基于不完备数据聚类的缺失数据填补方法

武 森[1)] 冯小东[1)] 单志广[2)]

[1)](北京科技大学东凌经济管理学院管理科学与工程系 北京 100083)

[2)](国家信息中心信息化研究部 北京 100045)

摘 要 缺失数据的处理是数据挖掘领域进行数据预处理的一个重要问题. 传统的缺失数据填补方法大部分是基于概率分布等一些统计假设,对于大数据集的数据挖掘不一定是最适合的方法. 受不完备数据分析(ROUSTIDA)未采用传统的概率统计学方法启发,提出基于不完备数据聚类的缺失数据填补方法(MIBOI),针对分类变量不完备数据集定义约束容差集合差异度,直接计算不完备数据对象集合内所有对象的总体相异程度,以不完备数据聚类的结果为基础进行缺失数据的填补. 采用UCI机器学习基准数据集进行实验表明,MIBOI对缺失数据的填补是有效可行的.

关键词 数据填补;不完备数据;聚类;约束容差集合差异度

中图法分类号 TP311 **DOI号**: 10.3724/SP.J.1016.2012.01726

Missing Data Imputation Approach Based on Incomplete Data Clustering

WU Sen[1)] FENG Xiao-Dong[1)] SHAN Zhi-Guang[2)]

[1)](*Department of Management Science and Engineering, Dongling School of Economics and Management, University of Science and Technology Beijing, Beijing* 100083)

[2)](*Informatization Research Department, State Information Center, Beijing* 100045)

Abstract Missing data processing is an important problem of data pre-processing in data mining field. Traditional missing data filling methods are mostly based on some statistical hypothesis, such as probability distribution, which might not be the most applicable approaches for data mining of large data set. Inspired by ROUSTIDA, an incomplete data analysis approach not using probability statistical methods, MIBOI is proposed for missing data imputation based on incomplete data clustering. Constraint Tolerance Set Dissimilarity is defined for incomplete data set of categorical variables, so the general dissimilarity of all the incomplete data objects in a set can be directly computed, and the missing data is imputed according to the incomplete data clustering results. The empirical tests using UCI benchmark data sets show that MIBOI is effective and feasible.

Keywords data imputation; incomplete data; clustering; constraint tolerance set dissimilarity

1 引 言

在实际数据分析中,经常遇到数据缺失问题. 作为机器学习领域基准数据库的UCI数据集中超过40%的数据库都含有缺失数据[1]. 这可能是由于数据获取限制,数据理解有误或数据漏读[2]等多方面原因造成的,使得在数据挖掘中面临的数据集往

收稿日期:2010-11-16;最终修改稿收到日期:2011-04-24. 本课题得到国家自然科学基金(70771007)、中央高校基本科研业务费专项资金(FRF-TP-10-006B)资助. **武 森**,女,1971年生,博士,教授,博士生导师,中国计算机学会(CCF)会员,主要研究方向为智能数据分析、知识发现和数据挖掘. E-mail: wusen@manage.ustb.edu.cn. **冯小东**,男,1988年生,博士研究生,主要研究方向为数据挖掘、聚类分析. **单志广**,男,1974年生,博士,研究员,中国计算机学会(CCF)理事,主要研究方向为信息系统性能评价.

往是不完整的.对这种有缺失数据的数据集,在此称为不完备的数据集.实际上,缺失数据是知识工程领域一个重要的热议问题[3].在数据挖掘中缺失数据经常处理不当,与缺失数据相关的有价值的知识往往被忽略了[4].本文针对缺失数据问题展开研究,提出基于不完备数据聚类的缺失数据填补方法 MIBOI (Missing data Imputation approach Based On Incomplete data clustering).该方法通过定义约束容差集合差异度,从集合的角度判断不完备数据对象的总体相异程度,进而以不完备数据聚类的结果为基础实现缺失数据的填补.采用 UCI 机器学习数据集中 4 个常用的分类变量基准数据集进行实验,结果表明:MIBOI 效率很高,且缺失数据填补的质量也比较高.

本文第 2 节总结缺失数据填补的相关工作;第 3 节定义基于不完备数据聚类进行缺失数据填补的基本概念;第 4 节提出并证明进行缺失数据填补的相关定理;第 5 节给出算法步骤;第 6 节用实验证明算法的有效性;第 7 节进行全文总结.

2　相关工作

对于存在缺失数据的不完备系统,多采用如下几种方法进行处理:(1)丢弃具有缺失数据的记录;(2)进行缺失数据的填补;(3)采用模型对缺失数据进行预测;(4)直接针对不完备数据进行分析.这些方法之间并不是相互排斥的,不同的方法之间在具体的实现算法上可能存在着紧密的联系.

丢弃具有缺失数据的记录[5-6]是应用中最简单的一种缺失数据处理方法.数据分析工作是在已经没有缺失数据的部分数据上完成的.这种方法在实际应用时也有不同的形式,例如在临床医学研究[7]中,具体又分为完全个体分析(complete case analysis)和可用个体分析(available case analysis).完全个体分析是将那些有缺失值的记录全部排除,只分析有完整数据的受试者,受试者如果有缺失值就要被排除在分析之外,会造成分析采用的实际样本量大大减少;可用个体分析是根据能够观察到的数据进行分析,只删掉那些需要用于分析的属性存在缺失数据的受试者,与完全个体分析相比可以更好地利用所有可用的信息,样本量会随着用于分析的属性不同而变化.总体而言,丢弃具有缺失数据的记录不能充分利用数据资源,而且可能会严重影响到数据的客观性和所研究问题结论的正确性.

对缺失数据进行填补,是为了在填补后的数据上完成具体问题的数据分析.简单而又常见的填补方法是全局常量填补法(global constant)和属性均值填补法(attribute mean)[6].文献[8]研究指出:在大多数情况下,这些方法同丢弃具有缺失数据的记录一样会生成有偏的结果,但有一些更复杂的填补方法,例如多重填补(multiple imputation),对缺失数据的处理更有效.多重填补不同于每一个缺失项填补一个值的单一填补(single imputation),其通过填补多个值以对填补的不确定性做出评价.热平台填补(hot deck imputation)[5]和冷平台填补(cold deck imputation)[5]都是典型的单一填补方法.热平台填补是将缺失值填补为与它最相似的一个对象的值.相似判定最常见的是使用相关系数矩阵来确定与缺失值所在属性最相关的属性,然后将所有对象按最相关属性值大小进行排序,将缺失值填补为排在它前面的对象值.与均值填补法相比,热平台填补后,变量的标准差与填补前比较接近,但这种方法使用不便,比较耗时.冷平台填补与热平台填补类似,不同之处在于填补值来自于其它数据源而不能是当前数据源.这些单一填补方法系统地低估了方差,而多重填补[9-10]是对单一填补的改进,它采用一系列可能的值来填补每一个缺失值,然后用标准的统计分析过程对多次填补后产生的若干个数据集进行分析,并将分析结果进行综合得到总体参数的估计值.多重填补反映了因缺失数据而导致的不确定性,统计推断更加有效.

采用模型对缺失数据进行预测的方法首先对输入的数据定义一个模型,然后基于该模型对未知参数进行极大似然(maximum likelihood)估计.期望值最大化(expectation maximization)算法[5]是缺失数据分析中用于极大似然估计的很常用的迭代算法.期望值最大化算法在某些情况下,当数据缺失比例很大时收敛可能很慢,各种改进算法或替代算法[11-12]相继提出.

直接针对不完备数据进行分析的方法,既不删除具有缺失数据的记录,也不进行缺失数据的填补.这种处理方法比较多地用在机器学习领域的分类问题研究中.这种不完备数据分析方法的一个重要特征是以提高分类性能为目标,旨在提高分类结果的精确性.例如,近年来直接扩充粗糙集[13-15]或决策树[16]模型来进行不完备信息系统研究取得了许多成果,这些成果主要是针对具有决策属性的决策表问题展开的,其实质是分类问题.

由于许多数据分析算法或具体应用问题要求输入的数据必须是完备的,所以直接针对不完备数据进行分析的方法就受到了限制.而其它方法大部分都是概率统计学方法,往往基于概率分布等一些统计假设,由于在数据挖掘领域数据集通常非常巨大,对这些分布假设的判定比较困难,因此这些传统的统计技术对数据挖掘应用中的缺失数据处理不一定是最适合的方法[17].另外,许多缺失数据填补方法都是针对连续变量提出的.尽管分类数据缺失无处不在,但是并没有处理分类变量缺失数据可用的原则性方法.当分类变量数目很大时,当前的缺失分类数据填补统计方法由于稳健性及实施困难等原因在应用中就受到了限制[10].

ROUSTIDA 不完备数据分析[17-18]未采用传统的概率统计学方法,而是在对粗糙集理论进行研究的基础上提出的.它是针对分类变量缺失数据进行填补质量和效率都比较好的一种方法.其基本思想是:缺失数据的填补应使填补后的数据集中具有缺失数据的对象与其它相似对象的属性值尽可能保持一致,亦即属性值之间的差异尽可能小.由于粗糙集理论中的差异矩阵反映了对象间的属性值差异,所以 ROUSTIDA 通过扩充差异矩阵来适合不完备数据集,进而实现缺失数据的填补.

