基于聚类关联规则的缺失数据处理研究*

方匡南　谢邦昌

内容提要:本文提出了基于聚类和关联规则的缺失数据处理新方法,通过聚类方法将含有缺失数据的数据集相近的记录归到一类,然后利用改进后的关联规则方法对各子数据集挖掘变量间的关联性,并利用这种关联性来填补缺失数据。通过实例分析,发现该方法对缺失数据处理,尤其是对在先验辅助信息缺失情况下的海量数据集具有较好的效果。

关键词:聚类;关联规则;缺失数据;插补
中图分类号:C816　　文献标识码:A　　文章编号:1002 - 4565(2011)02 - 0087 - 06

Research on Dealing with Missing Data Based on Clustering and Association Rule

Fang Kuangnan & Xie Bangchang

Abstract: This paper proposed a new method of dealing with missing data based on clustering and association rule. Firstly, we divided the original data set into several parts by clustering method, and then use the improved association rule to investigate useful rules between the variables on those child data sets, and use these rules to fill the missing data. We found that this method has a good result on handling massive data sets with missing data by empirical study.

Key words: Clustering; Association Rule; Missing Data

一、引言

数据缺失现象在社会经济研究、抽样调查、生物医药研究等诸多领域普遍存在。数据的缺失不仅影响了数据的质量,也可能造成统计分析结果的严重偏差。因此,对缺失数据的合理处理是一个非常重要的问题,是数据预处理的重要环节,也是提高数据质量的重要方法之一。缺失数据的处理一直是国际统计学界热点讨论的课题之一,但从目前的研究情况来看,我国学者对缺失数据处理的研究比较少,也不够广泛和深入,尤其是对缺失数据的处理方法的研究很少(金勇进、邵军,2009)。

对于缺失数据的处理方法有传统的个案删除法,也就是说如果一记录某个变量值缺失,就把该记录删除,只保留完全的记录。这种方法以减少样本量来换取信息的完备,丢弃了大量隐藏在含有缺失值对象中的信息,尤其当样本量较小时,该方法可能严重影响到数据的客观性和结果的正确性。目前,插补(imputation,也称填补)是处理缺失数据时普遍使用的一种技术,即采用一定的方式,为缺失数据确定一个合理的替代值,插补到原缺失数据的位置上。插补方法一方面可以减小由于缺失数据可能造成的估计量偏差,另一方面力图构造一个完整的数据集(完备信息系统),便于后续的统计分析。

插补方法众多,一般可以分为单一插补(single imputation)和多重插补(multiple imputation)。其中,单一插补是指对缺失数据构造单一替代值插补,常见的方法有众数插补法、均值插补法、回归插补法、热卡(hot deck)插补、冷卡(cold deck)插补法等。多重插补法是指用多个值来填充,然后用针对完整数据集的方法对它们进行分析得出综合的结果,常用的有趋势得分法、预测均值匹配法等。

单一均值、众数插补法的效果较差,而且插补值过于凝聚,扭曲了变量分布,低估了变量的方差;而回归插补需要找到有效的辅助变量,否则插补效果

* 本文为国家社科基金重点项目"国家统计数据质量管理问题研究"(09AZD0345)阶段性成果。

往往较差(金勇进、邵军,2009)。因此,学者们尝试从数据本身寻找变量之间的某种关系,氢变量间的这种关系作为新的辅助信息来进行插补,这样在先验知识缺失或者有限情况下,也可以充分挖掘变量间的辅助信息。Quinlan(2003)使用决策树模型来插补缺失值,但此方法缺点是只适用于单一属性含有缺失值时,如果遇到多个属性都有缺失值时,则无法建立正确决策树和计算属性之间的程度。

此外,也有部分学者尝试使用关联规则插补(shen 等,2007;Hong and Wu,2009;Wu and Chou,2004;Bashir 等,2006),但用传统的关联规则来做插补时,一方面可能规则数不够多,部分缺失值无法插补,另一方面可能会出现规则发生冲突的情况。即同一缺失值由不同的规则得到的插补值不同,这些缺陷致使插补效果较差,限制了关联规则在缺失数据处理中的研究。针对以上问题,本文提出了基于聚类和关联规则的缺失值处理方法。为了解决所挖掘规则数不够的问题,本文提出先对不完整资料根据数据特征聚成不同的类别,这样每一类中的数据都具有较高的相似度,所计算出来的支持度会较高。为了提高插补精度,传统方法往往利用某一个或几个辅助信息进行分层,而在缺乏有效辅助信息条件下,利用聚类方法,从数据本身特征出发,充分利用所有变量间的某种内在联系,将所研究的对象划分为不同的类,从而可以增加类内对象的同质性,扩大类间对象的异质性。为了更好地利用关联规则填补缺失数据,本文提出了一种新的最小支持度计算法。为了解不同规则在插补时的冲突问题,本文提出了关联规则得分方法。

