一种来自单元的反馈连接结构, 利用其反馈值来估计缺失值; Li 等[15]利用对抗神经网络创建一个不完整数据学习框架并对缺失数据进行插补.

随着缺失值估计方法研究的深入, 估计方法可以对缺失值产生一个估计值, 但不同的估计方法往往会产生不同的结果, 如果直接对带有估计值的数据进行聚类, 必然会改变不同类别的原始类中心继而对完整样本的准确聚类产生消极影响; 其次, 估计策略也会带来不确定问题, 因为缺失值有可能是对样本类别判定起关键作用的属性, 也有可能对样本划分结果产生很小甚至没有影响. 但是当缺失的属性极其重要时, 如果不对缺失值进行估计或者仅仅采取单一估计策略是不合理的. 例如样本 $\boldsymbol{x}_1=[4, 5, 1]$ 属于类别 ω_1, 样本 $\boldsymbol{x}_2=[4, 5, 2]$ 属于类别 ω_2, 此时对样本 $\boldsymbol{x}_3=[4, 5, ?]$ 进行单一估计有可能带来不准确, 因为对缺失值进行估计的样本极有可能来自不同类别; 第三, 无论采取何种估计策略将不完整样本强行划分给某个特定的单类都会增加样本被误分的风险, 特别是当该样本属于不同类别的概率没有明显差异时. 此时应该采取一种更为保险的策略来合理地表征这种不精确性. D-S 证据理论因其将传统辨识框架 Ω 扩展到幂集 2^{Ω} 中并能够通过复合类有效地表征由于多种原因导致的数据不确定性和不精确性问题而被广泛应用到各个领域[16]. 因此, 本文对不完整数据的研究工作基于 D-S 证据理论展开.

D-S 证据理论经过多年的发展, 已经在数据聚类[3]、数据分类[17]和决策融合[18]等领域得到了广泛的应用. Denœux 等[3]于 2008 年提出了一种针对目标数据的证据 c 均值(evidential c-means, ECM)聚类算法, 它被认为是 FCM 算法[2]和 NC 算法[19]在 D-S 证据理论框架下的扩展, ECM 的聚类结果中包含三个类别: 单类、复合类和噪声类. 其中复合类可以有效表征因为多种原因造成的样本类别的不确定性. 但是该方法在计算过程中仅将样本到单类或者复合类的类中心的距离作为获取信任值的依据, 所以当复合类和单类类中心相近的情况下往往会产生错误的聚类结果. 因此刘准钆等[20]针对 ECM 的缺点提出一种新的改进方法, 信任 c 均值(credal c-means, CCM), 认为样本到复合类的距离不仅仅与样本到复合类中心的距离有关还与样本到复合类所包含的单类的类中心的距离有关, 解决了 ECM 方法没有考虑到噪声的缺点. 文[21]提出了一种改进的证据 c 均值方法, 它将遗传算法与 ECM 相结合, 利用遗传算法具有良好的全局搜索能力, 克服了 ECM 结果局部最优的缺点. 虽然这些基于 D-S 证据理论的聚类方法能够合理地表征数据的不精确性, 但是没有考虑到数据缺失的问题, 也没有考虑到因为数据不完整造成的不精确问题.

针对这些在不完整数据聚类中可能遇到的问题, 本文提出了一种新的基于 D-S 证据理论的不完整数据混合分类算法(HCA). 为了避免估计值有可能改变数据原始类中心进而影响完整数据的聚类效果, 首先使用经典软聚类算法(例如 fuzzy c-means)对数据集中的完整样本进行聚类, 根据样本属于不同类别的概率选择一部分可靠样本作为训练集. 其次, 将剩余的完整样本和不完整样本作为测试集, 根据测试集中样本的现有属性利用训练集分别得到多个训练集, 并通过训练基础分类器(例如支持向量机[22]等)来对相应的样本进行分类; 然后采用一种模糊策略将类别难以区分的样本划分到相应的复合类以降低分类错误率; 最后, 对于复合类中的不完整样本, 在他们可能的类别中分别对缺失值进行 K 近邻插补, 得到多个带有估计

值的完整样本，并利用一种改进的 D-S 融合规则对多个样本的分类结果进行自适应融合以共同决定该不完整样本的最终类别. 该方法能够通过将那些难以被准确分类到某个单类的(带有估计值的)样本划分到特定的复合类用来表征由于缺失值引起的不确定性和不精确性，同时降低误分率.

1　相关工作

1.1　D-S 证据理论

D-S 证据理论(evidence reasoning)[23]是由 Dempster 在 1967 年最先提出的，经过 Shafer 的推广并在 1976 年形成证据推理理论，因此又称为 Dempster-Shafer 理论(Dempster-Shafer Theory，DST). D-S 证据理论对数据信息的不准确和矛盾等问题提供了显式估计，并且能够有效处理类别混合问题. D-S 证据理论将传统的辨识框架 Ω 扩展到幂集 2^{Ω}，延伸出复合类的概念，在 D-S 证据理论下，不仅可以将样本划分给单个类别，也可以将样本划分给 2^{Ω} 中的任何一个子集，拓展了硬划分和模糊划分的现有概念，这种灵活的划分方法能够使人们更加深入得了解数据，并提高对数据中异常值的鲁棒性.

在 Shafer 模型中，定义了一个元素集合 $\Omega=\{\omega_1, \cdots, \omega_c\}$，$\Omega$ 的所有子集构成的集合为 Ω 的幂集，表示为 2^{Ω}. 假设 $\Omega=\{\omega_1, \omega_2, \omega_3\}$，那么它的幂集 $2^{\Omega}=\{\phi, \omega_1, \omega_2, \omega_3, \omega_1\cup\omega_2, \omega_1\cup\omega_3, \omega_2\cup\omega_3, \Omega\}$，其中单个元素(例如，$\omega_i$，$i=1, \cdots, c$)代表一个特定的单类，同时定义 $\omega_i\cup\omega_j\triangleq\omega_{i,j}$($i, j=1, \cdots, c$)代表复合类，被划分到复合类 $\omega_{i,j}$ 中的样本的真实类别可能属于类别 ω_i，也可能属于类别 ω_j，表达一种局部未知. 一个证据的基本信任指派(basic belief assignment，BBA)就是从 2^{Ω} 到[0，1]上的一个映射函数 $m(\cdot)$，并且满足以下条件：

$$\begin{cases}\sum\limits_{A\in 2^{\Omega}} m(A)=1\\ m(\phi)=0\end{cases} \tag{1}$$

Shafer 还在 BBA 的基础上定义了信任函数 Bel(·)和似然函数 Pl(·)：

$$\mathrm{Bel}(A)=\sum_{A,\ B\in 2^{\Omega};\ B\subseteq A} m(B) \tag{2}$$

$$\mathrm{Pl}(A)=\sum_{A,\ B\in 2^{\Omega};\ A\cap B\neq\varnothing} m(B)=1-\mathrm{Bel}(\bar{A}) \tag{3}$$

其中，$\bar{A}$ 表示元素 A 在集合 Ω 中的补集. Bel(·)表达事件属于 A 的最低可信度，称为信任函数，Pl(·)表达事件属于 A 的最大可能性，称为似然函数；在信任函数和似然函数中必然存在 $\mathrm{Bel}(A)\leqslant\mathrm{Pl}(A)$[23]. 由信任函数 Bel(·)和似然函数 Pl(·)组成的信任区间[Bel(·)，Pl(·)]能够表达对某个假设的确认程度.