受 ROUSTIDA 不完备数据分析方法启发,本文提出的 MIBOI 不完备数据填补方法也未采用传统的概率统计学方法,而是基于不完备数据聚类提出的.通过针对分类变量不完备数据集定义约束容差集合差异度,从集合的角度判断不完备数据对象的总体相异程度,进而以不完备数据聚类的结果为基础进行缺失数据的填补.

3 基本概念

3.1 不完备信息系统

定义 1(不完备信息系统). 不完备信息系统定义为 $\boldsymbol{I}=(U,A,V,f)$,其中,对象集合 $U=\{x_1,x_2,\cdots,x_n\}$;对象属性集合 $A=\{a_1,a_2,\cdots,a_m\}$;V 是属性值的集合,$V=\bigcup V(a_k)$,$V(a_k)$ 是属性 a_k 的值域;f 是 $U\times A\rightarrow V$ 的映射函数,它为每个对象的每个属性赋予一个属性值,第 i 个对象的第 k 个属性值 $a_k(x_i)\in V(a_k)$,$i\in\{1,2,\cdots,n\}$,$k\in\{1,2,\cdots,m\}$,$a_k(x_i)=$"$*$"表示属性值缺失.

3.2 容差属性

定义 2(容差属性). 给定不完备信息系统 $\boldsymbol{I}=(U,A,V,f)$,对于非空对象集合 $X\subseteq U$,定义

$$\begin{aligned}T(X)=\{a_k \mid \forall_{x_i\in X,x_j\in X}a_k(x_i)&=a_k(x_j)\vee a_k(x_i)\\&=\text{“}*\text{”}\vee a_k(x_j)=\text{“}*\text{”}\}\end{aligned}\tag{1}$$

为集合 X 的容差属性集合,称 $a_k\in T(X)$ 为容差属性.

从上述定义可知,对于非空对象集合 X,其容差属性集合由满足下述条件的所有属性构成:X 中的任意两个对象在该属性的值没有明确不同的情况,即 X 中的任意两个对象在该属性的值相同或至少一个对象在该属性的值缺失.

如果 X 的容差属性集合 $T(X)\neq\varnothing$,对于容差属性 $a_k\in T(X)$,记

$$a_k(X)=\begin{cases}a_k(x_i), & \exists_{x_i\in X}a_k(x_i)\neq\text{“}*\text{”}\\ \text{“}*\text{”}, & \forall_{x_i\in X}a_k(x_i)=\text{“}*\text{”}\end{cases}\tag{2}$$

为容差属性 a_k 的容差值.相应地,记

$$V(X)=\{(k,a_k(X))\mid a_k\in T(X)\}\tag{3}$$

为容差属性集合 $T(X)$ 对应的(属性序号,容差值)二元组集合,简称为容差属性二元组集合.

3.3 约束容差属性

定义 3(约束容差属性). 给定不完备信息系统 $\boldsymbol{I}=(U,A,V,f)$,对于非空对象集合 $X\subseteq U$,$V(X)$ 为集合 X 的容差属性二元组集合,定义

$$S(X)=\begin{cases}V(X), & \exists_{(k,a_k(X))\in V(X)}a_k(X)\neq\text{“}*\text{”}\\ \varnothing, & \forall_{(k,a_k(X))\in V(X)}a_k(X)=\text{“}*\text{”}\end{cases}\tag{4}$$

为集合 X 的约束容差属性二元组集合,称 a_k 为约束容差属性.

从上述定义可知,对于非空对象集合 X,其约束容差属性集合中的属性都是容差属性,且受到这些容差属性中至少有一个属性的容差值不为"$*$"的约束.约束容差属性集合比容差属性集合更加严格.

如果 X 的约束容差属性二元组集合 $S(X)=\varnothing$,则包括如下两种情况:

(1) X 中没有容差属性.

(2) X 中有容差属性,但所有容差值都为"$*$".

3.4 约束容差集合差异度

定义 4(约束容差集合差异度). 给定不完备信息系统 $\boldsymbol{I}=(U,A,V,f)$,对于非空集合 $X\subseteq U$,$|X|$ 为 X 的基数,m 为属性数目,如果 X 的约束容差属性二元组集合 $S(X)\neq\varnothing$,则定义

$$D(X)=\frac{m-|S(X)|}{\sqrt{|X|\times|S(X)|}},\ X\neq\varnothing$$

且　　$$S(X)\neq\varnothing \tag{5}$$

为集合 X 的约束容差集合差异度.

约束容差集合差异度 $D(X)$反映了存在缺失数据的情况下集合 X 内各对象间的总体差异程度. $D(X)$越小,表明 X 集合内各对象间总体差异程度越小,各对象间越相似;$D(X)$越大,表明 X 集合内各对象间总体差异程度越大,各对象间越不相似.

如果 X 的约束容差属性集合 $S(X)=\varnothing$,则约束容差集合差异度为未定义,这体现在如下两种情况:

(1) X 中没有容差属性,即对于任意属性在 X 中都存在至少两个对象在该属性的值明确不同,则 X 内各对象间的总体差异程度很大,不再进行定义.

(2) X 中有容差属性,但所有容差值都为“$*$”,此时对 X 内所有对象没有一个属性取值明确相同,更重要的是所有容差值都为“$*$”对缺失数据的填补没有意义,所以不再进行定义.

3.5　约束容差集合精简

定义 5(约束容差集合精简). 给定不完备信息系统 $\boldsymbol{I}=(U,A,V,f)$,对于非空集合 $X\subseteq U$,$|X|$为 X 的基数,X 的约束容差属性二元组集合 $S(X)\neq\varnothing$,$D(X)$为约束容差集合差异度,则定义向量

$$\boldsymbol{R}(X)=(|X|,S(X),D(X)) \tag{6}$$

为集合 X 的约束容差集合精简.

根据上述定义,约束容差集合精简是一个向量,其包含 3 个分量:$|X|$,$S(X)$和 $D(X)$. 根据式(5)在计算约束容差集合差异度 $D(X)$时需要用到 m,$|X|$和$|S(X)|$. 由于属性数目 m 为已知常数,$|X|$和 $S(X)$是包含在约束容差集合精简中的前两个分量,所以 $D(X)$可以根据已知常数 m 及前两个分量 $|X|$和 $S(X)$直接计算得到.

这样,约束容差集合精简 $\boldsymbol{R}(X)=(|X|,S(X),D(X))$不仅包含了约束容差集合差异度 $D(X)$,表明了存在缺失数据的情况下集合 X 内各对象间的总体差异程度,同时也通过前两个分量$|X|$和 $S(X)$概括了计算约束容差集合差异度所需的全部对象信息. 本文提出的 MIBOI 方法在计算过程中只保留约束容差集合精简,而不保留所有对象的信息,从而在处理大数据集时数据处理量大规模减少.

约束容差集合精简是针对集合提出的,不受集合中包含对象数目的限制. 如果 X 中只包含一个对象,不妨记 $X=\{x\}$,则约束容差集合精简 $R(X)$包含了对象 x 的全部相关信息,如式(7)所示

$$\boldsymbol{R}(X)=\boldsymbol{R}(\{x\})=(1,\{(1,a_1(x)),(2,a_2(x)),\cdots,(m,a_m(x))\},0) \tag{7}$$

3.6　约束容差交运算

定义 6(约束容差交运算). 给定不完备信息系统 $\boldsymbol{I}=(U,A,V,f)$,非空集合 $X\subseteq U$ 的约束容差属性二元组集合记为 $S(X)$,对于 $X_1\subseteq U$ 和 $X_2\subseteq U$ 且 $X_1\cap X_2=\varnothing$,如果 $S(X_1)\neq\varnothing$且 $S(X_2)\neq\varnothing$,对于$(k,a_k(X_1))\in S(X_1)$和$(k,a_k(X_2))\in S(X_2)$且 $a_k(X_1)=a_k(X_2)\vee a_k(X_1)=\text{“}*\text{”}\vee a_k(X_2)=\text{“}*\text{”}$,记

$$\max(a_k(X_1),a_k(X_2))=\begin{cases}a_k(X_1),\ a_k(X_1)\neq\text{“}*\text{”}\wedge a_k(X_2)=\text{“}*\text{”}\\ a_k(X_2),\ a_k(X_1)=\text{“}*\text{”}\wedge a_k(X_2)\neq\text{“}*\text{”}\\ a_k(X_1)\text{ 或 }a_k(X_2),\ a_k(X_1)=a_k(X_2)\wedge a_k(X_1)\neq\text{“}*\text{”}\wedge a_k(X_2)\neq\text{“}*\text{”}\\ \text{“}*\text{”},\ a_k(X_1)=\text{“}*\text{”}\wedge a_k(X_2)=\text{“}*\text{”}\end{cases} \tag{8}$$

则定义

$$S(X_1)\bigcap{}^{T}S(X_2)=\{(k,\max(a_k(X_1),a_k(X_2)))\mid (k,a_k(X_1))\in S(X_1)\wedge(k,a_k(X_2))\in S(X_2)\wedge(a_k(X_1)=a_k(X_2)\vee a_k(X_1)=\text{“}*\text{”}\vee a_k(X_2)=\text{“}*\text{”})\} \tag{9}$$

为约束容差属性二元组集合 $S(X_1)$和 $S(X_2)$的容差交运算,并进一步定义

$$S(X_1)\bigcap{}^{C}S(X_2)=\begin{cases}S(X_1)\bigcap{}^{T}S(X_2),\\ \quad\exists_{(k,\max(a_k(X_1),a_k(X_2)))\in S(X_1)\cap^{T}S(X_2)}\\ \quad\max(a_k(X_1),a_k(X_2))\neq\text{“}*\text{”}\\ \varnothing,\\ \quad\forall_{(k,\max(a_k(X_1),a_k(X_2)))\in S(X_1)\cap^{T}S(X_2)}\\ \quad\max(a_k(X_1),a_k(X_2))=\text{“}*\text{”}\end{cases} \tag{10}$$

为约束容差属性二元组集合 $S(X_1)$和 $S(X_2)$的约束容差交运算.

