基于元组相似度的不完备数据填补方法研究

王俊陆[1] 王 玲[1] 王 妍[1,2] 宋宝燕[1]
(辽宁大学信息学院 沈阳 110036)[1] (东北大学信息与工程学院 沈阳 110819)[2]

摘 要 随着互联网及信息技术的发展,数据缺失、损坏等问题越来越普遍,尤其随着数据收集工作从人工转向机器,存储介质的不稳定性及网络传输出现遗漏等原因都导致数据缺失更加严重。数据库中大量的缺失值不仅严重影响了用户查询质量,还对数据挖掘与数据分析结果的正确性造成了影响,进而误导决策。目前,对缺失数据的填补还没有一种比较通用的方法,大部分策略都是针对某一类型的缺失值问题进行处理。因此,针对不同缺失类型同时出现在不完备数据中的复杂情况,提出了一种基于元组相似度的不完备数据填补方法(IATS)。采用数据挖掘的方法提取出不完备数据集中的加权关联规则,并根据此规则进行常规缺失数据的填补,而对于数据集的异常缺失问题,又引入数据推荐算法,采用推荐筛选策略进行元组相似度的计算并实现相应填补,在很大程度上提高了数据的有效利用率和用户查询结果的质量。实验表明,IATS策略在保证填补率的前提下具有更好的准确率。

关键词 海量数据,缺失类型,加权关联规则,元组相似度
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.02.013

Missing Data Imputation Approach Based on Tuple Similarity

WANG Jun-lu[1] WANG Ling[1] WANG Yan[1,2] SONG Bao-yan[1]
(School of Information, Liaoning University, Shenyang 110036, China)[1]
(School of Information Science and Engineering, Northeastern University, Shenyang 110819, China)[2]

Abstract With the development of Internet and information technology, the data loss, damage and other problems become more and more popular. Especially with data collection from the manual to machine, storage medium is not stability, transmission omissions appear and other reasons, resulting that missing data are more serious. A large number of missing values in the database not only seriously affect the quality of the query, but also affect the accuracy of the results of data mining and data analysis. At present, there is not a general method to deal with missing data. Most of the strategies are based on the problem of the missing value of a certain type. Therefore, in view of this complex situation of that the different deletion types also appear in the incomplete data at the same time, this paper put forward missing data imputation approach based on tuple similarity(IATS). Incomplete data sets of weighted association rules are extracted by the method of data mining, and according to the rules imputate normal missing data, and for abnormal missing data, this paper introduced data recommendation algorithm, the recommended screening strategy of tuple similarity calculation and the realization of the corresponding fill, and then it greatly improves the data effective utilization rate and user query result quality. The experimental results show that the IATS strategy has better accuracy under the premise of ensuring the filling ratio.

Keywords Massive data, Deletion type, Weighted association rules, Tuple similarity

1 引言

随着数据库技术和数据收集技术的快速发展,人们获取数据的速度呈指数级增长,对数据的分析与挖掘技术的要求也越来越高。数据质量管理开始得到各企业与研究机构的重视,其中,数据缺失是影响数据质量的重要因素之一。统计表明,在数据采集过程中大约有至少5%的缺失或损坏数据会被引入到数据集中[1-3],导致出现"数据丰富,但信息贫乏"[4]的问题。而数据预处理就是对这些缺失或损坏的数据进行相应的修复,其工作量在数据挖掘的过程中几乎占到了

到稿日期:2015-10-12 返修日期:2016-01-21 本文受国家自然科学基金项目(61472169,61472072),国家科技支撑计划项目(2012BAF13B08),国家"973"重点基础研究发展计划前期研究专项(2014CB360509),辽宁省科学事业公益研究基金项目(2015003003),辽宁大学科研基金(科技类)项目(LDQN2015001)资助。