来源于两个独立证据且在一个辨识框架下的基本信任函数 $m_1(\cdot)$ 和 $m_2(\cdot)$，可以利用 Dempster-Shafer 规则[24](DS 规则)进行融合，以获得融合后的基本信任函数：$m(\cdot)=[m_1\oplus m_2](\cdot)$，DS 融合规则为

$$m(A)=\begin{cases}\dfrac{\sum\limits_{B\cap C=A} m_1(B)m_2(C)}{1-K}, & A\neq\varnothing,\ B,\ C\in 2^{\Omega}\\ 0, & A=\varnothing\end{cases} \tag{4}$$

式中，K 为冲突因子，表示两个证据的冲突程度，定义为

$$K=\sum_{B\cap C=\varnothing} m_1(B)m_2(C) \tag{5}$$

D-S 理论中的 DS 融合规则对独立的证据进行融合具有以下优势：

1）可以将不同专家或者不同来源的信息进行综合，将多源信息进行融合，使得应用范围更加广泛；

2）在信息融合过程中，可以保留源信息所表达的不确定性，使得信息融合结果更加准确；

3）在表达信息融合结果时不但可以将结果划分到特定的单个元素构成的类别中，也可以将结果划分给由单个元素组成的类别，能够有效表达信息的不确定性并处理各个信息源之间的矛盾.

1.2　D-S 规则改进

D-S 规则将所有冲突信息 K 都按比例分配给其它集合. 只有当组合规则分母不为 0 时，即两个证据不完全冲突($K\neq1$)时，D-S 规则才有效. 但当两个证据冲突时($K=1$)，使用 D-S 规则对证据进行融合就会出现与常识相悖的情况，这是因为在归一化过程中组合规则将冲突信息完全忽略，在数学上引出不合常理的问题，所以 D-S 规则无法解决当两个证据冲突严重或者完全冲突的情况[24]. 为此，众多学者进行了研究，解决此问题的方法主要可以归纳为两大类：1）修改 D-S 合成规则. Yager 最早给出了对 D-S 规则的改进(即 Yager 规则)[25]，它重新将冲突信息全部划分给整个辨识框架，等待新的证据再做判断；孙全等[26]使用加权和平均的方法将冲突值重新分配；陈炜军等[27]引入证据距离函数的概念来计算权重，实现了对冲突值的分配. 2）修改证据源. 该方法认为问题在于证据本身，所以对证据源进行预处理，然后再使用 D-S 证据合成规则进行合成. Murphy 等[28]提出了一种对证据源进行平均然后多次合成的方法；邓勇等[29]利用证据距离函数计算证据的权重，提出了对证据源进行加权平均的方法. 卢正才等[30]将以上两类方法相结合，提出了一种新的处理冲突证据的方法. 本文采用 Yager 规则，证据组合规则为

$$m(A)=\begin{cases}\sum\limits_{B\cap C=A} m_1(B)m_2(C), & A\neq 2^{\Omega}\\ \sum\limits_{B\cap C=A} m_1(B)m_2(C)+k, & A=2^{\Omega}\end{cases} \tag{6}$$

其中，

$$k=\sum_{B\cap C=\varnothing} m_1(B)m_2(C) \tag{7}$$

2　基于 D-S 证据理论的不完整数据混合分类算法(HCA)

针对传统不完整数据插补聚类算法未考虑插补值对类中心的影响以及不完整样本建模带来的不确定性等问题，提出了一种基于 D-S 证据理论的不完整数据混合分类算法(HCA). 首先使用经典的软聚类方法对完整样本进行聚类

并选择一些可靠样本作为训练集；再根据剩余样本的现有属性，得到若干训练集来训练分类器并进行再分类；如果样本 $\boldsymbol{x}_i$ 对于不同类别的概率没有明显差异，在 D-S 证据理论框架下将样本划分到相应的复合类以降低误分率；在构成不完整样本所在复合类的单类中，分别对缺失值进行估计，得到多个完整样本，将得到的多个完整样本分类，并将多个分类结果自适应融合以共同决定这个不完整样本的最终类别. 算法将经典的软聚类方法与分类方法相结合，在 D-S 证据理论下解决不完整数据的无监督分类问题，解决了因为估计不完整样本而对类中心产生的影响. 在 D-S 理论框架下引入复合类的概念，在分类过程中允许样本属于几个单类构成的复合类，能够有效表达由于数据缺失和原始数据分布（不同类别数据在边缘产生重叠）产生的不精确性和不确定性，降低误分率. 对复合类中样本的缺失属性（可能是决定样本类别的关键属性）进行多估计，并基于 D-S 理论将分类结果自适应融合，以得到复合类中不完整样本更加精确的分类结果.

2.1 可靠样本的选择

考虑一个数据集 $X=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_l\}$，由两个部分组成：包含 m 个完整样本的数据集 $X_{\text{complete}}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_m\}$ 和 n 个含有缺失值样本的数据集 $X_{\text{missing}}=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$，$l=m+n$. 数据集 X_{complete} 中的样本先使用一些经典的软聚类方法对数据集进行聚类，在聚类结果中一些强烈支持属于一个类别的样本被选择为训练集，这些样本的聚类信息被认为是已经确定的，即可靠样本. 在无监督分类中，这些可靠样本及其标签信息可以被用来训练分类器. 这些软聚类方法包括概率框架下的聚类方法，例如 FCM[2]，NC（Noise-Clustering algorithm）[19] 和 D-S 证据理论框架下的 ECM[3]、CCM[20] 等方法. 对于概率框架下的软聚类方法，可以根据分类结果中样本对每一个类别的支持度来决定样本的可靠性，对于 D-S 证据理论下的软聚类方法，被划分到单类的样本可以被认为是可靠样本，被划分到复合类中的样本为不可靠样本. 本文使用 FCM 作为默认的软聚类方法，基于以下原因：①具有完善的理论基础；②算法流程简单；③实际应用最广泛. 通过 FCM，可以得到 X_{complete} 中 m 个样本的隶属度矩阵 $\boldsymbol{U}$，利用隶属度矩阵 $\boldsymbol{U}$ 可以获得可靠样本. 当一个完整样本 $\boldsymbol{x}_i$ 属于类别 ω_j 的隶属度值 $u_{i,j}$ 大于训练阈值 β 时，则认为该样本是可靠样本. 所有满足约束条件的样本都将被划分给训练集 X_{training}，约束条件表示如下：

$$X_{\text{training}}=\{\boldsymbol{x}_i\mid\boldsymbol{x}_i\in X_{\text{complete}},\ \max\{u_{i,j}\}\geqslant\beta\}\tag{8}$$

式中，β 为训练阈值. 由式（8）可知，训练集中样本的多少取决于训练阈值 β 的设定值. 在实际应用中，HCA 的训练阈值 β 可以根据需要在一定范围内进行调整. 理论上，只要训练阈值 $\beta>0.5$ 就可以认为样本的类别信息是确定的，即可靠样本. 当类中心距离较近时，如图 1 所示，相对较小的训练阈值（如 $\beta=0.7$）可能将处于不同类别重叠区域的样本划分到训练集中，降低训练集的可靠性和准确性，误导测试集样本的分类，而相对较大的训练阈值（如 $\beta=0.85$）可以在一定程度上避免这种情况，但如果训练阈值 β 取值过大（如 $\beta>0.95$），则训练集中只有少量样本，不能完整表达属于同一类别样本的属性特征从而降低训练集的普适性；当类中心距离较远时，如图 2 所示，即使训练阈值相对较小（如 $\beta=0.7$），训练集中也不会有重叠区域样本. 结合大量实验，本文确定的训练阈值范围为 $\beta\in[0.85,0.95]$.