从上述定义可知,对于非空约束容差属性二元组集合 $S(X_1)$和 $S(X_2)$,其容差交运算的结果由 $S(X_1)$和 $S(X_2)$中容差值不是明确不同的所有属性组成,即由 $S(X_1)$和 $S(X_2)$中容差值相同或至少一个为“$*$”的所有属性组成. 而约束容差交运算与容差交运算相比,增加了运算结果中至少有一个属性的容差值不为“$*$”的约束,比容差交运算更加严格.

4 相关定理

4.1 约束容差交运算定理

命题 1(约束容差交运算定理). 给定不完备信息系统 $\boldsymbol{I}=(U,A,V,f)$,非空集合 $X\subseteq U$ 的约束容差属性二元组集合为 $S(X)$,对于 $X_1\subseteq U$ 和 $X_2\subseteq U$ 且 $X_1\cap X_2=\varnothing$,如果 $S(X_1)\neq\varnothing$ 且 $S(X_2)\neq\varnothing$,则

$$S(X_1\cup X_2)=S(X_1)\bigcap{}^{C}S(X_2) \tag{11}$$

证明. 因为 $S(X_1)\neq\varnothing$ 且 $S(X_2)\neq\varnothing$,则根据式(4)有 $S(X_1)=V(X_1)$ 且 $S(X_2)=V(X_2)$,再根据容差属性二元组集合相关定义式(1)~(3)及容差交运算相关定义式(8),(9)可得

$$\begin{aligned}
&V(X_1\cup X_2)\\
&=\{(k,a_k(X_1\cup X_2))\mid\\
&\quad\forall_{x_i\in X_1\cup X_2,x_j\in X_1\cup X_2}a_k(x_i)=a_k(x_j)\vee\\
&\quad a_k(x_i)=\text{“}*\text{”}\vee a_k(x_j)=\text{“}*\text{”}\}\\
&=\{(k,\max(a_k(X_1),a_k(X_2)))\mid\\
&\quad(k,a_k(X_1))\in V(X_1)\wedge(k,a_k(X_2))\in V(X_2)\wedge\\
&\quad(a_k(X_1)=a_k(X_2)\vee a_k(X_1)=\\
&\quad\text{“}*\text{”}\vee a_k(X_2)=\text{“}*\text{”})\}\\
&=\{(k,\max(a_k(X_1),a_k(X_2)))\mid\\
&\quad(k,a_k(X_1))\in S(X_1)\wedge(k,a_k(X_2))\in S(X_2)\wedge\\
&\quad(a_k(X_1)=a_k(X_2)\vee a_k(X_1)=\\
&\quad\text{“}*\text{”}\vee a_k(X_2)=\text{“}*\text{”})\}\\
&=S(X_1)\bigcap{}^{T}S(X_2).
\end{aligned}$$

进一步根据约束容差属性二元组集合定义式(4)及约束容差交运算定义式(10)可得

$$S(X_1\cup X_2)=S(X_1)\bigcap{}^{C}S(X_2).$$ 证毕.

4.2 约束容差集合精简计算定理

命题 2(约束容差集合精简计算定理). 给定不完备信息系统 $\boldsymbol{I}=(U,A,V,f)$,非空集合 $X\subseteq U$ 的约束容差集合精简为 $\boldsymbol{R}(X)=(|X|,S(X),D(X))$,对于 $X_1\subseteq U$ 和 $X_2\subseteq U$,$X_1\cap X_2=\varnothing$,$S(X_1)\neq\varnothing$ 且 $S(X_2)\neq\varnothing$,如果 $S(X_1)\bigcap{}^{C}S(X_2)\neq\varnothing$,则

$$\boldsymbol{R}(X_1\cup X_2)=(|X_1\cup X_2|,S(X_1\cup X_2),D(X_1\cup X_2)),$$

其中,

$$\begin{cases}
|X_1\cup X_2|=|X_1|+|X_2|\\
S(X_1\cup X_2)=S(X_1)\bigcap{}^{C}S(X_2)\\
D(X_1\cup X_2)=\\
\qquad\dfrac{m-|S(X_1)\bigcap{}^{C}S(X_2)|}{\sqrt{|X_1|+|X_2|}\times|S(X_1)\bigcap{}^{C}S(X_2)|}
\end{cases} \tag{12}$$

证明.

(1)因为 $Y_1\subseteq X$ 和 $Y_2\subseteq X$ 且 $Y_1\bigcap Y_2=\varnothing$,所以 $|Y_1\cup Y_2|=|Y_1|+|Y_2|$.

(2)根据命题 1 有 $S(X_1\cup X_2)=S(X_1)\bigcap{}^{C}S(X_2)$.

(3)$D(X_1\cup X_2)$

$$\begin{aligned}
&=\frac{m-|S(X_1\cup X_2)|}{\sqrt{|X_1\cup X_2|}\times|S(X_1\cup X_2)|}\\
&=\frac{m-|S(X_1)\bigcap{}^{C}S(X_2)|}{\sqrt{|X_1|+|X_2|}\times|S(X_1)\bigcap{}^{C}S(X_2)|}.
\end{aligned}$$

证毕.

根据该定理,MIBOI 算法在将两个不相交的对象集合归入一类时,可以直接精确地进行约束容差集合精简的计算.因此,约束容差集合精简在处理不完备数据集时不仅可以大规模降低数据处理量,同时可以保证约束容差集合差异度计算的精确性.

4.3 约束容差集合精简不变定理

命题 3(约束容差集合精简不变定理). 给定不完备信息系统 $\boldsymbol{I}=(U,A,V,f)$,非空集合 $X\subseteq U$ 的约束容差集合精简为 $\boldsymbol{R}(X)=(|X|,S(X),D(X))$,$S(X)\neq\varnothing$,对于 $\forall(k,a_k(X))\in S(X)$,$\forall x_i\in X$,令

$$a_k(x_i^e)=\begin{cases}a_k(x_i),a_k(X)=\text{“}*\text{”}\vee a_k(x_i)\neq\text{“}*\text{”}\\ a_k(X),a_k(X)\neq\text{“}*\text{”}\wedge a_k(x_i)=\text{“}*\text{”}\end{cases} \tag{13}$$

缺失数据填补后的约束容差集合精简为

$$\boldsymbol{R}(X^e)=(|X^e|,S(X^e),D(X^e)) \tag{14}$$

则有

$$R(X^e)=R(X) \tag{15}$$

证明.

(1)因为式(13)进行的属性值变换不影响对象的数目,所以 $|X^e|=|X|$.

(2)对于式(13)的第 1 种情况,$a_k(x_i^e)=a_k(x_i)$,属性值不变,所以 $S(X^e)$ 不变.对于式(13)的第 2 种情况是在约束容差属性取值不为“$*$”($a_k(X)\neq$“$*$”),即容差值为明确值时,将各对象在该属性为“$*$”的值($a_k(x_i)=$“$*$”,即缺失值)用容差值替换,即 $a_k(x_i^e)=a_k(X)$.根据容差值定义式(2),替换后的 X^e 中所有对象在该属性的取值都为原来的容差值 $a_k(X)$,容差值不变,所以 $S(X^e)$ 不变.综合两种情况,$S(X^e)=S(X)$.

(3)根据式(5)在计算约束容差集合差异度 $D(X^e)$ 时需要用到 m,$|X^e|$ 和 $|S(X^e)|$.由于属性数目 m 为已知常数,$|X^e|=|X|$,$S(X^e)=S(X)$,所以 $D(X^e)=D(X)$.

综上所述,$\boldsymbol{R}(X^e)=\boldsymbol{R}(X)$. 证毕.

