重复数据中关键属性值缺失填补的改进ROUSTIDA算法

樊哲宁 杨秋辉 翟宇鹏 万 莹 王 帅
(四川大学计算机学院(软件学院) 成都 610065)

摘 要 随着数据分析研究的兴起,数据预处理越来越得到研究者的重视,其中缺失数据填补问题的重要性也逐渐显现。在ROUSTIDA数据补齐算法的基础上,针对具有关键属性的重复数据的特点,文中提出了一种改进的ROUSTIDA算法——Key&Rpt_RS算法。Key&Rpt_RS算法继承了ROUSTIDA算法的优势,同时考虑了目标数据的重复性特点,分析了关键属性对填补效果的影响,得到了更加准确且有效的填补结果。

关键词 数据预处理,重复数据,缺失填补,ROUSTIDA算法
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.02.005

Improved ROUSTIDA Algorithm for Missing Data Imputation with Key Attribute in Repetitive Data

FAN Zhe-ning YANG Qiu-hui ZHAI Yu-peng WAN Ying WANG Shuai
(College of Computer Science(Software Engineering),Sichuan University,Chengdu 610065,China)

Abstract With the rise of data analysis,the importance of data pre-processing has attracted more and more attention,especially the imputation of missing data. Based on the ROUSTIDA algorithm,this paper proposed an improved ROUSTIDA algorithm—Key&Rpt_RS algorithm. Key&Rpt_RS algorithm inherits the advantages of ROUSTIDA algorithm,considers the characteristic of repeatability in objective data,and analyzes the influence of key attribute on imputation effect. At last,this paper conducted the experiments based on the alarm data in communication network. The results show that Key&Rpt_RS algorithm outperforms the traditional ROUSTIDA algorithm in terms of the imputation effect for missing data.

Keywords Data pre-processing,Repeated data,Missing data imputation,ROUSTIDA algorithm

1 引言

数据分析是当前的热点研究领域之一,缺失数据处理的重要性逐渐凸显。数据缺失是数据分析领域的研究难题之一,普遍存在于社会学研究、经济分析、生物医药、工程应用等各个方面,影响数据的质量,进而造成统计分析结果存在偏差,严重者甚至会造成错误的结论。对缺失数据进行合理的完善,提高数据质量,是数据预处理的重要环节,也是数据分析过程的关键。因此,对缺失数据的研究具有重要的意义。

现实世界中存在的数据多种多样,并且数据所描述的对象、数据产生机制、缺失原因等不同,使得现有的缺失补齐方法不能满足所有数据的需要。本文立足于具有关键属性的重复数据,即大量除关键属性外其余属性值完全相同的重复数据,通过对其结构特点进行分析可以得出,关键属性将严重影响传统的数据补齐算法对数据之间相似性的计算。误差的形成原因主要有以下两点:

(1)当其中某条数据出现缺失时,重复产生的其他几条无缺失数据应当被作为与其最相似的数据来协助进行缺失补齐,然而这些数据可能会因为在关键属性上的差异被错误地认为相似度不高,从而降低补齐效果;

(2)当某条数据出现关键属性上的缺失时,应该根据其他重复数据的关键属性分布来进行补齐,但如果采用传统的数据补齐算法,则会出现因为多个相似数据在缺失属性上的值不同而导致无法补齐的情况。

基于此类数据在结构上的独特性,对其进行无差异处理可能会影响缺失数据补齐的效果。因此,本文针对此类包含关键属性且具有重复性的数据提出了缺失补齐的改进算法——Key&Rpt_RS算法,该算法改进了基于粗糙集理论的传统ROUSTIDA算法,且实验证明其得到了更优的补齐效果。

2 相关工作

自20世纪70年代后期,国外对数据缺失问题的研究日渐增多,以Rubin为代表的学者们陆续提出了各种对缺失数据进行处理的方法。已有的缺失数据处理方法大致可以分为4类:直接删除法、基于插补的方法、基于参数似然的方法和