本文的结构安排如下:第二部分含有缺失值数据表的描述;第三部分提出了基于聚类关联规则方法的插补法;第四部分是实例分析;第五部分是本文的小结与讨论。

二、含缺失值数据表的描述

通常数据被储存在数据库中,由一个数据表或者多个数据表组成,列表示属性(变量),行表示事务(对象),且每一行表示该对象的一条信息。对于含有缺失数据的数据表,传统的统计描述方法往往不够精确、简洁,为了便于阐述,本文引入粗糙集理论来描述数据表,用一个四元有序组来表示:

$$S = \{U, Q, V, f\}$$

S 是我们所研究的数据表(粗糙集理论也称之为信息系统),U 是对象(事务)的一个有限集$\{x_1, x_2, \cdots, x_n\}$,$n$ 为所研究的的样本记录(事务)总量,称 U 为论域;$Q = C \vee D$ 是描述对象(事务)属性的一个有限集合$\{q_1, q_2, \cdots, q_l\}$,称 Q 为属性集,其中 C 表示条件属性,D 表示决策(目标)属性集;当然属性集里可以没有目标属性,即 $D = \emptyset$;$V = \cup_{q \in Q} V_q$ 是属性项 $q \in Q$ 的值域;f 是 $U \times Q$ 到 V 的一个映射,即:

$$f: U \times Q \to V$$

且 $f(x, q) \in V_q$,它表示对象 x 关于属性 q 的取值,指的是 U 中每一个对象(事务)的属性值。在事务信息系统 S 中,对于 $q \in Q$、$v \in V$,称(q, v) 为该系统的描述。

如果一个信息系统 S 的属性值域 V_q 中至少有一个属性 q 的值为(用“?”表示)缺失数据,则称该信息系统是不完备的,也就是说含有缺失数据。

如表 1 就是一个事务数据库中的一张数据表,可以用信息系统 S 来描述,论域 U 包含 20 个事务项,即 $U = \{x_1, x_2, \cdots, x_{20}\}$,每个事物由包含 6 个属性项的集合 $Q = \{q_1, q_2, \cdots, q_6\} = \{A, B, C, D, E, F\}$,假设属性 F 是决策属性,其余为条件属性。属性 A 的值域是$\{0, 1, 2, 3, ?\}$,其中?表示缺失值。该数据表的一个事务项在某个属性上的取值看做是 $U \times Q \to V$ 的映射,比如 $f(x_1, A) = 2$,$f(x_{11}, C) = 2$。

表 1 带缺失数据的数据表

ID	A	B	C	D	E	F	ID	A	B	C	D	E	F
1	2	3	1	?	0	?	11	1	1	2	1	0	2
2	0	2	2	2	0	3	12	2	6	2	3	0	3
3	2	4	2	3	0	1	13	3	2	4	?	0	3
4	0	3	1	1	0	1	14	3	4	1	1	1	3
5	2	1	2	2	1	1	15	3	4	4	2	0	1
6	0	6	2	4	0	0	16	3	2	2	?	0	3
7	2	6	2	3	0	1	17	0	?	2	4	0	1
8	3	4	1	1	0	2	18	0	4	2	2	0	3
9	2	5	4	4	0	3	19	2	4	1	1	0	3
10	?	5	3	4	?	1	20	3	6	2	2	0	2