数据集 X 中的剩余样本被划分给测试集，测试集包含属性完整的不可靠样本和不完整样本两个部分，并在 2.2 小节中将被再次分类.

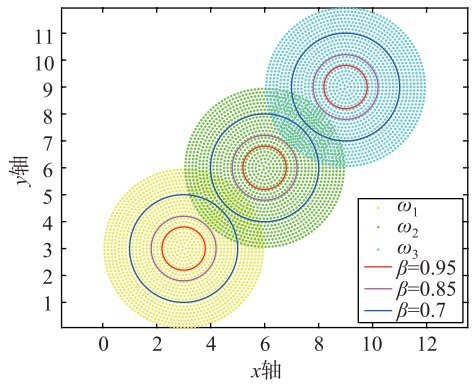

图 1 类中心距离较近时 β 取值示例

Fig.1 An example of β when class center distance is close

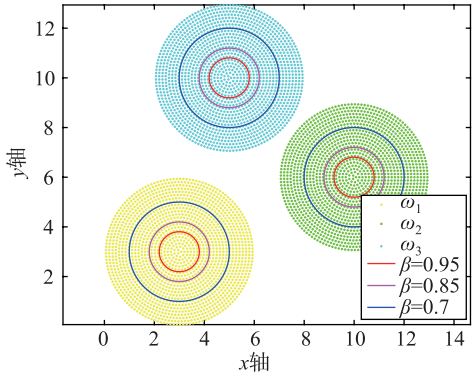

图 2 类中心距离较远时 β 取值示例

Fig.2 An example of β when class center distance is far

2.2 剩余样本的分类

剩余样本的分类采用经训练集训练的多个基础分类器实现. 剩余样本包括数据集 X_{complete} 中的不可靠样本和数据集 X_{missing} 中的样本. 分类器的选择没有特别要求，概率框架下的 SVM[22]、KNN[31]、决策树[32] 或者 D-S 证据理论框架下的 EK-NN[17]、BK-NN[33] 都可以使用，本文选用 KNN 和 SVM 作为基础分类器. 在训练分类器时，利用与测试集中样本现有属性对应的训练集属性训练多个基础分类器，然后使用相应的基础分类器对测试集中的样本进行分类，极端情况下，测试集中的每一个不完整样本都需要训练一个对应的基础分类器.

根据样本 $\boldsymbol{x}_i$ 的已知属性，基础分类器利用从 2.1 小节中得到的训练集进行训练，然后使用训练好的分类器来对样本 $\boldsymbol{x}_i$ 进行分类. 重复以上步骤直到所有不可靠样本和不

完整样本被分类完毕. 假设一个三维不完整样本 $\boldsymbol{x}_i=[?, x_{i2}, x_{i3}]$，第一维属性值 $\boldsymbol{x}_{i1}$ 丢失，那么基础分类器将使用训练集中所有样本的第二维和第三维属性进行训练，即 x_{j2} 和 x_{j3}，$\boldsymbol{x}_j\in X_{\text{training}}$. 所以在 HCA 中，测试集中的每一个样本都使用一个完整属性训练的分类器来进行分类.

分类器通常根据样本对类别的支持度将其划分给一个单类，但实际过程中会出现样本因为缺失值而同时支持多个单类且没有明显差别的情况. 例如一个 3 类问题，样本 $\boldsymbol{x}_i$ 对于 3 个类别的支持度为 $P(\boldsymbol{x}_i\in\omega_1)=0.1$，$P(\boldsymbol{x}_i\in\omega_2)=0.45$，$P(\boldsymbol{x}_i\in\omega_3)=0.45$，根据分类结果没有办法直接看出该样本是属于 ω_2 还是 ω_3. 如果将其直接划分给 ω_2 或者 ω_3 会增加错误分类的风险.

2.3　不精确样本的表征

为了合理表征不精确样本，利用复合类阈值 ε 将难以划分到单类的样本划分给相应复合类以降低错误率. 对于一个 c 类问题，首先找到样本 $\boldsymbol{x}_i$ 的最大支持度，最大支持度表示如下：

$$P^{\omega_{\max}}_{\max}(\boldsymbol{x}_i)=\max\{P^{\omega_i}(\boldsymbol{x}_i),\cdots,P^{\omega_k}(\boldsymbol{x}_i),\cdots,P^{\omega_c}(\boldsymbol{x}_i)\}\quad(9)$$

式中，$P^{\omega_{\max}}_{\max}(\boldsymbol{x}_i)$ 是样本 $\boldsymbol{x}_i$ 的最大支持度，$\omega_{\max}$ 为样本 $\boldsymbol{x}_i$ 最大支持度的类别，$P^{\omega_k}(\boldsymbol{x}_i)$ 是样本 $\boldsymbol{x}_i$ 对类别 ω_k 的支持度.

计算样本属于各类的支持度与最大类别支持度之间的差值，如果差值小于等于复合类阈值 ε，则将样本划分给相应复合类. 复合类表示如下：

$$\omega_{\max,\tau}=\{\boldsymbol{x}_i\mid P^{\omega_{\max}}_{\max}(\boldsymbol{x}_i)-P^{\omega_\tau}(\boldsymbol{x}_i)\leqslant\varepsilon\}\quad(10)$$

式中，$\varepsilon\in[0,1]$ 是一个可选择阈值，ω_τ 是一个可变集合，可能包含 $\omega_1,\cdots,\omega_c$ 中的一个或者多个元素，组成复合类的单类数量取决于满足式(10)单类的数量. 若 $\omega_i,\omega_j\in\omega_1,\cdots,\omega_c$ 且满足 $P^{\omega_{\max}}_{\max}(\boldsymbol{x}_i)-P^{\omega_i}(\boldsymbol{x}_i)\leqslant\varepsilon$ 和 $P^{\omega_{\max}}_{\max}(\boldsymbol{x}_i)-P^{\omega_j}(\boldsymbol{x}_i)\leqslant\varepsilon$，则 ω_τ 包含 ω_i 和 ω_j 两个元素，即 $\omega_\tau\Leftrightarrow\omega_i,\omega_j$，最终的复合类 $\omega_{\max,\tau}\Leftrightarrow\omega_{\max},\omega_i,\omega_j$.

从式(10)可知，如果阈值 ε 较小，样本更可能被划分给单类 $\omega_{\max}$，复合类中只有少量样本，这会增加一些不确定样本被错误分类的风险；如果 ε 值较大，更多的样本将被划分给复合类($\omega_{\max,\tau}$)，产生较大的不精确率. 因此 ε 应根据可接受的不精确率来进行调整. 为了尽量使更多的样本产生精确的结果，需要为复合类中的不完整样本添加更多信息来对其进行划分.