到稿日期:2017-12-05 返修日期:2018-03-22

樊哲宁(1994—),女,硕士,主要研究方向为软件分析与测试,E-mail:fanzheningchn@163.com;**杨秋辉**(1970—),女,副教授,CCF会员,主要研究方向为软件测试、经验软件工程、数据库系统及其应用,E-mail:yangqiuhui@scu.edu.cn(通信作者);**翟宇鹏**(1992—),男,硕士,主要研究方向为软件自动化测试;**万 莹**(1993—),女,硕士,主要研究方向为软件分析与测试;**王 帅**(1992—),男,硕士,主要研究方向为数据挖掘。

基于加权调整的方法[1-3]。

直接删除法：指在数据量极大而缺失数据量相对很小的情况下，可以将包含缺失值的记录直接丢弃，使其不参与后续分析的方法。该方法操作方便，但造成分析结果偏差的风险较大。

基于插补的方法：插补是指对缺失的数据进行填补，使之成为完备数据以用于后续的分析。金勇进[4]对缺失数据的插补调整做了详细的介绍，所涉及的插补方法有演绎估计、均值插补、随机插补、回归插补和多重插补等。

基于参数似然的方法：参数似然是在数据总体分布类型已知，假定模型正确并且缺失机制为随机缺失的情况下，通过已观测数据的分布对未知参数进行似然估计的方法。这一类的代表性算法有 Dempster[5] 提出的计算极大似然估计的 EM 算法等。

基于加权调整的方法：加权调整方法的原理是，当出现数据缺失时，采用某种方法将缺失记录所占权重分解到非缺失记录上，增大有观测值的记录的权重，以减少数据缺失可能对结果造成的误差。金勇进[6]介绍了几种常见的加权调整法，包括 Politz-Simmons 调整法、事后分层调整法和加权组调整法；另外，Robins 等[7]提出了加权估计方程(WEE)等方法。

除以上 4 种处理缺失数据的方法之外，近年来，随着数据挖掘与分析技术的兴起，基于粗糙集的数据补齐方法逐渐受到学者们的重视与关注。其中，ROUSTIDA 算法具有最为广泛的应用场景，越来越多的专家学者将其应用到缺失数据补齐领域，并针对不同的数据类型做了更为有效的改进。张振华等[8]在原 ROUSTIDA 算法的基础上引入决策规则独立原则，增强了算法的实用性；段鹏等[9]针对有缺失属性的对象与任何对象都不相似及与多个对象都相似时 ROUSTIDA 算法无法进行补齐的情况，扩充了原算法的适用范围；田树新等[10]将条件属性与决策属性区分对待，扩展了原 ROUSTIDA 算法的适用范围，并通过实例说明改进后的算法能获得更集中的决策规则；丁春荣等[11]依据决策规则独立原则，提出了一种基于相似关系向量的不完备信息系统数据补齐算法，该算法有效地解决了原 ROUSTIDA 算法可能存在的决策规则矛盾等问题。

尽管数据补齐预处理的方案多种多样，近年来 ROUSTIDA 算法也得到了广泛的应用，但由于现实生活中待处理数据对象的复杂性和多样性，使得难以期望某一算法对所有数据补齐都有较好的效果。本文亦基于以粗糙集理论(Rough Set Theory)为基础的 ROUSTIDA 算法，针对具有关键属性及重复性特点的数据，对算法进行了改进。

3　ROUSTIDA 算法的改进

3.1　ROUSTIDA 算法的相关概念

ROUSTIDA 算法是基于粗糙集理论的不完备数据分析方法，充分利用了粗糙集理论的优点，能够直接对原不完备数据集进行分析，而无需依赖于任何附加信息。粗糙集理论是由波兰数学家 Pawlak[12] 于 1982 年提出的一种对存在不精确、不确定性、模糊不清等问题的信息进行处理的数学工具，它的主要思想是通过已有的精确知识来对不精确的知识进行近似刻画，并在保持空间原有分类能力不变的前提下对空间对象的属性进行约简，从而发现蕴含的知识。该理论应用到对缺失数据进行补齐的问题上时，即表现为使得存在缺失值的对象与其他相似对象的差异性尽可能最小，从而减小缺失数据对后续分析结果的影响。