注:? 表示缺失。

三、缺失数据插补

本文提出的基于聚类关联规则的插补法,该方法主要思想是:先将含有缺失数据的数据集 S,利用合适的聚类方法聚成多个类别,这样每一类中的数据都具有较高的相似度,所计算出来的支持度会较高,所以挖掘出来的关联规则也会有较高的代表性,可以避免挖掘出偏颇的关联规则;又由于关联规则可以描述属性之间的关联性,然后在每一群中利用挖掘所得的关联规则对缺失值进行插补。为了更好地利用关联规则填补缺失数据,本文提出了一种新的最小支持度计算法,将缺失数据可能出现的机率一同考虑计算,以增加计算支持度的可信度。此外,在选取用米填补的规则时,提出了关联规则得分计算方法改善以往使用关联规则填补缺失数据所会遇到多个规则填补值冲突的问题。填补完全后再将数据集做合并,然后检验缺失数据填补的正确性以及对填补后的数据集做进一步的统计分析。从某种角度讲,本文提出的方法可以归类到热卡插补法,因为都是使用当期数据集的数据进行插补。

关联规则最早由 Agrawal 等(1993)提出来,主要用来研究事务数据库中属性之间的的关系。假设有一个数据表 $S=\{U,Q,V,f\}$,U 中的每一个事物 x 所包含的属性记为 T,即 $T\subset Q$。假设有一个属性集 A,一个事务的属性 T,即 $A\subset T$,则称事务 x 支持属性 A。关联规则是如下形式的一种含义:$A\rightarrow B$,其中 A,B 是两个属性集,$A\subset Q$,$B\subset Q$,且 $A\cap B=\Phi$。在这里我们称 A 为“前件”,B 为“后件”。一般用支持度(support)、可信度(confidence)等参数来描述关联规则的性质。

关联规则的传统参数设定方法对于含有缺失值的数据集的规则提取往往效率低下,而且用来插补缺失值时可能面临规则冲突等问题,因此,本文提出了缺失概率支持度、最小支持度和规则填补得分方法,并给出了定义和计算方法。

(一)关联规则插补法若干概念和方法

1. 缺失概率支持度(support)。

设 U 中有 $s\%$ 的事务同时支持项集 A 和 B,称 s 为关联规则 $A\rightarrow B$ 的支持度,记为 $s(A\rightarrow B)$。实际上,支持度也就是“前件”和“后件”并集中观测的比例,即 $P(A\cap B)$。不同于传统的的支持度计算,本文将缺失数据在属性中可能出现各个值的机率也考虑进来计算,实际是经缺失概率权重调整后的支持度,因此称为缺失概率支持度。

定义 1　缺失概率支持度:假设 $S=\{U,Q,V,f\}$,其中 $Q=\{q_1,q_2,\cdots,q_l\}$,$U=\{x_1,x_2,\cdots,x_n\}$,$n(v_j^i)=n_j^i$ 表示属性 q_i 第 j 个取值的对象个数,$n_?^i$ 表示属性 q_i 缺失的对象个数,则属性 q_i 的第 j 个取值的缺失概率支持度为

$$\sup_j^{q_i}=n_j^i\frac{n_j^i}{n^i-n_?^i}\Big/\sum_{j=1}^{T_i}n_j^i\frac{n_j^i}{n^i-n_?^i}$$

其中 Ti 表示属性 q_i 的非缺失的属性取值数。

例如,表 1 为一含有缺失数据的数据库,共有 20 笔数据,7 个遗失值,6 个属性为 A、B、C、D、E、F。属性 A 中,缺失数据为 1 笔,取值为 0、1、2、3 的分别有 5、1、7、6 笔。则该带缺失数据的数据表中属性 A 取值为 0、1、2、3 的概率分别为 $\frac{5}{19},\frac{1}{19},\frac{7}{19},\frac{6}{19}$,将缺失数据的机率支持数加入之后,属性 A 取值为 0、1、2、3 的概率则分别为 $5\frac{5}{19}/111$,$1\frac{1}{19}/111$,$7\frac{7}{19}/111$,$6\frac{6}{19}/111$。

2. 最小支持度(min-sup)。

在关联规则的分析中,最小支持度的设定将决定所挖掘出来的规则是否符合要求和规则数目的多少,因此设定最小支持度是非常重要的。本文针对不同的数据集,计算每个属性值在其属性中所出现的比例,并选取一个能够囊括用户所希望的数据百分比,综合所有属性之后,选择最低的百分比作为最小支持度。