2.4　不精确样本的划分

为了对复合类中不完整样本尽可能精确划分，采用基于 D-S 理论的方法对复合类中的不完整样本插补、分类、融合. 假设一个属于 c 类问题(即存在类别 $\omega_1,\omega_2,\cdots,\omega_c$)的三维样本 $\boldsymbol{x}_k=[x_{k1},x_{k2},?]$，它只有前两维属性(即 x_{k1} 和 x_{k2})，第三维属性(即 x_{k3})丢失. 根据样本 $\boldsymbol{x}_k$ 的现有属性，在 2.3 小节中被划分到复合类 $\omega_{\max,\tau}$ 中(假设这个复合类中包含两个单类，即 $\omega_{\max,\tau}\Leftrightarrow\omega_i,\omega_j$，$i,j=1,\cdots,c$). 在类别 ω_i 和类别 ω_j 中分别对不完整样本 $\boldsymbol{x}_k$ 的缺失属性 x_{k3} 进行填补，得到两个估计值 x^i_{k3} 和 x^j_{k3}(x^i_{k3} 表示缺失值在类别 ω_i 中的估计值，x^j_{k3} 表示缺失值在类别 ω_j 中的估计值)，并分别组成两个完整样本 $\boldsymbol{x}^i_k$ 和 $\boldsymbol{x}^j_k$($\boldsymbol{x}^i_k$，$\boldsymbol{x}^j_k$ 分别表示不完整样本 x_k 在类别 ω_i，ω_j 中估计后得到的完整的样本). 这里使用 K 近邻插补方法对不完整样本的缺失值进行估计，然后使用相应的基础分类器对填补后的完整样本 $\boldsymbol{x}^i_k$ 和 $\boldsymbol{x}^j_k$ 进行分类，得到两个分类结果 $\boldsymbol{P}(\boldsymbol{x}^i_k)=[P^{\omega_1}(\boldsymbol{x}^i_k),\cdots,\boldsymbol{P}^{\omega_c}(\boldsymbol{x}^i_k)]$ 和 $\boldsymbol{P}(\boldsymbol{x}^j_k)=[\boldsymbol{P}^{\omega_1}(\boldsymbol{x}^j_k),\cdots,\boldsymbol{P}^{\omega_c}(\boldsymbol{x}^j_k)]$. 将分类结果 $\boldsymbol{P}(\boldsymbol{x}^i_k)$ 和 $\boldsymbol{P}(\boldsymbol{x}^j_k)$ 进行融合得到不完整样本 $\boldsymbol{x}_k$ 的最终分类结果. 融合分类结果前，首先判断分类结果 $\boldsymbol{P}(\boldsymbol{x}^i_k)$ 和 $\boldsymbol{P}(\boldsymbol{x}^j_k)$ 是否冲突，其方法为：设 $\boldsymbol{P}^{\omega_{\max}}_{\max}(\boldsymbol{x}^i_k)$ 和 $\boldsymbol{P}^{\omega_{\max}}_{\max}(\boldsymbol{x}^j_k)$ 是由式(9)得到的两个最大支持度，且 $\boldsymbol{P}^{\omega_{\max}}_{\max}(\boldsymbol{x}^i_k)=\boldsymbol{P}^{\omega_a}_{\max}(\boldsymbol{x}^i_k)$，$\boldsymbol{P}^{\omega_{\max}}_{\max}(\boldsymbol{x}^j_k)=\boldsymbol{P}^{\omega_b}_{\max}(\boldsymbol{x}^j_k)$，$\omega_a$ 和 ω_b 分别代表样本 $\boldsymbol{x}^i_k$ 和样本 $\boldsymbol{x}^j_k$ 分类结果中的最大支持度类别. 若 $\omega_a=\omega_b$，表示样本 $\boldsymbol{x}^i_k$ 和 $\boldsymbol{x}^j_k$ 支持相同的类别，即分类结果 $\boldsymbol{P}(\boldsymbol{x}^i_k)$ 和 $\boldsymbol{P}(\boldsymbol{x}^j_k)$ 不冲突；若 $\omega_a\neq\omega_b$，则样本 $\boldsymbol{x}^i_k$ 和 $\boldsymbol{x}^j_k$ 支持不同的类别，即分类结果 $\boldsymbol{P}(\boldsymbol{x}^i_k)$ 和 $\boldsymbol{P}(\boldsymbol{x}^j_k)$ 冲突. 然后，依据分类结果是否冲突选用融合规则：①若分类结果不冲突，使用D-S规则对分类结果进行融合；②若分类结果相互冲突，使用改进的D-S规则——Yager 规则对分类结果进行融合. 最终分类结果可以将本身处于复合类中的不完整样本进行重新划分，得到更精确的分类结果.

为了方便表示，HCA 方法的伪代码如算法 1 所示.

算法 1　HCA 方法

输入：

　　待聚类数据集：$X=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_l\}$

参数：

　　c：聚类类别数 $2\leqslant c\leqslant l$

　　β：训练阈值(默认值 $\beta\in[0.85,0.95]$)

　　ε：复合类阈值(默认值 $\varepsilon\in[0,1]$)

Start

　　将数据集 X 划分成 X_{complete} 和 X_{missing}；

　　使用 FCM 对数据集 X_{complete} 进行聚类，根据式(8)构建训练集，剩余样本构成测试集；

　　根据测试集中样本的现有属性，利用训练集构造多个训练集，并训练多个基础分类器；

　　使用多个基础分类器对测试集中相应样本进行分类；

　　根据式(10)将样本提交给复合类 $\omega_{\max,\tau}$；

　　对复合类中的不完整样本在其可能的单类中分别估计并分类；

　　根据分类结果是否冲突使用 D-S 规则或 Yager 规则对多个分类结果自适应融合.

End

3　实验与结果分析

采用 3 个实验验证 HCA 算法的性能. HCA 算法的性能将通过与均值插补(MI)、K 近邻插补(KNNI)和 FCM 插补(FCMI)方法进行对比来验证. 在 MI 中，利用对应于缺失属性的所有样本属性的平均值来代替缺失值. 在 KNNI 中，利用与含有缺失值样本最相似的 K 个样本来对缺失值进行估计. 在 FCMI 中，根据 FCM 产生的聚类中心，利用含有缺失值样本与类中心之间的距离来估计缺失值. 实验中为了对比各算法的性能，采用误分率 R_e 和不精确率 R_i

两个指标进行评估，R_e 和 R_i 的定义为

$R_e=\frac{N_e}{T}$（N_e：错分的样本数量，T：数据集中样本的总数）；

$R_i=\frac{N_i}{T}$（N_i：划分到复合类的样本数量）.

3.1 人工合成数据示例

使用模拟数据来验证 HCA 与其它方法的区别. 考虑一个三类问题 $\Omega=\{\omega_1,\omega_2,\omega_3\}$，样本数据呈圆形如图 3 所示，每类包含 1 950 个样本. 圆形的圆心分别为 $c_1=(3,8)$，$c_2=(7,3)$，$c_3=(11,8)$，半径为 $r=3$. 训练集和测试集分布如图 4 所示. 假设测试集中一半样本的第二维属性（对应于 y 轴坐标）丢失，属性丢失的测试样本仅根据第一维属性值（对应于 x 轴坐标值）进行分类. 使用 SVM 作为基础分类器，设定训练阈值 $\beta=0.9$ 和复合类阈值 $\varepsilon=0.2$，将 HCA 与均值插补，K-NN 插补和 FCM 插补三种方法进行对比. 在 K-NN 插补中，选择 $K=5$. 使用不同方法的测试样本分类结果如图 5 ~ 图 8 所示. 为了方便表示，定义 ω^{te} 代表 ω^{test}（测试集），ω^{tr} 代表 $\omega^{training}$（训练集），$\omega_{i,\cdots,k}$ 代表 $\omega_i\cup\cdots\cup\omega_k$（复合类）.