在介绍使用 ROUSTIDA 算法进行数据补齐之前，首先介绍粗糙集理论中的相关概念。

(1)信息系统

信息系统常被称为信息表、属性值或者数据表，是将数据对象转化为数学方法进行的抽象描述。信息系统[13]是一个四元组 $S=\{U,A,V,f\}$，表示一组对象(或事例)的非空有限集合，其中，U 是对象的非空有限集合，也称为论域；A 是属性的非空有限集合；V 表示属性的值域，即所有属性的任何可能取值的集合 $f:U\times A\to V$ 为信息函数，即给定对象的属性到其对应属性值的映射。

(2)可辨识矩阵

可辨识矩阵也称为差异矩阵，是 Skowron 教授在 20 世纪 90 年代初基于决策表系统提出的[14]。上文中，信息系统的定义实际上是对决策表系统的扩充，在决策表系统中，属性集合 A 由条件属性 P 和决策属性 D 组成，而信息系统不对其做区分。因此，这里将基于信息系统的可辨识矩阵称为扩充可辨识矩阵，矩阵单元的内容是属性集，表示两组对象(或事例)在该属性上的值不同，反映了对象间的属性差异。

扩充可辨识矩阵的定义为：

$$M(i,j)=\begin{cases}a_k\mid a_k\in A\wedge a_k(x_i)\neq a_k(x_j)\wedge a_k(x_i)\neq *\wedge \\ \quad a_k(x_j)\neq *, & i\neq j\\ \varnothing, & i=j\end{cases} \tag{1}$$

其中，$i=1,2,\cdots,m$；$j=1,2,\cdots,n$；“$*$”表示缺失。

除可辨识矩阵外，ROUSTIDA 算法还涉及到以下几种变量。

数据对象 x_i 的遗失属性集 MAS_i：

$$MAS_i=\{a_k\mid a_k(x_i)=*,k=1,\cdots,m\} \tag{2}$$

无差别对象集 NS_i(也称为相似对象集)：

$$NS_i=\{j\mid M(i,j)=\varnothing,i\neq j,j=1,\cdots,n\} \tag{3}$$

信息系统 S 的遗失对象集 MOS：

$$MOS=\{i\mid MAS_i\neq\varnothing,i=1,\cdots,n\} \tag{4}$$

3.2　ROUSTIDA 算法的流程

ROUSTIDA 算法的一般流程[15]如图 1 所示。

输入：不完备信息系统 $S^0=\langle U^0,A,V,f^0\rangle$；

输出：完备的信息系统 $S^r=\langle U^r,A,V,f^r\rangle$。

生成中间信息系统 S^{r+1} 的规则如下：

(1)对于 $i\notin MOS^r$，有 $a_k(x_i^{r+1})=a_k(x_i^r)$，$k=1,2,\cdots,m$。

(2)对于 $i\in MOS^r$，对所有 $k\in MAS_i^r$ 做循环。

1)如果 $|NS_i^r|=1$，设 $j\in NS_i^r$：

若 $a_k(x_j^r)=*$，则 $a_k(x_i^{r+1})=*$；否则，$a_k(x_i^{r+1})=a_k(x_j^r)$。

2)否则：

若存在 $j_0,j_1(j_0,j_1\in NS_i^r)$，满足 $(a_k(x_{j_0}^r)\neq *)\wedge(a_k(x_{j_1}^r)\neq *)\wedge(a_k(x_{j_1}^r)\neq a_k(x_{j_0}^r))$，则 $a_k(x_i^{r+1})=*$；

否则，若存在 $j_0\in NS_i^r$，满足 $(a_k(x_{j_0}^r)\neq *)$，则 $a_k(x_i^{r+1})=a_k(x_{j_0}^r)$；