定义 2　最小支持度(min-sup)。假设 $S=\{U,Q,V,f\}$,其中 $Q:\{q_1,q_2,\cdots,q_l\}$,$U=\{x_1,x_2,\cdots,x_n\}$,$V^i=\{v_1^i,\cdots,v_{T_i}^i\}$ 表 q_i 的属性值域,T_i 表示 q_i 属性的取值个数。$n_{(j)}^i$ 表示属性 q_i 取值的对象个数的第 j 个次序统计最,且 $\sum_{j=1}^{T_i}n_{(j)}^i=n$,则最小支持度为

$$\sup_{\min}=\min\left\{\frac{n_{(m_1)}^1}{n^1},\cdots,\frac{n_{(m_l)}^l}{n^l}\right\}$$

其中,$\sum_{j=1}^{m_i}\frac{n_{(j)}^i}{n^i}\geqslant\lambda\geqslant\sum_{j=1}^{m_i-1}\frac{n_{(j)}^i}{n^i}$,$\lambda$ 是根据实际需要没定的阈值。

例如,假设希望表 1 每个属性能够找到至少 70% 的数据。以 A 属性为例,取值为 0,1,2,3 的分别有 3,1,6,4 笔,缺失数据的笔数为 1,经由排序之

后,若要涵括 70% 资料则表示门坎值至少设在 20%。依此类推,可得所有属性的最小支持度,属性 B、C、D、E 的最小支持度分别为 10%、25%、20%、85%、35%,这六个属性中最小支持度最小的是属性 B,为 10%,将此作为该数据集的最小支持度。

3. 规则填补得分(socre)。

使用关联规则填补缺失数据时,可能会发生多条规则可以填补到相同缺火数据,而多条规则所要填补的值有可能会不相同,因此在这种情况下,本文提出一个新的计算关联得分方法。首先,计算规则的有效填补分数,并用填补分数高的规则来做填补。当缺失数据被填补过后,能够再填补到其他缺失数据时,填补分数的算法将有所改变。

第一次填补分数计算方式:

$$appl(X_j^i = X_j^i) = \begin{cases} 1, \text{为非缺失值} \\ 0, \text{为缺失值} \end{cases}$$

$$appl(t) = \sum_{j=1}^{Wt} appl(X_j^i = X_j^i)$$

$$score(t) = \frac{appl(t)}{Wt} \times \left[lift(X \rightarrow Y) \times \sup(X \rightarrow Y) \right]^{\frac{Wt}{n}}$$

其中,X_j^i 表示在第 j 笔数据中第 i 个属性的项(item);$appl(X_j^i = X_j^i)$ 表示在数据库中属性值不缺失就给予基本分数 1 分,如果缺失就不给予分数,即 0 分。并加总此条规则中的 $appl(X_j^i = X_j^i)$ 分数,Wt 表示此条规则所使用数据库中的属性个数,因此 $\frac{appl(t)}{Wt}$ 可以代表此规则的明确程度。$lift(X \rightarrow Y)$ 则表示规则 $X \rightarrow Y$ 的相关程度;而 n 为数据库中的所有属性个数,因此,$\frac{Wt}{n}$ 可用来表示此规则所使用的属性在数据库中所有属性的比例。

第 k 次填补分数计算方式:

$$appl(X_j^i = X_j^i) = \begin{cases} 1, \text{为非缺失值} \\ 0, \text{为缺失值} \\ (0.5)^p, \text{缺失值} \rightarrow \text{非缺失值}, p = k - 1 \end{cases}$$

当数据库经由第一次填补缺失数据后,若没有将所有缺失数据填补完全,经由本研究所定之填补流程,进入到第 k 轮填补缺失数据,当有缺失数据经由上一轮填补而成为非缺失数据时,appl 的分数则调整为 $(0.5)^p$,其中 p 为第 $k-1$ 次填补。

(二)基于聚类关联规则的缺失数据插补算法

为了更加凝练表示基于聚类关联规则的缺失数据插补方法,本文提出了如下可执行的算法。

输入:不完备信息系统 $S = \{U, R, V, f\}$,输出:填充后完备信息系统 $S' = \{U', R, V, f'\}$。

步骤 1:计算信息系统 S 的缺失对象集 MOS。

步骤 2:对不完备信息系统 S 进行聚类,把相似的事务项聚成一类,得到一系列子信息系统 $S_1, S_2, \cdots, S_k$。

步骤 3:for $i = 1$ to k do

对聚类后的子信息系统 S_i 挖掘关联规则 $\{A_{ti} \rightarrow B_{ti}\}, t = 1, \cdots, L_i$,$L_i$ 表示第 i 个子信息系统 S_i 挖掘出来的关联规则数。