该定理是基于不完备数据聚类进行缺失数据填补的重要基础.根据该定理,对于约束容差属性,将缺失值用明确的容差值替换后,约束容差集合精简不变,约束容差集合差异度不变,所以缺失数据的填补能较好地保持原有不完备数据的聚类特征.

5　算法描述

MIBOI 算法基于不完备数据聚类对缺失数据进行填补.在聚类过程中,从采用第一个对象创建第一个类开始,通过一次扫描不完备数据对象完成全部新类的创建及对象到类的归并.对于已创建的类,仅保留约束容差集合精简,不保留全部对象的信息.是否创建新类取决于预先指定的约束容差集合差异度上限 u.对于每一个被扫描到的对象,寻找使得并入后约束容差集合差异度最小的已创建类.如果该最小约束容差集合差异度不大于上限 u,则并入该类;否则创建新类.在聚类完成后逐一对各个类进行缺失数据的填补.对于每个约束容差属性,如果其容差值不为"*",即容差值为明确值而不是空值时,将该类中各对象在该属性为"*"的值(即缺失值)用该容差值替换,实现缺失数据的填补.由于不一定所有的属性都是约束容差属性,且约束容差属性的容差值也可能是空值,而 MIBOI 算法只对容差值不为空的约束容差属性进行缺失数据填补,所以算法不能保证缺失数据填补后的信息系统是完备的.

MIBOI 算法步骤如下所述.

1. 对不完备信息系统 $\boldsymbol{I}=(U,A,V,f)$,根据约束容差集合精简定义式(6),(7)得到由第一个对象创建的类 $X_1=\{x_1\}$ 的约束容差集合精简 $\boldsymbol{R}(X_1)=(1,S(X_1),0)$,类似地得到 $\boldsymbol{R}(\{x_2\})=(1,S(\{x_2\}),0)$;

2. 根据约束容差交运算式(8)～(10)计算 $S(X_1)\cap^C S(\{x_2\})$,如果 $S(X_1)\cap^C S(\{x_2\})=\varnothing$,创建新类 $X_2=\{x_2\}$,类的数目 $c=2$,转入步 4;

3. 根据约束容差集合精简计算定理,应用式(12)计算 X_1 和 $\{x_2\}$ 合并后的约束容差集合精简 $\boldsymbol{R}(X_1\cup\{x_2\})=(2,S(X_1\cup\{x_2\}),D(X_1\cup\{x_2\}))$,如果约束容差集合差异度 $D(X_1\cup\{x_2\})\leq u$,将 x_2 并入已创建类 X_1 中,$X_1=\{x_1,x_2\}$,类的数目 $c=1$;否则,创建新类 $X_2=\{x_2\}$,类的数目 $c=2$;

4. $i=3$;

5. 如果 $i>n$,转到步 8;否则,对于每一个已创建类 X_t,$t=1,2,\cdots,c$,根据约束容差交运算式(8)～(10)计算 $S(X_t)\cap^C S(\{x_i\})$,如果 $S(X_t)\cap^C S(\{x_i\})=\varnothing$ 对 $t=1,2,\cdots,c$ 都成立,则创建新类 $X_{c+1}=\{x_i\}$,类的数目 $c=c+1$,转入步 7;

6. 对于 $t=1,2,\cdots,c$ 且 $S(X_t)\cap^C S(\{x_i\})\neq\varnothing$,根据约束容差集合精简计算定理,应用式(12)计算 X_t 和 $\{x_i\}$ 合并后的约束容差集合精简 $\boldsymbol{R}(X_t\cup\{x_i\})=(|X_t|+1,S(X_t\cup\{x_i\}),D(X_t\cup\{x_i\}))$,寻找 τ 使得 X_τ 和 $\{x_i\}$ 合并后的约束容差集合差异度最小,即 $D(X_\tau\cup\{x_i\})=\min_{(t=1,2,\cdots,c)\wedge S(X_t)\cap^C S(\{x_i\})\neq\varnothing} D(X_t\cup\{x_i\})$,如果约束容差集合差异度 $D(X_\tau\cup\{x_i\})\leq u$,将 x_i 并入已创建类 X_τ 中,$X_\tau=X_\tau\cup\{x_i\}$,类的数目不变;否则,创建新类 $X_{c+1}=\{x_i\}$,类的数目 $c=c+1$;

7. $i=i+1$,转到步 5;

8. $t=1$;

9. 如果 $t>c$,转到步 11;否则,对于类 X_t,如果 $S(X_t)\neq\varnothing$,根据约束容差集合精简不变定理,应用式(13)的第 2 种情况对缺失数据进行填补,即对于 $\forall(k,a_k(X_t))\in S(X_t)$,$\forall x_i\in X_t$,如果 $a_k(X_t)\neq$"*"且 $a_k(x_i)=$"*",则将 $a_k(x_i)$ 的值填补为 $a_k(X_t)$;

10. $t=t+1$,转到步 9;

11. 结束.

从上述步骤可知 MIBOI 算法具有如下特点:

(1)基于不完备数据的聚类结果对缺失数据进行填补.根据约束容差集合精简不变定理,在缺失数据填补后,各个类的约束容差集合精简保持不变.约束容差集合精简包含 3 个分量:集合的基数,约束容差属性二元组集合和约束容差集合差异度,这种不变性使缺失数据的填补能较好地保持原有不完备数据的聚类特征;

(2)完成不完备数据聚类仅需扫描数据一次,在聚类过程中只需保留已创建各类的约束容差集合精简,而不必存储全部对象的所有信息,对数据进行了高度压缩,数据处理量显著降低,并且约束容差集合精简计算定理保证了在合并两个不相交的对象集合时约束容差集合精简可以直接进行精确的计算,保证了约束容差集合差异度计算的精确性,因此这种数据压缩不会降低不完备数据聚类的质量,从而保证了缺失数据填补的质量;

(3)算法的主要计算时间在于不完备数据的聚类.由于在此过程中不必计算两两对象间的距离,扫描到的对象只需与已创建的各个类进行并入后约束容差集合精简的计算以确定扫描到的对象是并入已创建的某个类中还是创建一个新类,算法的计算时间复杂度为 $O(cm_cn)$,其中 c 为类的数目,m_c 为平均约束容差属性数目,n 为对象数目.在实际的数据挖掘应用中 c 和 m_c 一般远小于 n.

6　实验结果

为了检验 MIBOI 算法进行缺失数据填补的效率和效果,从 UCI 机器学习数据集中选择 4 个常用的分类变量的基准数据集针对 MIBOI 算法和经典

的 ROUSTIDA 算法及属性均值填补算法进行对比实验. 实验硬件环境为: Intel(R) Core(TM)2 Duo CPU E7400@2.80GHz 处理器, 2.00GB 内存; 软件环境为: Windows XP 操作系统, Microsoft Visual C++2008 Express Edition 编程语言.

6.1 数据集

采用 UCI 数据集中著名的 Small Soybean、Zoo、SPECT Heart(train)、Image Segmentation 数据集进行 MIBOI 算法的有效性检验. Small Soybean 数据集中包含 47 个对象, 共分为 4 类, 每类对应一种大豆作物病害, 描述对象的属性共有 35 个, 各属性的取值都可以作为分类变量. Zoo 数据集中包含 101 个对象, 共分为 7 类, 每类对应一种动物类型, 描述对象的属性共有 16 个, 其中 15 个为二态属性, 值域为{0,1}, 还有 1 个为数值属性, 描述动物腿的数目, 值域为{0,2,4,5,6,8}, 可以视为分类变量. SPECT Heart(train)数据集中包含 80 个对象, 共分为 2 类, 将病人分为正常和不正常, 描述对象的属性共有 22 个, 都是对心脏单质子发射计算机层析成像(Single Proton Emission Computed Tomography, SPECT)图像诊断数据处理后得到的二态属性, 值域为{0,1}. Image Segmentation 数据集一共包含 2310 个对象, 19 个属性, 共分为 7 类, 描述 7 种户外图像的像素属性. 本文从每个类中随机抽取了 50 个对象, 共 350 个对象, 值域为{0,1}二态属性的数据作为实验对象.

为了充分检验 MIBOI 算法进行缺失数据填补的有效性, 将完备的 Small Soybean、Zoo、SPECT Heart (train)、Image Segmentation 数据集分别从 5%的缺失比率开始随机将数据置为"*", 并按 5%的缺失比率递增, 直至缺失比率达到 70%. 针对每一个数据集的每一种数据缺失比率, 分别采用 MIBOI 算法、经典的 ROUSTIDA 算法、属性均值填补法(MEANS, 用数值属性均值或分类属性众数来替代缺失数据)及其结合方法进行缺失数据的填补. 每次实验中, 对象随机排序, 缺失数据随机生成, 对 100 次随机实验结果进行对比. 从补齐率、填补正确率、填补后聚类正确率、时间效率和参数分析 5 个方面评价缺失数据填补算法的性能.