否则，$a_k(x_i^{r+1})=*$。

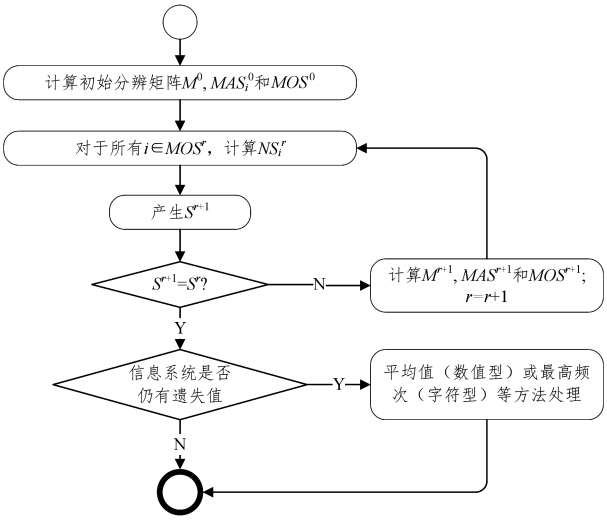

图 1 ROUSTIDA 算法的流程图

Fig. 1 Flowchart of ROUSTIDA algorithm

ROUSTIDA 算法通过对扩充差异矩阵的多次计算和完整化分析，来对不完备信息系统进行逐步填补，直至终止条件成立，实现了完备的信息系统。该算法实现了对缺失值的填补，使得完整化后的信息系统中具有缺失值的对象与其他相似对象的属性值之间的差异尽可能保持最小[16]。

3.3 考虑关键属性影响的 Key&Rpt_RS 算法

结合关键属性考虑的 Key&Rpt_RS 算法的基本思想包含以下两个方面。

(1)当数据中存在关键属性上的缺失时，在关键属性不参与计算的前提下，找出与其最相似的数据对象，并选择相似对象关键属性的随机值进行补齐；当数据在其他属性上存在缺失时，也应当忽略关键属性，从而更加准确地计算相似性，进而提高补齐的准确率。

(2)当缺失填补完成后，原 ROUSTIDA 算法对未能有效填补的缺失数据进行了平均值或最高频次值的填补。而从原算法生成 S^{r+1} 的规则(2)中的 2)可以发现，当缺失对象 i 的无差别对象集中存在两条以上的数据对象，且在缺失属性上的值不空且不等时，缺失数据将不被填补，如表 1 所列。

表 1 ROUSTIDA 算法无法填充的情况

Table 1 Example of ROUSTIDA algorithm without imputation

U	a_1	a_2	a_k	a_n
i	1	0	*	0
j_0	1	0	1	*
j_1	1	0	0	0

针对原算法的上述缺陷，本文提出改进思路：若无差别对象集 NS_i 中存在多于两种不同的相似对象，则出现次数越少的对象 k 的属性值被遗失的概率就越大[17]。因此，在原算法的基础上，在对无差别对象集 NS_i 做循环的同时统计了不同的属性 k 值出现的次数，并选取出现次数最少的属性 k 值进行缺失填补。这一改进有效地解决了原算法对部分缺失无法填充的问题，且证明了该方法的命中率远高于平均值或最高频次填补。