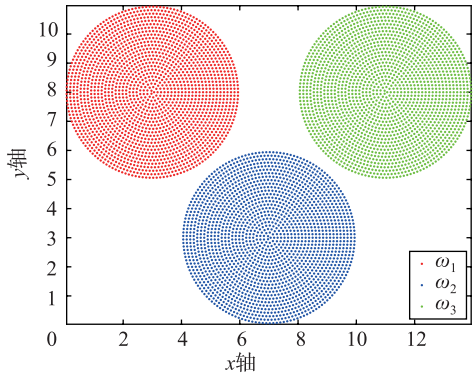

图 3 原始数据集

Fig.3 Original data set

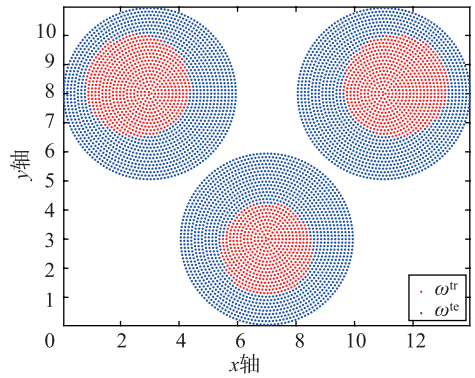

图 4 训练集和测试集

Fig.4 Training data and test data

从图 3 可以看出，类别 ω_2 在 x 轴对应的属性上与类别 ω_1 和类别 ω_3 在边界区域有部分重叠. 处于重叠区域的样本仅仅根据 x 轴的属性值很难被正确分类，因为无法仅根据第一维属性（x 轴）区分这些样本是属于类别 ω_1 还是 ω_2（或者 ω_2 和 ω_3）. 传统的 MI、KNNI 和 FCMI 方法直接对样本的缺失值进行估计，在估计过程中为样本带来了不确定性，传统概率框架下将所有样本分到一个确定的单类中，使得处于不同类别重叠区域的样本被强行划分，没有办法真实反映出由于缺失值估计和样本原始分布带来的不精确性和不确定性，带来了较大的误分率. HCA 算法在 D-S 证据理论框架下，能够将大部分处于不同类别重叠区域的样本有效划分到相应的复合类 $\omega_{1,2}$ 或 $\omega_{2,3}$ 中，这样能够有效表达样本类别的不精确性并降低了误分率，结果如图 8 所示.

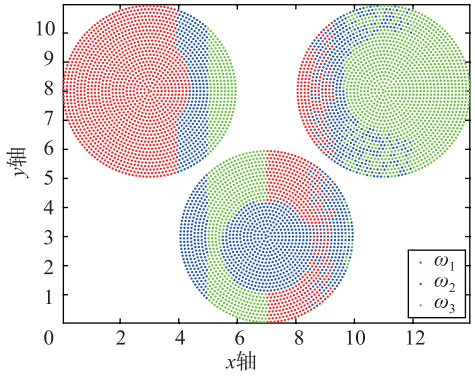

图 5 MI 估计实验结果（$R_e=36.99$）

Fig.5 Result by method with MI estimation ($R_e=36.99$)

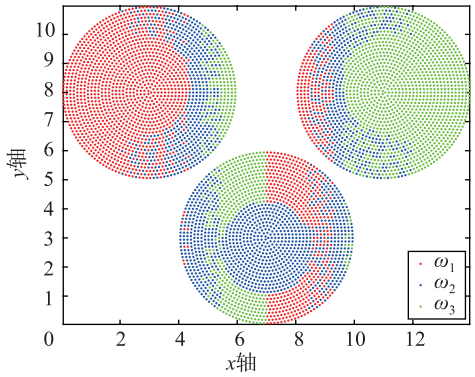

图 6 KNNI 估计实验结果（$R_e=37.88$）

Fig.6 Result by method with KNNI estimation ($R_e=37.88$)

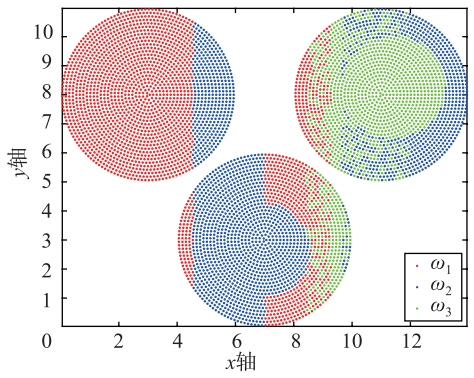

图 7 FCMI 估计实验结果（$R_e=33.30$）

Fig.7 Result by method with FCMI estimation ($R_e=33.30$)

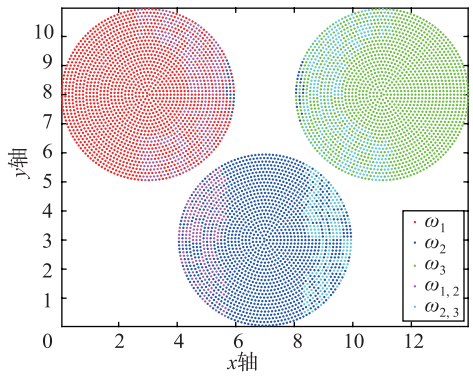

图 8　HCA 实验结果（$R_e=0.7$，$R_i=20.19$）

Fig.8　Result by HCA（$R_e=0.7$，$R_i=20.19$）

3.2　参数实验

从 UCI 数据集中选取 8 个真实数据集来验证提出算法的性能并与其它三种方法进行对比，同时对复合类阈值参数 ε 进行讨论. 实验所用 8 个数据集的基本信息在表 1 中给出. 实验中，使用 SVM 作为基础分类器，采用 2 折交叉验证法，每个数据集有 n 个缺失属性的样本，缺失值在所有样本的所有属性上随机产生. 默认训练阈值为 $\beta=0.9$，不同复合类阈值下的分类结果如表 2 所示.