表 1 数据集描述

	数据集	对象数	属性数	类别数
1	Small Soybean	47	35	4
2	Zoo	101	16	7
3	SPECT Heart(train)	80	22	2
4	Image Segmentation	350	19	7

6.2 算法补齐率分析

对不完备信息系统 $\boldsymbol{I}=(U,A,V,f)$, 应用算法进行缺失数据填补, 记对象个数为 N, 数据总量为 M, 缺失数据的比率为 α, 实验运行次数为 n, 时间消耗总计为 T, 正确填补数据总量为 C, 错误填补数据总量为 F.

补齐率 CR(Complete Rate)为平均每次运行填补数据量占缺失数据总量的比率. 平均每次运行填补数据量为平均每次运行正确填补数据量和平均每次运行错误填补数据量之和

$$CR=(C/n+F/n)/(M\times\alpha) \tag{16}$$

利用式(16)计算单独利用 MIBOI 和 ROUSTIDA 针对不同数据集在不同缺失比率下的补齐率, 结果如表 2 和图 1 所示. MIBOI 的补齐率明显高于 ROUSTIDA.

表 2 补齐率随数据缺失比率的变化

缺失比率/%	补齐率/%							
	Small Soybean		Zoo		SPECT Heart(train)		Image Segmentation	
	MIBOI	ROUSTIDA	MIBOI	ROUSTIDA	MIBOI	ROUSTIDA	MIBOI	ROUSTIDA
5	85.32	3.29	84.63	57.87	79.13	29.13	95.65	77.54
10	87.13	4.21	86.22	62.11	82.74	32.22	96.90	73.33
15	86.89	7.68	85.28	65.12	78.63	36.27	96.66	72.69
20	85.63	14.41	86.12	65.02	79.28	40.87	96.83	68.82
25	84.30	26.01	87.62	65.52	79.20	45.24	96.89	66.39
30	83.83	27.77	87.38	56.28	79.28	48.61	96.85	62.03
35	83.09	52.82	86.68	50.28	79.49	50.94	97.08	52.80
40	82.35	60.94	87.47	37.97	81.86	49.82	97.57	42.41
45	80.76	72.15	87.80	28.50	80.41	47.11	97.35	30.50
50	80.79	73.64	88.12	20.31	81.01	41.53	97.35	19.84
55	79.45	74.90	87.97	13.93	81.95	33.72	97.45	11.38
60	78.69	68.42	88.34	8.15	82.25	26.23	97.55	7.28
65	77.90	59.90	88.53	5.92	83.02	19.56	97.61	4.96
70	76.85	54.41	89.22	4.67	83.09	15.25	97.56	4.18

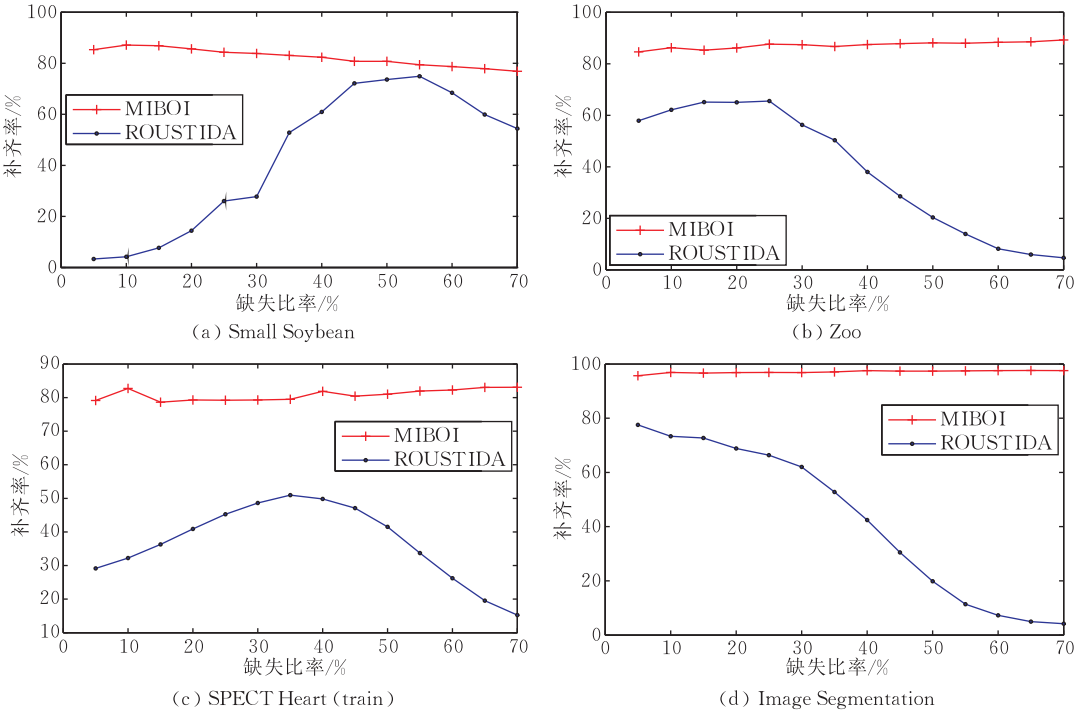


图 1　补齐率随数据缺失比率的变化

6.3　算法填补正确率分析

算法填补正确率(Accuracy Rate)为平均每次运行正确填补数据量与缺失数据总量的比率，计算公式如下：

$$AR = (C/n)/(M \times \alpha) \tag{17}$$

对于采用 MIBOI 和 ROUSTIDA 两种算法得到的不完备信息系统，继续采用均值填补法进行填补，形成完备信息系统，再利用式(17)计算填补正确率. 分别采用均值填补(MEANS)、先 ROUSTIDA 再均值填补(RA＋MS)、先 MIBOI 再均值填补(MI＋MS)及先 ROUSTIDA 再 MIBOI 和均值填补(RA＋MI＋MS)共 4 种方法得到完备信息系统进行填补正确率比较.

利用式(17)计算 4 种方法针对不同数据集在不同缺失比率下的填补正确率，结果如表 3 和图 2 所示. 总体上看，4 种方法填补正确率的高低顺序依次为：RA＋MI＋MS＞MI＋MS＞RA＋MS＞MS. 可以看出：为了得到完备信息系统，采用 MI＋MS 方法在填补正确率指标上一般优于 RA＋MS 方法及直接采用 MEANS 方法. 而采用 RA＋MI＋MS 方法一般更优于 MI＋MS 方法.

表 3　填补正确率随数据缺失比率的变化

缺失比率/%	填补正确率/%															
	Small Soybean				Zoo				SPECT Heart(train)				Image Segmentation			
	MEANS	RA＋MS	MI＋MS	RA＋MI＋MS	MEANS	RA＋MS	MI＋MS	RA＋MI＋MS	MEANS	RA＋MS	MI＋MS	RA＋MI＋MS	MEANS	RA＋MS	MI＋MS	RA＋MI＋MS
5	74.02	74.51	86.02	86.24	65.38	80.63	85.75	86.15	77.95	78.41	87.36	90.09	68.89	90.79	98.29	98.29
10	75.24	75.30	85.77	85.48	68.20	83.54	85.85	85.80	78.18	78.18	81.45	82.68	71.67	91.51	97.16	97.46
15	75.28	75.53	86.16	86.23	68.26	83.18	84.50	84.65	77.05	75.23	79.06	79.26	65.34	88.20	95.66	96.33
20	73.95	75.59	83.81	84.02	69.38	84.15	83.43	84.68	77.44	76.36	79.59	79.89	67.42	89.29	95.72	96.30
25	73.80	76.79	83.45	83.76	68.56	81.51	82.27	82.68	76.14	77.23	77.81	78.55	67.17	87.05	93.60	94.35
30	73.57	77.00	82.68	83.18	68.86	78.68	81.13	81.58	75.53	76.59	77.59	77.68	67.46	83.57	93.75	94.39
35	73.15	78.87	80.78	82.42	67.88	75.58	79.56	80.56	75.52	75.42	76.72	76.98	67.46	82.22	92.73	93.38
40	73.48	79.04	80.14	81.84	68.85	73.48	78.21	78.47	75.11	76.51	77.63	78.17	68.26	77.39	91.45	92.07
45	73.76	79.64	79.08	80.87	68.36	70.84	76.97	76.79	75.00	76.01	76.40	76.65	67.56	73.16	91.33	91.78
50	73.49	77.38	77.02	78.88	67.56	68.68	75.09	75.18	75.07	76.57	78.15	78.61	68.32	70.80	90.18	90.49
55	72.58	77.58	75.92	78.76	67.77	68.73	72.92	72.67	75.83	77.09	78.50	79.12	67.42	68.69	88.26	88.38
60	73.07	75.61	75.00	77.42	67.46	67.59	71.17	70.64	75.59	76.95	77.10	77.64	67.61	67.90	86.66	86.75
65	73.39	73.93	73.74	74.71	67.66	67.67	69.31	69.13	73.65	74.51	74.83	75.11	68.08	68.15	84.49	84.50
70	73.23	73.87	72.39	73.24	67.43	67.31	67.48	67.29	73.15	73.94	75.18	75.33	67.53	67.54	82.74	82.74