综合上述改进思路，图 2 给出了 Key&Rpt_RS 算法的流程。

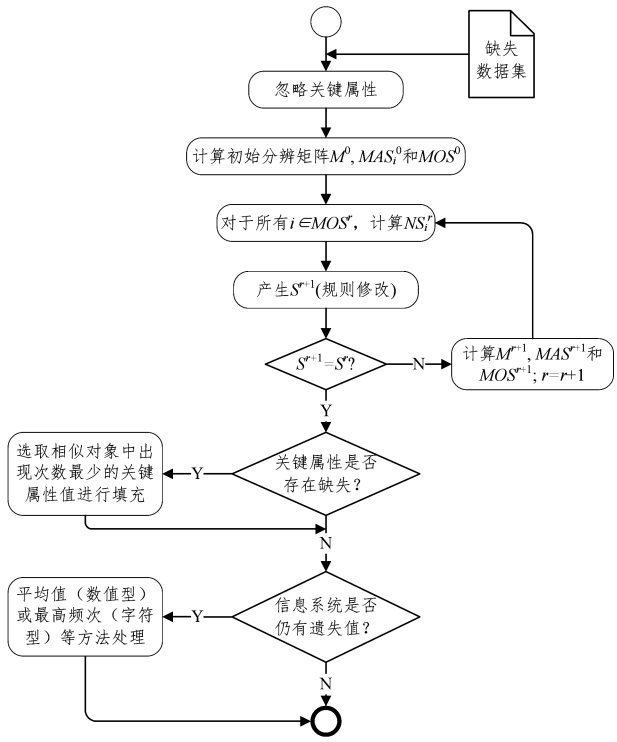

图 2 Key&Rpt_RS 算法的流程图

Fig. 2 Flowchart of Key&Rpt_RS algorithm

其中，原 S^{r+1} 生成规则(2)中的 2)被改进为：

对 $j\in NS_i^r$ 做循环，统计不同的属性 k 值出现的次数，并记录出现次数最少的无差别对象 j_m。

若存在 j_0，$j_1\in NS_i^r$，满足$(a_k(x_{j_0}^r)\neq *)\wedge(a_k(x_{j_1}^r\neq *)\wedge(a_k(x_{j_1}^r)\neq a_k(x_{j_0}^r))$，则 $a_k(x_i^{r+1})=a_k(x_{j_m}^r)$；

否则，若存在 $j_0\in NS_i^r$，满足$(a_k(x_{j_0}^r)\neq *)$，则 $a_k(x_i^{r+1})=a_k(x_{j_0}^r)$；

否则，$a_k(x_i^{r+1})=*$。

该改进思想是针对具有关键属性的重复数据缺失填补的研究提出的，经下文实验验证可以得到，Key&Rpt_RS 算法对该类数据具有较优的填补效果。但由于 Key&Rpt_RS 算法具有较强的针对性，即对数据本身结构特征的要求较高，但对其他结构的数据缺失填补效果未知，其适应性与通用性或有降低。

4 实验验证

本节以通信网络的告警数据为例，对本文提出的 Key&Rpt_RS 算法和改进前的 ROUSTIDA 算法进行了比较实验。

4.1 实验数据

实验使用的数据为某通信公司提供的告警数据以及告警仿真数据。当通信网络发生故障时会触发相应的告警动作，这些告警数据中都包含着故障发生的时间信息。在现实使用中，某一故障发生时可能导致重复多次产生告警数据，这些数据所描述的对象几乎是一致的，但却存在时间上的间隔，这里的时间即为关键属性。

表 2 列出了 2016 年 12 月 26 日的一小段告警数据(因篇幅原因，此处只列出几条告警数据以及其中的几个属性)。实验从所有数据中抽出 3 组，分别包含 1000，2000，3000 条告警数据，从每组数据中随机抽取 200 个属性值并将其设置为空；

然后分别使用 ROUSTIDA 和 Key&Rpt_RS 方法进行补齐；最后比较各方法补齐的效果。

表 2　告警数据的样例

Table 2　Example of alarm data

id	time	name	level	n_loc	e_loc
14001	2017-04-06 20:26:42	Slap 链路故障	重要	Node31	框号 0 槽号 1
14001	2017-04-06 20:26:44		重要	Node36	框号 0 槽号 2
10304	2017-04-06 20:26:46	Mtp 链路拥塞	重要	Node37	框号 0 槽号 3
1111	2017-04-06 20:26:50	other1	提示	Node2	框号 1 槽号 1
11608		GTPU 路径故障	重要	Node8	框号 2 槽号 2
11609	2017-04-06 20:26:54	GTPC 路径故障	重要	Node24	框号 2 槽号 3