表 1　实验用数据集的基本信息

Tab.1　Basic information of the used data sets

数据集名称	类别数	属性数	样本数
Breast	2	9	699
Heart	2	13	270
Pima	2	8	768
Wbc	2	9	683
Bupa	2	6	345
Wine	3	13	178
Iris	3	4	150
Seeds	3	7	210

表 2　不同复合类阈值下的数据集分类结果（%）

Tab.2　Classification results of different methods with different meta-class threshold（%）

Data	n	MI	KNNI	FCMI	ε	HCA		ε	HCA		ε	HCA	
						R_e	R_i		R_e	R_i		R_e	R_i
Breast	3	4.86	4.86	4.86	0.1	**4.29**	1.43	0.3	**3.58**	4.58	0.5	**2.00**	9.87
	6	5.08	5.15	5.01		**4.58**	1.57		**3.72**	4.58		**2.15**	10.01
	10	5.29	5.58	5.37		**4.86**	1.57		**3.86**	4.58		**2.43**	10.16
Heart	3	41.30	41.48	41.11	0.1	**39.26**	4.07	0.3	**34.33**	12.45	0.5	**31.48**	20.37
	6	41.85	42.41	41.48		**39.82**	4.07		**34.89**	12.59		**31.85**	20.37
	10	42.20	42.78	41.85		**40.00**	4.81		**36.12**	12.78		**32.96**	20.74
Pima	3	34.24	34.24	34.31	0.1	**33.40**	2.34	0.3	**31.03**	7.90	0.5	**29.39**	12.72
	6	34.31	34.38	34.51		**33.79**	2.34		**31.19**	7.80		**29.52**	12.72
	10	34.51	34.38	34.58		**33.92**	2.34		**31.51**	7.85		**30.08**	12.96
Wbc	3	4.54	4.54	4.54	0.1	**4.25**	1.61	0.3	**3.37**	4.69	0.5	**1.61**	10.25
	6	4.69	4.62	4.62		**4.39**	1.61		**3.51**	4.69		**1.76**	10.25
	10	4.83	4.76	4.69		**4.54**	1.61		**3.66**	4.69		**2.05**	10.25
Bupa	3	46.09	46.82	46.25	0.1	**45.8**	2.03	0.3	**45.06**	4.93	0.5	**43.48**	7.25
	6	46.38	46.96	46.38		**45.95**	2.03		**45.5**	4.93		**44.06**	7.54
	10	46.96	47.25	46.67		**46.24**	2.03		**45.64**	5.22		**44.06**	7.58
Wine	3	31.46	32.02	31.18	0.1	**26.78**	14.79	0.3	**20.78**	26.12	0.5	**19.86**	27.53
	6	32.30	32.58	33.15		**28.37**	15.17		**22.47**	27.57		**21.35**	27.77
	10	32.58	32.58	33.71		**29.59**	15.61		**23.6**	27.9		**23.04**	28.65
Iris	3	10.00	10.00	8.67	0.1	**6.00**	7.34	0.3	**2.67**	29.17	0.5	**2.67**	41.34
	6	10.67	10.34	10.00		**7.67**	8.34		**4.00**	29.35		**2.67**	42.00
	10	11.33	11.33	10.33		**8.89**	10.22		**7.00**	29.67		**3.78**	42.11
Seeds	3	10.95	10.95	11.05	0.1	**8.41**	12.5	0.3	**3.81**	40.11	0.5	**2.22**	50.79
	6	11.43	11.43	11.90		**9.76**	12.62		**5.24**	40.48		**3.33**	50.95
	10	11.90	11.66	12.48		**10.7**	12.62		**7.14**	40.57		**5.24**	50.57

从表 2 中可以看出，HCA 方法相比于其他三种方法（MI、KNNI 和 FCMI）错误率更低，但与此同时因为引入复合类产生了一些不精确的分类结果. 一些样本因为属性值的缺失或处于不同类别的重叠区域难以被正确分类到单类从而被划分给了相应的复合类. D-S 证据理论下的 HCA 方法能够有效表达由于数据缺失和数据原始分布造成的不确定性和不精确性并降低误分率. 随着测试数据集中缺失样本数量的增加，误分率也相应增加，同时也会产生更多的不精确率. 这种情况是合理的，因为样本的缺失属性越多，造成样本分类的不确定性就越大，也更容易引起误分. 从表 2 中还可以看出，随着复合类阈值 ε 的变化，误分率和不精确率会相应的变化，当 ε 越大时，只有分类结果中属于一个单类的样本才会被划分到相应的单类中，误分率会降低，不精确率会上升；当 ε 越小时，只要样本对不同类别有一定的区分度就可以将其划分到相应的单类中，误分率会上升，不精确率会下降. 所以，对于复合类

阈值 ε 的取值，需要在误分率和不精确率之间找到一个折中.

3.3 适用性实验

实验使用表 1 中的 8 个数据集，通过与 MI、KNNI 和 FCMI 方法进行比较来验证 HCA 方法的性能并说明 HCA 方法的适用性. KNN 和 SVM 在实验中作为基础分类器. 选择复合类阈值 $\varepsilon=0.3$，训练集阈值 $\beta=0.9$. 平均误分率 R_e(%)和不精确率 R_i(%)在表 3 中给出，为了方便表示，定义 BC 代表基础分类器，A 代表 KNN 作为基础分类器，B 代表 SVM 作为基础分类器.

表 3 不同基础分类器的数据集分类结果(%)

Tab.3 Classification results of different methods with different basic classifier (%)

Data	(n, BC)	MI	KNNI	FCMI	HCA	
					R_e	R_i
Breast	(3, A)	6.01	6.15	6.15	**5.29**	1.43
	(3, B)	4.86	4.86	4.86	**3.58**	4.58
	(6, A)	6.15	6.29	6.29	**5.44**	1.57
	(6, B)	5.08	5.15	5.01	**3.72**	4.58
	(10, A)	6.29	6.44	6.44	**5.58**	1.57
	(10, B)	5.29	5.58	5.37	**3.86**	4.58
Breast	(3, A)	42.22	42.02	41.85	**39.08**	3.68
	(3, B)	41.30	41.48	41.11	**33.33**	12.45
	(6, A)	42.59	42.22	42.22	**39.63**	4.07
	(6, B)	41.85	42.41	41.48	**33.89**	12.59
	(10, A)	42.78	42.59	42.59	**39.63**	4.44
	(10, B)	42.20	42.78	41.85	**36.12**	12.78
Pima	(3, A)	34.11	34.18	34.11	**33.72**	0.65
	(3.B)	34.24	34.24	34.31	**31.03**	7.9
	(6, A)	34.38	34.31	34.45	**33.98**	0.78
	(6, B)	34.31	34.38	34.51	**31.19**	7.8
	(10, A)	34.45	34.51	34.77	**33.98**	0.91
	(10, B)	34.51	34.38	34.58	**31.51**	7.85
Wbc	(3, A)	5.93	5.86	5.86	**5.27**	1.61
	(3, B)	4.54	4.54	4.54	**3.37**	4.69
	(6, A)	6.08	6.00	6.15	**5.42**	1.61
	(6, B)	4.69	4.62	4.62	**3.51**	4.69
	(10, A)	6.15	6.15	6.23	**5.56**	1.61
	(10, B)	4.83	4.76	4.69	**3.66**	4.69
Bupa	(3, A)	45.37	45.22	45.22	**44.21**	1.16
	(3, B)	46.09	46.82	46.25	**45.06**	4.93
	(6, A)	45.95	46.09	45.51	**45.21**	1.16
	(6, B)	46.38	46.96	46.38	**45.50**	4.93
	(10, A)	46.53	46.38	46.09	**45.64**	1.60
	(10, B)	46.96	47.25	46.67	**45.64**	5.22
Wine	(3, A)	32.02	31.46	31.46	**30.53**	0.56
	(3, B)	31.46	32.02	31.18	**20.78**	26.12
	(6, A)	32.87	33.15	32.31	**32.02**	0.78
	(6, B)	32.30	32.58	33.15	**22.47**	27.57
	(10, A)	33.15	33.99	33.15	**32.58**	0.94
	(10, B)	32.58	32.58	33.71	**23.60**	27.9
Iris	(3, A)	11.00	10.67	10.33	**9.56**	3.34
	(3, B)	10.00	10.00	8.67	**2.67**	29.17
	(6, A)	11.33	11.67	11.33	**9.56**	4.44
	(6, B)	10.67	10.34	10.00	**4.00**	29.35
	(10, A)	11.67	13.33	12.00	**11.55**	6.22
	(10, B)	11.33	11.33	10.33	**7.00**	29.67
Seeds	(3, A)	11.33	11.43	11.33	**11.11**	1.43
	(3, B)	10.95	10.95	11.05	**3.81**	40.11
	(6, A)	11.74	11.90	11.67	**11.24**	1.43
	(6, B)	11.43	11.43	11.90	**5.24**	40.48
	(10, A)	12.62	12.38	12.38	**11.90**	3.49
	(10, B)	11.90	11.66	12.48	**7.14**	40.57