End do

步骤 4:if 无法用关联规则填补的属性,then 使用众数填补。

Else if 可以用规则填补,且为单一规则,then 直接用规则填补;Else if 若有多条规则,且彼此不冲突,then 直接用规则填补,Else if 若有多条规则且发生冲突情形,then 进行第 k 次($k = k + 1$)填补的得分计算,并选择较高得分的规则来填补。

步骤 5:填补完所有数据之后,检验步骤 4 后的信息系统是否完备,若不完备则重复步骤 3 - 步骤 4。

步骤 6:当无法再有缺失数据可被填补,则执行结束。

本文的方法与分层热卡插补法、序贯热卡插补法等类似,难以给山明确的均方误差估计公式,因此主要通过模拟或者实证研究来分析其插补效果。

四、实例分析

本文对表 1 的缺失数据利用关联规则方法,设定最小支持度为 15%,提取了 46 条关联规则,详见表 2。进行第一轮填补时,从第一笔数据开始,非缺失值的属性项目有 A_2、B_3、C_1 与 E_0,而为缺失值的属性为 D 与 F,因此在规则中,挑选出前项为 A、B、C 与 E 的规则,以及后项为属性 D 与 F 的所有规则。接着,分别将前项所找出来的规则与后项的规则做交集,所得到的规则就是前项为 A、B、C、E 且后项必为属性 D 与 F 的规则。而这些规则最后还须要与这笔数据的项目做交集,如此一来才能够找到前项必为 A_2、B_3、C_1、E_0 且后项必为属性 D 与 F 的规则。若发现只有一条规则可填补时,则选择直

接填补;有多条规则可填补时,则先检查规则是否彼此冲突(同一个缺失值有不同填补的值),没有冲突则直接填补,有冲突时则使用填补分数计算法来计算出较高分的规则,并以高分规则来做填补。经第一轮填补后,若数据集还有缺失值(填补率大于0%)则进行第二轮的填补,并重复至无法填补为止。最后,无法用规则填补的缺失值用众数填补。含有缺失数据的事务项填补后的结果见表 3。

表 2　　由表 1 挖掘所得的关联规则

1	$A_0 \rightarrow C_2$	17	$C_4 \rightarrow E_0$	33	$B_4 D_1 \rightarrow C_1$
2	$A_0 \rightarrow E_0$	18	$D_1 \rightarrow E_0$	34	$B_6 C_2 \rightarrow E_0$
3	$D_4 \rightarrow A_2$	19	$D_2 \rightarrow E_0$	35	$B_6 E_0 \rightarrow C_2$
4	$A_2 \rightarrow E_0$	20	$D_3 \rightarrow E_0$	36	$C_1 D_1 \rightarrow E_0$
5	$A_3 \rightarrow E_0$	21	$D_4 \rightarrow E_0$	37	$C_1 E_0 \rightarrow D_1$
6	$B_2 \rightarrow E_0$	22	$F_1 \rightarrow E_0$	38	$D_1 E_0 \rightarrow C_1$
7	$B_2 \rightarrow F_3$	23	$F_2 \rightarrow E_0$	39	$C_2 D_2 \rightarrow E_0$
8	$B_4 \rightarrow E_0$	24	$F_3 \rightarrow E_0$	40	$D_2 E_0 \rightarrow C_2$
9	$B_6 \rightarrow C_2$	25	$A_0 C_2 \rightarrow E_0$	41	$C_2 D_3 \rightarrow E_0$
10	$B_6 \rightarrow E_0$	26	$A_0 E_0 \rightarrow C_2$	42	$D_3 E_0 \rightarrow C_2$
11	$C_1 \rightarrow D_1$	27	$A_2 D_3 \rightarrow E_0$	43	$C_2 D_3 \rightarrow E_0$
12	$D_1 \rightarrow C_1$	28	$D_3 E_0 \rightarrow A_2$	44	$D_3 E_0 \rightarrow C_2$
13	$C_1 \rightarrow E_0$	29	$A_2 F_3 \rightarrow E_0$	45	$C_2 F_1 \rightarrow E_0$
14	$D_2 \rightarrow C_2$	30	$B_3 E_0 \rightarrow F_3$	46	$C_2 F_3 \rightarrow E_0$
15	$D_3 \rightarrow C_2$	31	$B_2 F_3 \rightarrow E_0$		
16	$C_2 \rightarrow E_0$	32	$B_4 C_1 \rightarrow D_1$		