4.2　实验过程

4.2.1　数据预处理

由于数据对象的属性值包含中文字符，因此需要首先对包含中文字符的属性列进行如下替换。

name{Slap 链路故障，MTP2 链路拥塞，GTPC 路径故障，…}数值化为{0，1，2，…}；

level{重要，提示}数值化为{0，1}；

e_loc{框号 p 槽号 q}字符化为{pq}。

4.2.2　ROUSTIDA 填补

由于篇幅限制，这里设计了由 6 条数据组成的示例来说明算法的计算过程，如表 3 所列，其中 a_1 为关键属性。

表 3　信息系统 S^0(ROUSTIDA 算法)

Table 3　Information system S^0(ROUSTIDA algorithm)

编号	a_1	a_2	a_3	a_4	a_5
1	*	0	1	0	0
2	0	0	*	0	1
3	1	0	1	0	0
4	1	0	1	0	1
5	1	1	1	0	1
6	1	0	1	0	*

计算：M^0，MAS_i^0，MOS^0，NS_i^0：$M^0(1,1)=\varnothing$，$M^0(1,2)=\{a_5\}$，$M^0(1,3)=\varnothing$，$M^0(1,4)=\{a_5\}$，$M^0(1,5)=\{a_2,a_5\}$，$M^0(1,6)=\varnothing$，…；$MAS_1^0=\{a_1\}$，$MAS_2^0=\{a_3\}$，$MAS_3^0=\varnothing$，$MAS_4^0=\varnothing$，$MAS_5^0=\varnothing$，$MAS_6^0=\{a_5\}$；$MOS^0=\{1,2,6\}$；$NS_1^0=\{3,6\}$，$NS_2^0=\varnothing$，$NS_3^0=\{1,6\}$，$NS_4^0=\{6\}$，$NS_5^0=\varnothing$，$NS_6^0=\{1,3,4\}$。

由此可以得到 S^1，如表 4 所列。

表 4　信息系统 S^1(ROUSTIDA 算法)

Table 4　Information system S^1(ROUSTIDA algorithm)

编号	a_1	a_2	a_3	a_4	a_5
1	1	0	1	0	0
2	0	0	*	0	1
3	1	0	1	0	0
4	1	0	1	0	1
5	1	1	1	0	1
6	1	0	1	0	*

继续重复上述过程得到 S^2，会发现 S^2 与 S^1 完全相同，即达到循环终止条件。表 4 中依然存在缺失值，这里采用最高频次法进行填补，得到表 5 所列的完备信息系统。

表 5　完备信息系统 S(ROUSTIDA 算法)

Table 5　Complete information system S(ROUSTIDA algorithm)

编号	a_1	a_2	a_3	a_4	a_5
1	1	0	1	0	0
2	0	0	1	0	1
3	1	0	1	0	0
4	1	0	1	0	1
5	1	1	1	0	1
6	1	0	1	0	1

4.2.3　Key&Rpt_RS 填补

根据所提出的 Key&Rpt_RS 算法思想，首先应忽略关键属性，对表 6 所列的不完备信息系统进行 Key&Rpt_RS 填补。

表 6　信息系统 S^0(Key&Rpt_RS 算法)

Table 6　Information system S^0(Key&Rpt_RS algorithm)

编号	a_1	a_2	a_3	a_4
1	*	0	1	0
2	0	0	*	0
3	1	0	1	0
4	1	0	1	0
5	1	1	1	0
6	1	0	1	0

计算 M^0，MAS_i^0，MOS^0，NS_i^0：$M^0(1,1)=\varnothing$，$M^0(1,2)=\varnothing$，$M^0(1,3)=\varnothing$，$M^0(1,4)=\varnothing$，$M^0(1,5)=\{a_2\}$，$M^0(1,6)=\varnothing$，…；$MAS_1^0=\{a_1\}$，$MAS_2^0=\{a_3\}$，$MAS_3^0=\varnothing$，$MAS_4^0=\varnothing$，$MAS_5^0=\varnothing$，$MAS_6^0=\varnothing$；$MOS^0=\{1,2\}$；$NS_1^0=\{2,3,4,6\}$，$NS_2^0=\{1\}$，$NS_3^0=\{1,4,6\}$，$NS_4^0=\{1,3,6\}$，$NS_5^0=\varnothing$，$NS_6^0=\{1,3,4\}$。