从表 3 中可以看出，不论是 SVM 还是 KNN 作为基础分类器，HCA 方法的误分率都比其它方法(MI、KNNI 和 FCMI)的误分率低，这说明 HCA 方法对于分类器具有良好的适应性. 与此同时，一些因为属性值缺失而难以直接划分到确定单类中的样本被划分到相应的复合类中. 随着测试集中缺失值数量(即 n)的增加，HCA 方法会得到更高的误分率和不精确率. 这种情况是正常的，因为随着缺失属性的增加，会给样本带来更多的不确定性，从而产生更多的误分率. 对于最终处于复合类中的样本来说，仅仅根据现有的属性信息不能够准确的对其进行类别划分，所以在进行决策判断时对于这些样本要更加谨慎. 如果需要得到更加精确的分类结果，那么就需要借助其它的技术手段或者其它信息源对这些样本进行再分类.

4 结论

本文提出了一种新的基于 D-S 证据理论的不完整数据混合分类算法，能够将几个单类支持度没有明显区别的样本划分到相应复合类，表达由于样本属性值缺失和数据原始分布造成的不确定性和不精确性，并降低分类错误率. 算法先使用软聚类方法将数据集中的完整样本聚类，并从中选择部分样本作为训练集，其余样本作为测试集. 根据测试集中样本的现有属性，使用相应训练集训练得到的分类器进行再分类. 如果一个样本对多个不同单类具有相近概率，这表明很难仅根据现有信息将样本正确分类. 在这种情况下，将其划分到相应的复合类以降低分类错误率. 最后，为了得到复合类中不完整样本更精确的分类结果，将复合类中的不完整样本在构成其所在复合类的单类中分别进行填补，并分类填补后的多个完整样本，不完整样本的最终分类结果由多个分类结果经 DS 规则或 Yager 规则自适应融合得到. 通过三个实验将 HCA 与其它方法进行对比，实验结果证实，HCA 算法能够合理表征由缺失值和原始数据分布引起的不确定性及不精确性，并降低误分率. 虽然本算法取得了较好的效果，但是仍有一些问题需要进一步研究：1)针对每一个不完整样本都要训练相应的分类器导致运算负担较大，如何训练一个可以适应不同完整样本的分类器并降低计算负担是下一步需要研究的问题之一；2)最终结果将一些支持度相近的样本保留到了相应的复合类中，下一步工作考虑如何更好得挖掘数据的现有信息，以得到在复合类中样本更加精确的分类结果.

参考文献

[1] Little R J A, Rubin D B. Statistical analysis with missing data[M]. 2nd ed. New York, NY, USA: Wiley, 2002.

[2] Bezdek J C, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm[J]. Computers & Geosciences, 1984, 10(2/3): 191 – 203.

[3] Marie-Hélène M, Denœux T. ECM: An evidential version of the fuzzy c-means algorithm[J]. Pattern Recognition. 2008, 41 (4): 1384 – 1397.

[4] Pedro J G, José L S, Aníbal R F. Pattern classification with missing data: A review[J]. Neural Computer, 2010, 19(2): 263 – 282.

[5] Qiang Y, Ling C, Chai X, et al. Test-cost sensitive classification on data with missing values[J]. IEEE Transactions on Knowledge & Data Engineering, 2006, 18(5): 626 – 638.

[6] 钱晓东，罗彦福. 基于互信息属性排序的不完整数据聚类算法[J]. 信息与控制，2019，48(1)：80 – 87.
Qian X D, Luo Y F. Incomplete data clustering algorithm based on mutual information attributes ranking[J]. Information and Control, 2019, 48(1): 80 – 87.

[7] 邰凌楠，王春雨，田茂再. 缺失数据下的逆概率多重加权分位回归估计及其应用[J]. 统计研究，2018，35(9)：115 – 128.
Tai L N, Wang C Y, Tian M Z. Inverse probability multiple weighted quantile regression estimation and its application with missing data[J]. Statistical Research, 2018, 35(9): 115 – 128.

[8] Imbert A, Valsesia A, Le G C, et al. Multiple hot-deck imputation for network inference from RNA sequencing data[J]. Bioinformatics, 2018, 34(10): 1726 – 1732.

[9] 于力超，金勇进. 基于分层模型的缺失数据插补方法研究[J]. 统计研究，2018，35(11)：93 – 104.
Yu L C, Jin Y J. Research on comparison of missing data imputation methods based on multilevel models[J]. Statistical Research, 2018, 35 (11): 93 – 104.

[10] Aljuaid T, Sasi S. Proper imputation techniques for missing values in data sets[C]// 2016 International Conference on Data Science and Engineering. Piscataway, NJ, USA: IEEE, 2016: 1 – 5.

[11] Julián L, Josá A S, Herrera F. Missing data imputation for fuzzy rule-based classification systems[J]. Soft Computing, 2012, 16(5): 863 – 881.

[12] Samad T, Harp S A. Self-organization with partial data[J]. Network: Computation Neural Systems, 2009, 3(2): 205 – 212.

[13] Huang M W, Lin W C, Tsai C F. Outlier removal in model-based missing value imputation for medical datasets[J]. Journal of Healthcare Engineering, 2018, 2018: 1 – 9.

[14] Kim H G, Jang G J, Choi H J, et al. Recurrent neural networks with missing information imputation for medical examination data prediction [C]// 2017 IEEE International Conference on Big Data and Smart Computing (BigComp). Piscataway, NJ, USA: IEEE, 2017: 317 – 323.

[15] Li C X, Jiang B, Marlin B. MisGAN: Learning from incomplete data with generative adversarial networks[C/OL]//2019 International Conference on Learning Representations(ICLR 2019). (2019 – 2 – 24)[2019 – 10 – 10]. https: //openreview. net/forum? id = S1lDV3RcKm.

[16] 柯小路，马荔瑶，李子懿，等. 证据推理规则的性质研究及方法修正[J]. 信息与控制，2016，45(02)：165 – 170.
Ke X L, Ma L Y, Li Z Y, et al. Property research and approach modification of evidential reasoning rule[J]. Information and Control, 2016, 45(2): 165 – 170.

[17] Denœux T. A k-nearest neighbor classification rule based on dempster-shafer theory[J]. IEEE Transactions on Systems, Man and Cybernetics, 1995, 25(5): 804 – 813.