表 3　　表 1 缺失部分数据插补结果

ID	A	B	C	D	E	F
1	2	3	1	**1**	0	**3**
10	2	5	3	4	0	1
13	3	2	4	**1**	0	3
16	3	2	2	**2**	0	3
17	0	**4**	2	4	0	1

注:表中数字粗体的为填补值。

为了进一步说明基于聚类关联规则的缺失数据插补方法,以及比较与其他插补方的优劣,本文选用了 Suduku、Chess 与 German 等 3 个数据集。首先,将原始数据分为训练集和测试集,抽取 80% 左右的数据作为训练集,剩余的 20% 作为测试集,利用训练集挖掘出来的关联对测试集的缺失数据进行插补。具体数据特征描述见表 4。

表 4　　数据集特征描述

数据集名称	条件属性		目标属性	训练集样本数	测试集样本数	样本总数
	个数	变量类别				
Suduku	9	全是分类变量	1	767	191	958
Chess	36	全是分类变量	1	2557	639	3196
German	20	13 个分类变量,7 个数值变量	1	800	200	1000

注:数据来源于加州大学欧文分校机器学习数据库 http://archive.ics.uci.edu/ml/。

除了本文提出的聚类关联规则法(CAR)和关联规则(AR)法外,还选用了众数插补法、SMV 插补法、FRCAR 插补法等几种比较常用的插补方法作为标杆进行比较。各种方法的描述如下:

表 5　　插补方法

插补方法	说明
CAR	利用聚类方法,先将原信息系统 S 聚成不同的子信息系统,再使用本文提出的关联规则插补法
AR	不对原信息系统聚类,直接在 S 上使用本文所提出的关联规则插补法
Mode	全部缺失数据都采用属性之众数插补
SMV	使用 SMV 算法来插补
FRCAR	使用 shen 等提出的 FRCAR 算法来插补

聚类关联规则(CAR)方法是首先把训练数据集分为不同的类,本文使用的聚类方法是 Kohonen 法,不同群的训练数据集分别挖掘关联规则来做填补,接着使用各群当中的训练数据所挖掘出来的关联规则对测试数据集进行填补,填补完各群之后,再将各群的数据合并,并检验其填补的正确率。

表 4 中的数据集实际上都是完备信息系统,都不存在缺失数据,为了验证各种插补方法的优劣,本文使用随机数种子随机打空数据集,分别产生含有 5%、10%、15% 及 20% 缺失数据的共 12 个数据集。对于属性为数值变量的,需要通过分组离散化。各种方法填补准确率如表 6 所示:

从表 6 可以看出,本文提出的聚类关联规则(CAR)方法对三个数据集在不同的缺失比例下的插补准确率都是最高,其次是关联规则(AR)方法,而比较常用的众数(Mode)插补法效果最差。

表 6　　缺失数据填补正确率　　(%)

数据集	Suduku				German				Chess			
缺失比例(%)	5	10	15	20	5	10	15	20	5	10	15	20
CAR	42.11	48.77	44.62	46.18	43.66	45.26	42.08	44.29	75.80	74.49	74.76	73.12
AR	37.11	43.50	41.09	42.56	41.15	41.75	41.99	41.89	70.66	68.68	70.86	70.76
Mode	29.34	35.84	37.18	37.00	38.09	37.30	35.36	38.71	64.42	61.20	67.77	65.47
SMV	34.34	41.93	40.67	41.32	38.35	36.27	35.07	39.00	68.03	62.28	68.78	66.73
FRCAR	36.05	43.91	41.30	42.63	37.95	37.65	38.14	39.52	70.55	66.49	70.60	70.30

五、小结与讨论

本文提出了基于聚类和关联规则方法的缺失数据插补法。该方法先对原始数据集进行聚类，将具有相同特点的数据聚成一类，然后利用改进的关联规则来填补缺失数据；为了更好地利用关联规则填补缺失数据，本文提出了一种新的最小支持度的设定方法，将缺失数据在数据库中可能出现的机率一同考虑计算，以增加计算支持度的可信度；在选取用来填补的规则时，提出了规则得分计算方法，解决了多规则插仆冲突的问题。通过本文的实例数据分析可以得知，本文提出的基于聚类关联规则的插补方法具有较好的效果，优于其他几种插补方法。与均值、众数插补法相比，本文的方法得到的插补值更加分散，没有像均值插补法那样插补值过于凝聚，扭曲了变量的样本分布和低估了插补方差。本文提山的方法可以在缺少先验辅助信息条件下，根据数据本身的特征，充分挖掘数据内部变最和数值间的联系，利用这种数据内部间的联系米进行插补。