由计算可得，数据 x_1 存在 4 条相似数据：x_2，x_3，x_4，x_6。这 4 条数据在属性 a_1 存在两种不同取值 0 和 1 的情况下分别对两种取值的出现次数进行计数，取出现次数最少的属性值对 $a_1(x_1)$进行填补。由此，可以得到 S^1，如表 7 所列。

表 7　信息系统 S^1(Key&Rpt_RS 算法)

Table 7　Information system S^1(Key&Rpt_RS algorithm)

编号	a_1	a_2	a_3	a_4
1	0	0	1	0
2	0	0	1	0
3	1	0	1	0
4	1	0	1	0
5	1	1	1	0
6	1	0	1	0

从表 7 可以看出，S^1 已不存在缺失，可知 $S^2=S^1$。这时，考虑关键属性中的缺失状况，选择相似对象 x_1，x_3，x_4 中出现次数最少的关键属性值进行填补，得到表 8 所列的结果，该信息系统中已不存在缺失值，即算法结束。

表 8　引入关键属性的完备信息系统 S(Key&Rpt_RS 算法)

Table 8　Complete information system S considering key attribute (Key&Rpt_RS algorithm)

编号	a_1	a_2	a_3	a_4	a_5
1	0	0	1	0	0
2	0	0	1	0	1
3	1	0	1	0	0
4	1	0	1	0	1
5	1	1	1	0	1
6	1	0	1	0	1

4.3 实验结果

本实验使用命中率作为评价两种算法补齐效果的指标，命中率的计算公式如下：

$$命中率=\frac{正确补齐的属性值数量}{总缺失属性值数量}\times 100\% \tag{5}$$

为验证算法的效果，实验提前随机设置了一些缺失值，填补结束后，将填补的结果与真实值做对比，从而得到两种算法对缺失数据填补的命中率，实验结果如图 3 所示。

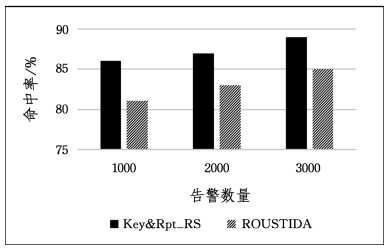

图 3 两种算法对缺失数据填补的命中率结果

Fig. 3 Hit rate of two algorithms for missing data imputation

从图 3 可以看出，在本文所述的包含关键属性且具有重复性特点的数据集上，改进后的 Key&Rpt_RS 算法对缺失数据集的填补效果明显优于传统的 ROUSTIDA 算法。

结束语 本文提出的 Key&Rpt_RS 算法是一种针对包含关键属性且具有重复特性的数据缺失更加有效的填补算法，多适用于工程技术领域内对故障告警数据的预处理过程。相较于传统的 ROUSTIDA 算法，所提算法大大提高了缺失数据填补的命中率和准确性；并且，当数据量越多时，由于出现与缺失数据相似的数据的概率越大，因此命中率也会得到提高；另外，该算法继承了粗糙集理论的优点，对数据的先验性要求不高，且处理结果具有客观性，有一定的借鉴价值。本文提出的方法针对具有关键属性的重复性数据，虽然提高了缺失补齐的命中率，但对数据的结构特点的要求较高，具有一定的局限性。未来，需要对不同结构特点的数据缺失进行高命中率的补齐，进而提出统一的缺失补齐算法。

参 考 文 献

[1] RUBIND B. Multiple imputation for nonresponse in surveys[J]. Journal of Marketing Research, 1987, 137(1): 180.