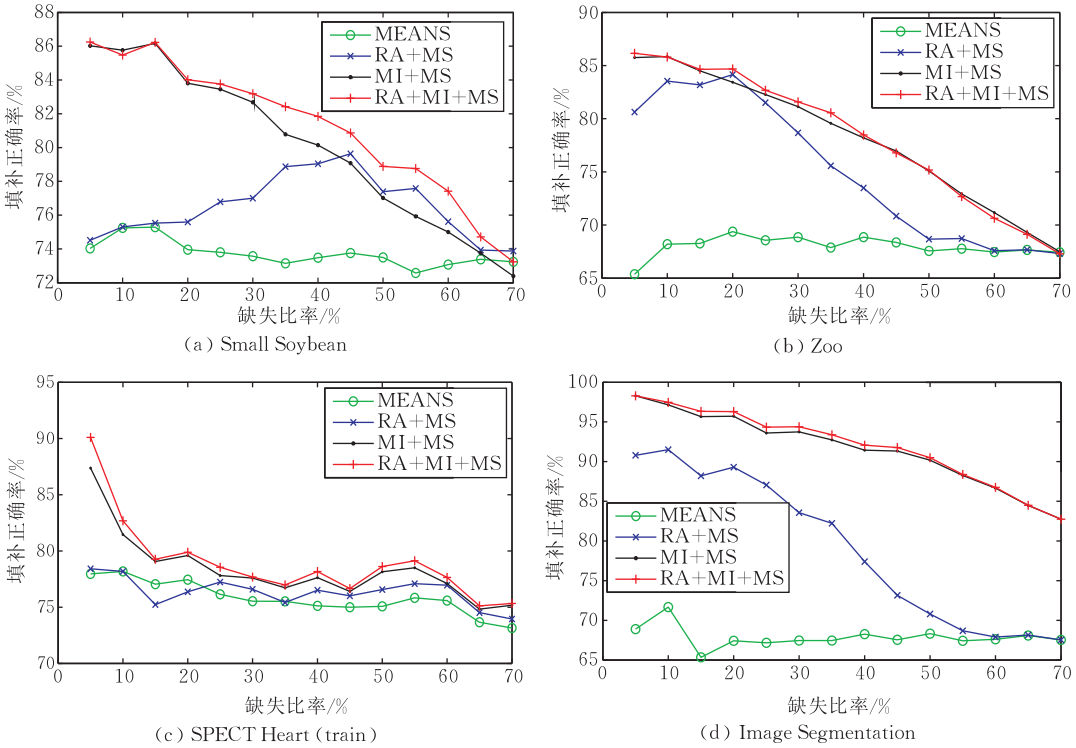


图 2 填补正确率随数据缺失比率的变化

6.4 填补后聚类正确率分析

利用填补后的聚类正确率从另一个角度分析填补的质量，聚类正确率公式为

$$p = \left(\sum_{i=1}^{k} b_i\right)/N \tag{18}$$

其中，N 表示对象的总个数，b_i 表示正确聚类到第 i 类的对象数，k 表示聚类个数. 同上一小节，本节对 MEANS、RA+MS、MI+MS 及 RA+MI+MS 共 4 种方法进行填补后形成的完备信息系统，运用经典 K-Modes[19] 聚类算法比较分析其聚类正确率.

利用式(18)计算 4 种方法针对不同数据集在不同缺失比率下的填补后的聚类正确率，结果如表 4 和图 3 所示. 总体上看，4 种方法填补后聚类正确率的高低顺序依次为 RA+MI+MS>MI+MS>RA+MS>MS. 其中，在缺失比率较低(小于 20%)的情况下，4 种方法的聚类效果差距不明显. 但随着缺失比率的增加，RA+MI+MS 和 MI+MS 对于 MEANS 和 RA+MS 的优势增大，这种趋势针对后 3 个数据集的结果更明显.

表 4 填补后聚类正确率随数据缺失比率的变化

缺失比率/%	填补后聚类正确率/%															
	Small Soybean				Zoo				SPECT Heart(train)				Image Segmentation			
	MEANS	RA+MS	MI+MS	RA+MI+MS	MEANS	RA+MS	MI+MS	RA+MI+MS	MEANS	RA+MS	MI+MS	RA+MI+MS	MEANS	RA+MS	MI+MS	RA+MI+MS
5	76.72	76.81	78.81	78.77	76.14	78.02	77.52	78.33	61.60	62.61	62.30	62.35	72.49	77.26	77.27	77.57
10	74.81	74.89	77.06	77.79	73.71	76.25	76.70	78.02	61.55	62.98	63.25	63.35	69.77	74.89	75.09	76.09
15	72.09	72.85	75.36	75.02	73.76	76.05	75.88	76.91	60.45	61.69	62.75	64.35	68.86	71.06	72.86	72.36
20	69.06	69.91	76.49	75.34	70.49	74.52	75.32	75.72	58.45	59.65	61.85	63.05	67.34	69.09	69.63	70.74
25	67.32	70.70	76.38	75.89	66.56	73.88	74.65	75.39	56.85	57.79	60.75	61.95	65.57	67.80	69.26	70.34
30	63.57	69.53	74.89	76.72	62.95	70.73	72.93	74.38	54.50	55.43	60.00	60.50	63.11	64.86	68.83	69.09
35	60.15	71.15	72.94	75.66	58.72	68.04	72.45	72.34	52.75	53.60	59.25	59.85	61.73	63.15	68.95	70.48
40	56.85	67.79	69.77	72.79	57.37	62.82	71.23	70.26	51.32	52.14	58.45	58.85	58.71	61.26	67.34	68.21
45	55.64	70.28	68.87	71.83	54.51	57.35	69.50	69.82	50.25	50.98	57.84	58.42	57.09	58.26	65.46	66.66
50	51.47	66.28	63.98	68.94	53.36	53.75	68.06	67.43	49.85	50.37	57.81	58.25	55.21	56.46	65.87	66.79
55	48.74	61.19	60.17	66.68	51.12	51.49	62.40	61.67	47.25	47.57	57.35	57.25	52.43	53.67	65.06	65.42
60	47.13	57.34	57.60	64.19	50.28	50.50	60.22	59.58	46.51	46.63	55.56	55.55	51.37	51.67	65.15	65.37
65	45.60	47.74	53.23	56.28	50.47	50.47	57.15	56.06	46.95	47.05	53.70	54.00	49.18	50.26	62.48	62.50
70	44.70	45.94	46.64	50.28	48.94	48.92	51.59	51.40	46.55	46.60	53.50	53.60	48.12	48.30	60.82	61.22

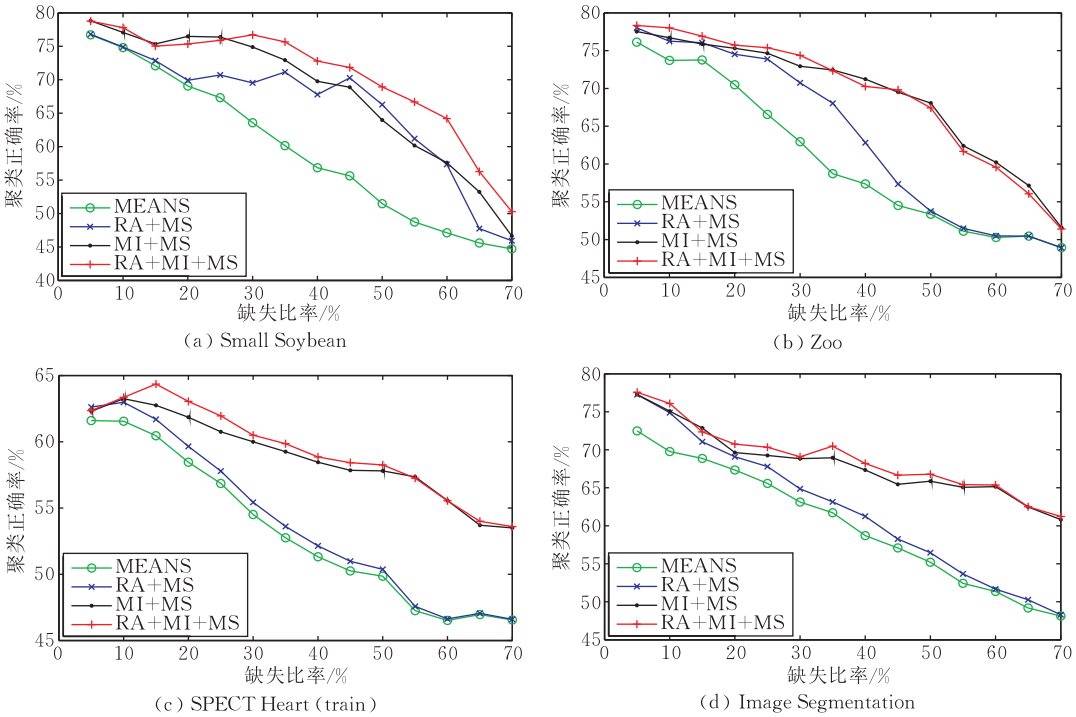


图 3　填补后聚类正确率随数据缺失比率的变化

6.5　算法时间效率分析

算法时间效率指标选用平均时间(Average Time,AT)：

$$AT = T/n \tag{19}$$

平均时间为平均每次运行时间消耗.