此外，本文提山的方法不仅是针对分类变量的插补，对于数值变最也是同样适用，需在填补缺失数据之前，将数值属性离散化成类别属性，再进行关联规则插补，用插补值加上离散区间长度的均匀随机数。比如将数值属性按区间长度 c 划分为不同的区间，利用本文提出的聚类关联规则插补后的值加上 $[0,c]$ 的均匀分布随机数 e_i，如果区间长度 c 选取合适，往往具有良好的插补效果。

本文的不足之处在丁当某一记录里缺失值较多时，可能会缺少关联“前件”而无法插补，因此，有时需要结合均值、随机均值或众数等其他插补法。如何将本方法和已有插补方法有效结合起来，将本方法作为一种挖掘变量间辅助信息的手段，充分利用各种方法的优势，提高插补的效果，是今后要努力的方向之一。

参考文献

[1] 金勇进. 缺失数据的插补调整 [J]. 数理统计与管理，2001(5)：47 – 53.

[2] 金勇进、邵军. 缺失数据的统计处理 [M]. 北京：中国统计出版社. 2009(1).

[3] 张其文、李明. 一种缺失数据的填补方法 [J]. 兰州理工大学学报. 2006(4)：102 – 104.

[4] 金勇进. 调查中的数据缺失及处理(I)—缺失数据及影响 [J]. 数理统计与管理. 2001(01)：59 – 62.

[5] Baraldi A. N. Enders C. K. An introduction to modern missing data analyses [J]. Journal of School Psychology. 2010(48)：5 – 37

[6] Angiulli F. Ianni G. Palopoli L. On the complexity of inducing categorical and quantitative association rules [J]. Theoretical Computer Science. 2004(314)：217 – 249.

[7] Huang, C. C., A Case – Based Reasoning Model for Supporting Feature Weight and Missing Value Completion [J], Industrial and Information Management, NCKU. 2005.

[8] Gustavo E. A. P. A. Batista and Maria Carolina Monard, An Analysis of Four Missing Data Treatment Methods for Supervised Learning [J], Applied Artificial Intelligence, 2003(17): 519 – 533.

[9] Liu, W. Z, White, A. P., Thompson, S. G. and Bramer, M. A. Techniques for Dealing with Missing Values in Classification [J], International Symposium on intelligent Data Analysis, 1997：527 – 536.

[10] Liang, T. H., Wang, C. Y., and Yang, Y. H. A study of Imputation Missing Data for Household Income [J], Journal of Data Analysis, 2006(4)：75 – 101.

[11] Agrawal, R. and Srikant, R., Fast Algorithm for Mining Association Rules [C], Proc. 20th Int1 Conf. Very Large Data Bases, Santiago, Chile, 1994. 487 – 499.

[12] Shen, J. J. Chang C. C. and Li Y. C., Combined association rules for dealing with missing values [J], Journal of Information Science, 2007(33): 468 – 480.

[13] Hong, T. P., and Wu, C. W., Data Mining from an Incomplete Data Set [C], The 14th Conference on Artificial Intelligence and Application, 2009.

[14] Wu, C. H., Wun, C. H., Chou, H. J., Using Association Rules for Completing Missing Data [C], Proceeding of the Fourth International Conference on Hybrid Intelligent System, 2004.

[15] Shariq B., Saad R., Umer M., Sonya T., A. Rauf B. Using Association Rules for Better Treatment of Missing Value [C]. 10th WSEAS Conference on Communication & Compute, 2006.

作者简介

方匡南，男，27 岁，浙江省人，2010 年毕业于厦门大学经济学院计划统计系，获经济学博士学位，现为厦门大学经济学院助理教授。研究方向为数据挖掘和经济计量。

谢邦昌，男，48 岁，中国台湾省人，1991 年毕业于台湾大学统计系，获统计学博士，现为台湾辅仁大学统计资讯系教授，厦门大学讲座教授。研究方向为数据挖掘与商业智能。

（责任编辑：周 晶）