[2] SHUAI P, LI X S, ZHOU X H, et al. Theresearchprocesson statistical processing of missing data[J]. Chinese Journal of Health Statistics, 2013, 30(1): 135-139. (in Chinese)
帅平，李晓松，周晓华，等. 缺失数据统计处理方法的研究进展[J]. 中国卫生统计，2013，30(1)：135-139.

[3] YUE Y, TIAN K C. Review of data missing and its imputation method[J]. Journal of Preventive Medicine Information, 2005, 21(6): 683-685. (in Chinese)
岳勇，田考聪. 数据缺失及其填补方法综述[J]. 预防医学情报杂志，2005，21(6)：683-685.

[4] JIN Y J. Imputation adjustment method for missing data[J]. Journal of applied statistics and management, 2001, 20(6): 47-53. (in Chinese)
金勇进. 缺失数据的插补调整[J]. 数理统计与管理，2001，20(6)：47-53.

[5] DEMPSTER A P. Maximum likelihood estimation from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society, 1977, 39(1): 1-38.

[6] JIN Y J. Adjusting for Missing Data by Weighting in Survey Analysis[J]. Journal of applied statistics and management, 2001 (5): 61-64. (in Chinese)
金勇进. 缺失数据的加权调整(系列之 IV)[J]. 数理统计与管理，2001(5)：61-64.

[7] ROBINS J M, ROTNITZKY A, ZHAO L P. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed[J]. Journal of the American Statistical Association, 1994, 89(427): 846-866.

[8] ZHANG Z H, LIU W Q. An Improved Algorithm Based on the Incomplete Data of the Rough Set Theory[J]. Computer Engineering & Science, 2002, 24(4): 41-42. (in Chinese)
张振华，刘文奇. 一种基于粗集理论不完备数据的改进算法[J]. 计算机工程与科学，2002，24(4)：41-42.

[9] DUAN P, ZHUANG H L, HE L, et al. Improved algorithm based on incomplete data analysis method[J]. Computer Engineering and Design, 2009, 30(7): 1681-1684. (in Chinese)
段鹏，庄红林，何磊，等. 不完备数据分析方法(ROUSTIDA)的改进算法[J]. 计算机工程与设计，2009，30(7)：1681-1684.

[10] TIAN S X, WU X P, WANG H X. Improved method for data reinforcement based on ROUSTIDA[J]. Journal of Naval University of Engineering, 2011, 23(5): 11-15. (in Chinese)
田树新，吴晓平，王红霞. 一种基于改进的 ROUSTIDA 算法的数据补齐方法[J]. 海军工程大学学报，2011，23(5)：11-15.

[11] DING C R, LI L S. Improved ROUSTIDA algorithm based on similarity relation vector[J]. Computer Engineering and Applications, 2014, 50(13): 133-136. (in Chinese)
丁春荣，李龙澍. 基于相似关系向量的改进 ROUSTIDA 算法[J]. 计算机工程与应用，2014，50(13)：133-136.

[12] PAWLAK Z. Rough set[J]. International Journal of Computer & Information Sciences, 1982, 11(5): 341-356.

[13] 张文修. 粗糙集理论与方法[M]. 北京：科学出版社，2001.

[14] SKOWRON A, RAUSZER C. The Discernibility Matrices and Functions in Information Systems[M] // Intelligent Decision Support. Springer, Dordrecht, 1992: 331-362.

[15] 王国胤. Rough 集理论与知识获取[M]. 西安：西安交通大学出版社，2001.

[16] ZHANG W, LIAO X F, WU Z F. An incomplete data analysis approach based on rough set theory[J]. 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[J]. 模式识别与人工智能，2003，16(2)：158-163.

[17] MENG J, LIU Y C, MO H B. New method of packing missing data based on rough set theory[J]. Computer Engineering and Applications, 2008, 44(6): 175-177. (in Chinese)
孟军，刘永超，莫海波. 基于粗糙集理论的不完备数据填补方法[J]. 计算机工程与应用，2008，44(6)：175-177.