[18] 俞志富，李俊武，王利华. 一种基于云模型和证据理论的融合识别方法[J]. 信息与控制，2014，43(1)：30 – 36.
Yu Z F, Li J W, Wang L H. A fusion recognition method based on cloud model and evidence theory[J]. Information and Control, 2014, 43 (1): 30 – 36.

[19] Sen S, Dave R N. Clustering of relational data containing noise and outliers[C]// IEEE World Congress on IEEE International Conference on Fuzzy Systems. Piscataway, NJ, USA: IEEE, 1998: 1411 – 1416.

[20] Liu Z G, Pan Q, Dezert J, et al. Credal c-means clustering method based on belief functions[J]. Knowledge-Based Systems, 2015, 74(1): 119 – 132.

[21] 王彤，李卫伟. 一种改进的证据理论 C 均值分割方法[J]. 计算机应用与软件，2011，28(9)：255 – 256，297.
Wang T, Li W W. A segmentation method based on improved evidential C-means[J]. Computer Applications and Software, 2011, 28(9): 255 – 256, 297.

[22] Shevade S K, Keerthi S S, Bhattacharyya C, et al. Improvements to the SMO algorithm for SVM regression[J]. IEEE Transactions on Neural Networks, 2000, 11(5): 1188 – 1193.

[23] Shafer G. A mathematical theory of evidence[M]. Princeton, NJ, USA: Princeton University Press, 1976: 297.

[24] 尹慧琳，王磊. D-S 证据推理改进方法综述[J]. 计算机工程与应用，2005(27)：22 – 24.
Yin H L, Wang L. A review if modification methods of D-S evidence theory[J]. Computer Engineering and Applications, 2005(27): 22 – 24

[25] Yager R. On the Dempster-Shafer framework and new combination rules[J]. Information Sciences, 1987, 41 (2): 93 – 137.

[26] 孙全，叶秀清，顾伟康. 一种新的基于证据理论的合成公式[J]. 电子学报，2000(8)：117 – 119.
Sun Q, Ye X Q, Gu K W. A new combination rules of evidence theory[J]. Acta Electronica Sinica, 2000(8): 117 – 119.

（下转第 471 页）

[12] Zhu J L, Ge Z Q, Song Z H. Distributed parallel PCA for modeling and monitoring of large-scale plant-wide processes with big data[J]. IEEE Transaction on Industrial Informatics, 2017, 13(4): 1877 – 1885.
[13] Xu C, Zhao S Y, Liu F. Distributed plant-wide process monitoring based on PCA and minimal redundancy maximal relevance[J]. Chemometrics and Intelligent Laboratory Systems, 2017, 169: 53 – 63.
[14] Jiang Q C, Yan X F. Plant-wide process monitoring based on mutual information-multiblock principal component analysis[J]. ISA Transactions, 2014, 53(3): 1516 – 1527.
[15] Ge Z Q, Song Z H. Distributed PCA model for plant-wide process monitoring[J]. Industrial & Engineering Chemistry Research, 2013, 52(5): 1947 – 1957.
[16] Li S, Zhou X F, Pan F C, et al. Correlated and weakly correlated fault detection based on variable division and ICA[J]. Computers & Industrial Engineering, 2017, 112: 320 – 335.
[17] Jiang Q C, Wang B, Yan X F. Multiblock independent component analysis integrated with Hellinger distance and Bayesian inference for Non-Gaussian plant-wide process monitoring[J]. Industrial & Engineering Chemistry Research, 2015, 54(9): 2497 – 2508.
[18] Wang B, Yan X F, Jiang Q C. Independent component analysis model utilizing de-mixing information for improved non-Gaussian process monitoring[J]. Computers & Industrial Engineering, 2016, 94: 188 – 200.
[19] Cheng G, Macvoy T J. Predictive on-line monitoring of continuous processes[J]. Journal of Process Control, 1998, 8(6): 409 – 420.
[20] Tong C D, Palazoglu A, Yan X F. Improved ICA for process monitoring based on ensemble learning and Bayesian inference[J]. Chemometrics and Intelligent Laboratory Systems, 2014, 135: 141 – 149.
[21] 杨泽宇，王培良. 基于核独立成分分析和支持向量数据描述的非线性系统故障检测方法[J]. 信息与控制，2017，46(2)：149 – 156
Yang Z Y, Wang P L. Fault detection method for non-linear systems based on kernel independent component analysis and support vector data description[J]. Information and Control, 2017, 46(2): 149 – 156.
[22] Mustapha A, Abdelmalek K, Azzeddine B. A combined monitoring scheme with fuzzy logic filter for plant-wide Tennessee Eastman process fault detection[J]. Chemical Engineering Science, 2018, 187: 269 – 279.
[23] Downs J J, Vogel E F. A plant-wide industrial process control problem[J]. Computers & Chemical Engineering, 1993, 17(3): 245 – 255.

作者简介

万新春(1990 –)，男，硕士生. 研究领域为过程故障检测.
童楚东(1988 –)，男，博士，副教授. 研究领域为过程故障诊断与分析.
史旭华(1967 –)，女，博士，教授. 研究领域为计算智能与过程控制.

(上接第 463 页)

[27] 陈炜军，景占荣，袁芳菲，等. D-S 证据理论的不足及其数学修正[J]. 中北大学学报(自然科学版)，2010，31(2)：161 – 168.
Chen W J, Jing ZH R, Yuan F F, et al. Shortcoming of D-S evidence theory and its mathematic modification[J]. Journal of North University of China (Natural Science Edition), 2010, 31(2): 161 – 168.
[28] Murphy C K. Combining belief functions when evidence conflicts[J]. Decision Support Systems, 2000, 29(1): 1 – 9.
[29] 邓勇，施文康，朱振福. 一种有效处理冲突证据的组合方法[J]. 红外与毫米波学报，2004，23(1)：27 – 32.
Deng Y, Shi W K, Zhu Z F. Efficient combination approach of conflict evidence[J]. Journal of Infrared and Millimeter Waves, 2004, 23(1): 27 – 32.
[30] 卢正才，覃征. 证据合成的一般框架及高度冲突证据合成方法[J]. 清华大学学报(自然科学版)，2011，51(11)：1611 – 1615，1626.
Lu Z C, Qin Z. General framework for evidence combination and its approach to highly conflicting evidence fusion[J]. Journal of Tsinghua University (Science and Technology), 2011, 51(11): 1611 – 1615, 1626.
[31] Cover T M, Hart P E. Nearest neighbor pattern classification[J]. IEEE Transactions on Information Theory, 1967, 13(1): 21 – 27.
[32] Quinlan J R. Simplifying decision trees[J]. International Journal of Man-Machine Studies, 1987, 27(3): 221 – 234.
[33] Liu Z G, Pan Q, Dezert J. A new belief-based K-nearest neighbor classification method[J]. Pattern Recognition, 2013, 46(3): 834 – 844.

作者简介

段中兴(1969 –)，男，博士，教授，博士生导师. 研究领域为智能系统与智能信息处理，智能检测与视觉，建筑环境控制与节能优化，嵌入式技术与智能系统.
毕瀚元(1996 –)，男，硕士生. 研究领域为模式识别，机器学习.
张作伟(1988 –)，男，博士生，助教. 研究领域为模式识别，机器学习，D-S 证据理论及其应用.