上两节选择的 4 种方法的基础是 ROUSTIDA、MIBOI 和 MEANS 这 3 种算法. 由于 MEANS 算法只是根据某个属性纵向计算众数填补，无复杂的数据迭代过程，因此处理时间与 MIBOI 和 ROUSTIDA 相比很小，故不予考虑. 通过 ROUSTIDA 和 MIBOI 两种算法的效率可以扩展计算上述 4 种方法的效率.

利用式(19)计算 MIBOI 和 ROUSTIDA 针对不同数据集在不同缺失比率下的平均时间，结果如表 5 和图 4 所示. 除了针对 Small Soybean 数据集在缺失比率为 5%的情况下，MIBOI 的平均时间略高于 ROUSTIDA 以外，MIBOI 的平均时间都明显低于 ROUSTIDA. 而针对 Small Soybean 数据集在缺失比率为 5%的情况下，MIBOI 的平均时间略高于 ROUSTIDA 的原因，是在此情况下 MIBOI 的平均每次运行填补数据量远高于 ROUSTIDA. 因此，MIBOI 的时间效率明显优越于 ROUSTIDA，进而有 MI+MS 的时间效率优于 RA+MS，RA+MI+MS 时间效率与 RA+MS 差别不明显.

表 5　平均时间随数据缺失比率的变化

缺失比率/%	平均时间/s							
	Small Soybean		Zoo		SPECT Heart(train)		Image Segmentation	
	MIBOI	ROUSTIDA	MIBOI	ROUSTIDA	MIBOI	ROUSTIDA	MIBOI	ROUSTIDA
5	0.0131131	0.0129556	0.0162254	0.0927672	0.0192329	0.0735226	0.0619321	0.5218761
10	0.0117771	0.0176370	0.0166684	0.1208486	0.0171604	0.0843587	0.0616965	0.6440620
15	0.0087403	0.0219360	0.0175031	0.1457246	0.0173548	0.0893062	0.0625065	0.7856314
20	0.0093193	0.0243008	0.0167371	0.1456594	0.0161800	0.0924595	0.0659766	0.9251275
25	0.0094550	0.0315406	0.0156688	0.1521954	0.0156049	0.1131494	0.0665631	1.0993335
30	0.0093229	0.0365858	0.0164809	0.1533735	0.0157433	0.1023212	0.0632263	1.1785693
35	0.0084011	0.0435361	0.0164376	0.1586361	0.0152395	0.1218758	0.0639240	1.2753069
40	0.0087856	0.0462890	0.0155978	0.1589187	0.0169334	0.1414599	0.0655622	1.3288523
45	0.0078838	0.0474545	0.0154916	0.1628987	0.0159857	0.1446286	0.0597393	1.3836006
50	0.0080644	0.0487682	0.0152555	0.1449884	0.0167692	0.1443256	0.0627980	1.3164211
55	0.0072038	0.0517491	0.0153124	0.1427352	0.0140706	0.1464181	0.0614751	1.1922182
60	0.0064598	0.0548814	0.0133016	0.1342913	0.0145347	0.1330675	0.0595110	0.9574924
65	0.0063797	0.0508105	0.0127699	0.1436160	0.0134001	0.1248297	0.0551150	0.8367161
70	0.0059796	0.0537372	0.0128770	0.1457940	0.0133480	0.1415762	0.0524282	0.6286149

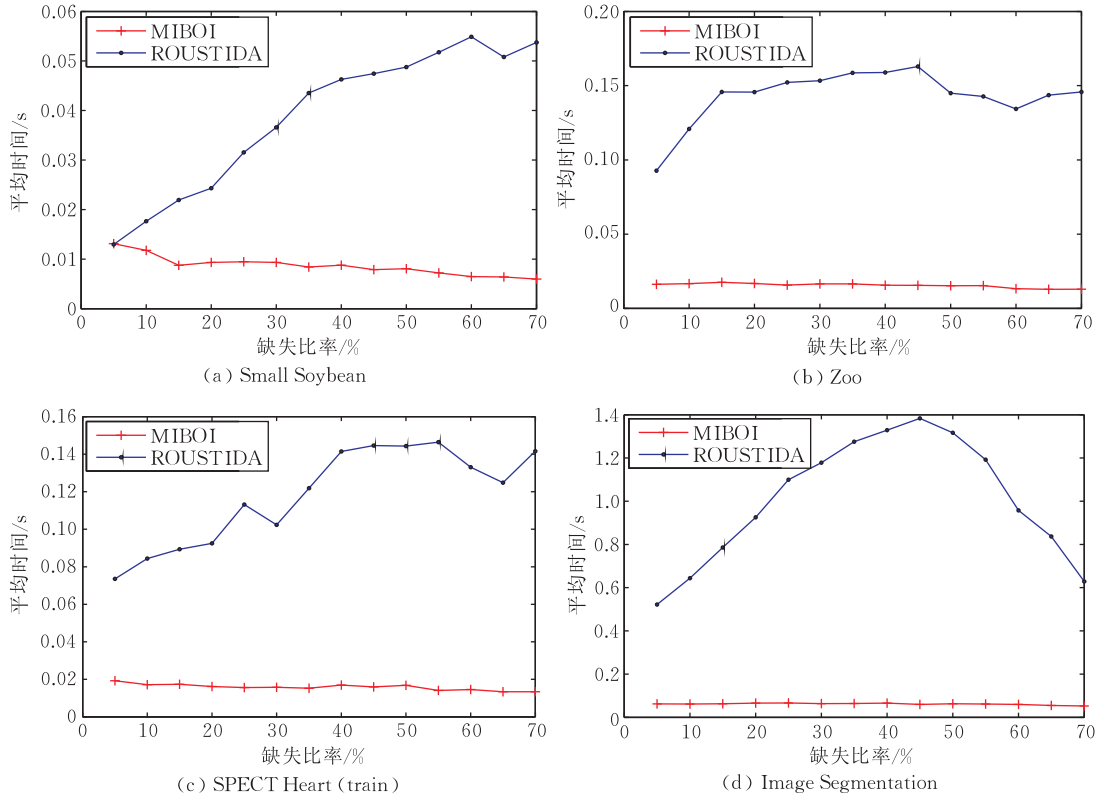


图 4 平均时间随数据缺失比率的变化

6.6 算法参数分析

MIBOI 算法有一个需要预先指定的参数，即约束容差集合差异度上限 u. 每次随机实验中，MIBOI 算法的约束容差集合差异度上限 u 皆取使得缺失数据填补绝对正确率达到最高的最佳值.

针对不同数据集在不同缺失比率下 100 次随机实验，u 最佳值的平均值、标准差和标准差系数结果如表 6、图 5 所示. 从针对 Small Soybean、Zoo、SPECT Heart(train)这 3 个数据集(由于 Image Segmentation 数据集同一个类内差异很小的对象较多，实验结果显示 u 最佳取值均为 0，故本部分不讨论该数据集)的实验结果看：对于每 100 次随机实验，参数 u 最佳值的平均值都随着数据缺失比率的增加而降低. 标准差随着数据缺失比率的增加没有统一的规律，3 个数据集的标准差系数随着缺失比率的增加而增加.

表 6 u 最佳值的平均值和标准差系数随数据缺失比率的变化

缺失	Small Soybean			Zoo			SPECT Heart(train)		
比率/%	均值	标准差	标准差系数/%	均值	标准差	标准差系数/%	均值	标准差	标准差系数/%
5	0.1170	0.0149	12.74	0.0854	0.0320	37.47	0.2074	0.0498	24.01
10	0.1151	0.0151	13.12	0.0845	0.0257	30.41	0.1748	0.0366	20.94
15	0.1141	0.0148	12.97	0.0783	0.0220	28.10	0.1620	0.0430	26.54
20	0.1113	0.0172	15.45	0.0766	0.0235	30.68	0.1482	0.0307	20.72
25	0.1120	0.0177	15.80	0.0713	0.0242	33.94	0.1215	0.0294	24.20
30	0.1083	0.0205	18.93	0.0697	0.0209	29.99	0.1239	0.0267	21.55
35	0.1080	0.0242	22.41	0.0654	0.0222	33.95	0.1060	0.0317	29.91
40	0.1013	0.0214	21.13	0.0528	0.0221	41.86	0.0938	0.0299	31.88
45	0.1000	0.0263	26.30	0.0563	0.0230	40.85	0.0841	0.0242	28.78
50	0.0989	0.0265	26.79	0.0463	0.0240	51.84	0.0742	0.0288	38.81
55	0.0988	0.0306	30.97	0.0392	0.0220	56.12	0.0634	0.0267	42.11
60	0.0958	0.0312	32.57	0.0380	0.0227	59.74	0.0514	0.0271	52.72
65	0.0946	0.0314	33.19	0.0260	0.0224	86.15	0.0393	0.0266	67.68
70	0.0924	0.0291	31.49	0.0227	0.0209	92.07	0.0284	0.0228	80.28