王俊陆(1988—),男,硕士生,主要研究方向为大数据处理、数据库理论,E-mail:wangjunlu@lnu.edu.cn;王 玲(1989—),女,硕士生,主要研究方向为不完整数据填补,E-mail:wlcomeon_2014@163.com;王 妍(1978—),女,博士生,副教授,主要研究方向为海量数据查询处理、感知数据处理,E-mail:wang_yan@lnu.edu.cn;宋宝燕(1965—),女,博士,教授,CCF会员,主要研究方向为数据库理论与技术、RFID事件流处理技术、海量数据处理技术、图数据处理技术等,E-mail:bysong@lnu.edu.cn。

80%[5]，因此，在数据预处理阶段对不完备数据集数据进行适当的填补，对后续的数据分析、聚类和挖掘等操作有很大益处[6]。

针对上述情况，提出了一种基于元组相似度的不完备数据填补方法 IATS(Imputation Approach based on Tuple Similarity)，该方法主要解决不同数据缺失类型同时出现在不完备数据中的复杂问题。采用加权关联规则对常规缺失数据进行填补，而对于数据集中的异常缺失问题，本文又引入数据推荐算法，采用推荐筛选策略找到相似元组，计算元组相似度并建立相似矩阵，根据最大相似元组中的完整数据值对目标元组进行相应填补，很大程度地提高了数据的有效利用率和用户查询结果的质量。

本文第 2 节总结目前国内外对缺失数据填补的相关工作；第 3 节介绍数据缺失的类型及加权关联规则的填补策略；第 4 节提出基于元组相似度的方法进行缺失数据填补；第 5 节通过实验，从填补率、准确率和时效性 3 个方面证明 IATS 方法的有效性；最后对全文进行总结。

2　相关工作

目前，国内外很多专家学者对数据缺失填补问题进行了大量研究，缺失值的填补是提高数据挖掘结果质量的重要任务。许多缺失值填补算法已经应用于各种领域[7-10]，其中最简单的方法是将含有缺损属性的数据看成是对数据挖掘的正确性无影响的数据，将其直接删除，实验研究[11-12]已表明这种方法对于缺损属性的元组相对全体元组所占比例较小的数据库比较有效，如果所有的元组都有数据缺损，这种方法将删除大量数据，导致有用信息丢失，最终影响到数据挖掘的精度。同样地，文献[13]通过多个实验比对，证明均值填补也是一种不可靠的填补技术。另一常用的填充算法就是回归方法[14-16]，包括线性回归、多元回归、logistic 回归等，这一类方法用完整数据上建立的拟合模型来预测缺失变量的取值，拟合的优劣取决于自变量的选择和训练集的完备程度。

另一类方法有：1）基于关联规则集的缺失值填补方法(Missing Values Completion，MVC)[17]，该方法利用 RAR 算法[20]生成的关联规则集来填补缺失值，相对于缺失值填补技术的有效性[4]，MVC 仅仅增加预测的准确性，而没有考虑缺失值的填补率，致使推理时的关键属性可能被忽略[4,18]，所以本文提出加权关联规则来确保关键属性不被忽略；2）期望最大化填补(EMI)[19]，EMI 算法克服均值填补、特殊值填补等常用方法预测精度低的缺点。这种方法的主要缺点是填补缺失值时，EMI 算法使用的信息来自整个数据集，可能陷入局部极值，收敛速度较慢，且计算复杂，因此只适合表现出很强的相关性的属性数据。基于 k 最近邻算法的缺失值填补(kNNI)[21]，kNNI 不需要利用整个数据集的信息，使用 k 个相似记录；3）填补缺失值，该方法简单并且在局部相关性很强的数据集中表现很好，而在大数据集上能耗非常大；4）基于决策树的缺失值填补(DMI)[22]，该方法利用决策树和 EMI 算法进行数据缺失填补，在时效性方面表现出很好的效果，但是没有解决一条记录中存在多个缺失值的问题。