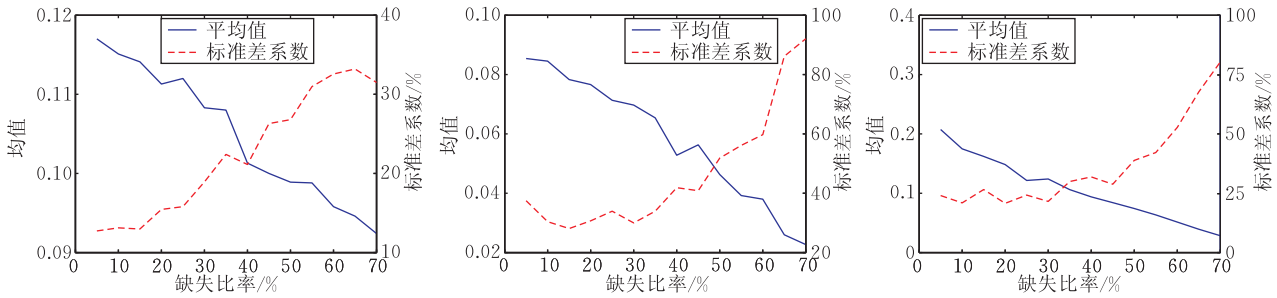


图 5　u 最佳值的平均值和标准差系数随数据缺失比率的变化

7　结束语

在数据挖掘实际应用中面临的往往是不完备数据集，对缺失数据的处理已成为数据预处理中一项非常重要的工作. 针对分类变量缺失数据处理问题，本文未采用传统的概率统计学方法，而是基于不完备数据聚类提出一种缺失数据填补的新方法 MIBOI. 该方法通过定义约束容差集合差异度，从集合的角度判断不完备数据对象的总体相异程度，进而根据不完备数据聚类的结果实现缺失数据的填补. 为了检验所提出的 MIBOI 方法的有效性，采用 UCI 机器学习数据集中 4 个常用的基准数据集进行实验，分别从算法补齐率、填补正确率、填补后聚类正确率和时间效率 4 个方面将 MIBOI 与同样未采用传统概率统计学方法的经典不完备数据分析方法 ROUSTIDA 及采用统计学方法的 MEANS 方法之间的结合方法进行对比分析. 结果表明，MIBOI 从算法补齐率和时间效率两个方面总体优于经典 ROUSTIDA 方法，MIBOI 与 MEANS 结合的 MI+MS 在填补正确率和填补后聚类正确率两方面优于 ROUSTIDA 与 MEANS 结合的 RA+MS 及 MEANS 单独填补方法，而先 ROUSTIDA 再 MIBOI 和均值填补(RA+MI+MS)具有最好的结果，并且在时间效率上没有明显的劣势.

进一步的研究工作：MIBOI 方法有一个需要预先指定的参数，即约束容差集合差异度上限 u，缺失数据填补的结果受该参数 u 的影响. 每次随机实验中，u 都是经过尝试选取使得正确率达到最高的最佳值. 总体而言，u 最佳值的平均值随着数据缺失比率的增加而降低，标准差系数随着缺失比率的增加而增加. 在实际应用中 u 最佳值的预先指定还是有一定困难的. 如何高效合理地确定参数 u 或彻底地消除参数 u 的影响是需要继续研究的一个内容.

参　考　文　献

[1] Garcia-Laencina P J, Sancho-Gomez J L, Figueiras-Vidal A R, Verleysen M. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing, 2009, 72(7-9): 1483-1493

[2] Gu Yu, Yu Ge, Li Xiao-Jing, Wang Yi. RFID data interpolation algorithm based on dynamic probabilistic path-event model. Journal of Software, 2010, 21(3): 438-451 (in Chinese)
(谷峪，于戈，李晓静，王义. 基于动态概率路径事件模型的 RFID 数据填补算法. 软件学报，2010，21(3)：438-451)

[3] Tseng S, Wang K, Lee C. A pre-processing method to deal with missing values by integrating clustering and regression techniques. Applied Artificial Intelligence, 2003, 17(5-6): 535-544

[4] Wang H, Wang S. Discovering patterns of missing data in survey databases: An application of rough sets. Expert Systems with Applications, 2009, 36(3): 6256-6260

[5] Little R, Rubin D. Statistical Analysis with Missing Data. New York, NY: John Wiley & Sons Inc, 2002

[6] Han J, Kamber M. Data Mining Concepts and Techniques. 2nd Edition. San Diego, USA: Academic Press, 2006

[7] Nakagawa S, Freckleton R P. Missing inaction: The dangers of ignoring missing data. Trends in Ecology and Evolution, 2008, 23(11): 592-296

[8] Donders A R T, Heijden G J M G, Stijnen T, Moons K G M. Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology, 2006, 59(10): 1087-1091

[9] Penn D A. Estimating missing values from the general social survey: An application of multiple imputation. Social Science Quarterly, 2007, 88(2): 573-584

[10] Gebregziabher M, DeSantis S M. Latent class based multiple imputation approach for missing categorical data. Journal of Statistical Planning & Inference, 2010, 140(11): 3252-3262

[11] Bernaards C A, Sijtsma K. Influence of imputation and EM methods on factor analysis when item nonresponse in questionnaire data is nonignorable. Multivariate Behavioral Research, 2000, 35(3): 321-364

[12] Tang M L, Wang Ng K, Tian G L, Tan M. On improved EM algorithm and confidence interval construction for incomplete tables. Computational Statistics & Data Analysis, 2007, 51(6): 2919-2933

[13] Kryszkiewicz M. Rough set approach to incomplete information system. Information Sciences, 1998, 112(1-4): 39-49

[14] Rady E A, Abd El-Monsef M M E, Abd El-Latif W A. A modified rough set approach to incomplete information systems. Journal of Applied Mathematics and Decision Sciences, 2007: 1155-1168

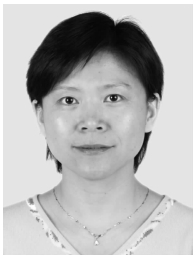
[15] Wang G, Guan L, Hu F. Rough set extensions in incomplete information systems. Frontiers of Electrical and Electronic Engineering in China, 2008, 3(4): 399-405

[16] Twala B E T H, Jones M C, Hand D J. Good methods for coping with missing data in decision trees. Pattern Recognition Letters, 2008, 29(7): 950-956

[17] Wang Guo-Yin. Rough Set Theory and Knowledge Acquisition. Xi'an: Xi'an Jiaotong University Press, 2001(in Chinese)
(王国胤. Rough 集理论与知识获取. 西安: 西安交通大学出版社, 2001)

[18] Zhang Wei, Liao Xiao-Feng, Wu Zhong-Fu. An incomplete data analysis approach based on Rough set theory. Pattern Recognition and Artificial Intelligence, 2003, 16(2): 158-163(in Chinese)
(张伟, 廖晓峰, 吴中福. 一种基于 Rough 集理论的不完备数据分析方法. 模式识别与人工智能, 2003, 16(2): 158-163)

[19] Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 1998, 2(3): 283-304



WU Sen, born in 1971, professor, Ph. D. supervisor. Her current major research interests include intelligent data analysis, knowledge discovery and data mining.

FENG Xiao-Dong, born in 1988, doctoral candidate. His current research interests include data mining, clustering analysis.

SHAN Zhi-Guang, born in 1974, Ph. D., research fellow. His current major research interest is information system evaluation.

Background

This paper focuses on missing data imputation with the aim to recover the missing values of categorical attributes as close to the original ones as possible. Handling missing data is a very important task of the data pre-processing process in the data mining field to ensure the good data quality. A number of researches have been made on dealing with missing values. The simple and broadly used methods include complete data analysis or available case analysis, global constant and attribute mean imputation. These methods normally lead to biased estimate results. The more sophisticated methods, typically multiple imputation and expectation maximization, give better results. But they are mostly based on some statistical hypothesis such as probability distribution. These methods might not be perfect for data mining applications generally, because it is commonly hard to determine probability distribution when the data set is very large. Different from these methods, another existing incomplete data analysis approach ROUSTIDA does not use traditional probability statistical techniques, handling missing data based on rough set theory. It is applicable for data mining applications. However the efficiency and effectiveness of the proposed methods are still far from satisfaction. Getting hints form ROUSTIDA not using traditional probability statistical approaches, this paper proposes MIBOI approach for missing data imputation based on incomplete data clustering. It significantly improves the efficiency of the existing missing data processing methods.

The research was supported by the National Natural Science Foundation of China under Grant No. 70771007. The project is mainly about fundamental and technical problems of clustering knowledge discovery form huge data especially when data is with high dimensionality. The research group mainly works on dissimilarity measures for cluster analysis, missing data processing, data dimensionality reduction and high dimensional data clustering algorithms.