综上所述，对上述各个方法存在的问题进行分析比较，以填补率和准确率进行评价，通过实验证明，本文提出的 IATS 填补策略为缺失数据问题的研究提供了一种较优的方法。

3　基于加权关联规则的数据填补方法

数据缺失机制的概念首先是由 Rubin[23]于 1976 年提出的。为了更好地描述数据缺失的内在原因，且各种处理缺失数据的方法都是建立在数据缺失类型的某种假定上，故本节先对数据缺失类型进行分类，按照属性之间是否具有相关性数据缺失类型分成两类，然后对常规型缺失采用加权关联规则进行数据缺失的填补。

3.1　数据缺失类型

数据缺失类型其实是在含有缺失值的数据集中归纳总结空缺与非空缺值的关系而形成的。本文将数据集中不含缺失值的变量(属性)称为完全变量，数据集中含有缺失值的变量称为不完全变量。缺失类型主要可以分为以下两种。

设 $A=(A_{ij})$ 表示全体数据，$M=(M_{ij})$ 表示缺失模型，M_{ij} 表示对应的 A_{ij} 是否缺失，本文数据缺失模型就是以基于 A 的 M 的条件分布来表示，即 $f(M|A,\Phi)$，其中 Φ 代表缺失值。A_{obs} 和 A_{mis} 分别代表 A 中完全变量和不完全变量部分。

(1)常规缺失(Missing In Normal，MIN)：如果某一数据的缺失与其他完全属性值有关，且缺失值可以由完全属性预测，则称这种缺失为常规型数据缺失。

$$f(M|A,\Phi)=f(M|A_{obs},\Phi) \tag{1}$$

(2)异常缺失(Missing In Abnormal，MIA)：如果某一数据的缺失与其他属性值无关，且缺失值不可以由完全属性预测，但可以从自身属性值中进行预测，则称这种缺失为异常型数据缺失。

$$f(M|A,\Phi)=f(M|A_{mis},\Phi) \tag{2}$$

常规型数据缺失表明了属性之间具有一定的关联，那么采用加权关联规则进行缺失值填补。而异常型数据缺失需要去发现新信息，根据这一特点本文基于元组相似度进行相应缺失数据的填补工作。本文方法的整体结构如图 1 所示。

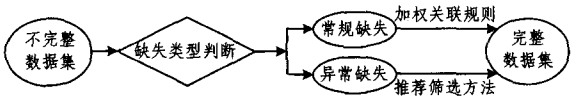

图 1　缺失数据的填补流程

如图 1 所示，本文首先从不完整数据集中判断数据缺失的类型，应用既有的方法[24-25]进行判断，对此本文不做重点阐述。针对两种缺失类型同时出现的情况，本文提出的 IATS 策略采用加权关联规则对常规缺失进行填补，又采用数据推荐筛选策略对异常缺失进行填补，最后生成完整数据集。

3.2　基于加权关联规则的填补方法

关联规则的挖掘能有效建立数据库中属性间的关联，因此被挖掘出来的关联规则通常能用于恢复数据库中丢失的属性值。但某些时候用户可能对某些项目更感兴趣，希望加强这些项目对规则的影响，而对另外一些项目不感兴趣，要求削弱这些项目的作用；同时在挖掘出的关联规则中可能会出现多条规则填补同一处缺失值的情况，例如 $A1=c\Rightarrow A3=d$ 与 $A2=a\Rightarrow A3=d$，是应用第一条还是第二条规则填补 $A3$?为此，本文针对用户对于项目的感兴趣程度对不同的项目设置

不同的权重，以提高规则的有用性和填补的准确率。

3.2.1 加权关联规则

设关系数据库 D 可用一个二维表 S 来表示，如表 1 所列，表 2 是带有缺失值的二维表 S'。

表 1 完整数据表 S

ID	A1	A2	A3	A4
1	a	a	a	c
2	a	a	b	c
3	a	b	c	c
4	a	b	d	c
5	b	b	e	d
6	b	b	f	d
7	b	b	g	d
8	b	c	h	d
9	d	d	i	e
10	c	f	j	f

表 2 不完整数据表 S'

ID	A1	A2	A3	A4
1	?	a	a	c
2	a	a	b	?
3	a	b	c	c
4	a	b	d	c
5	?	b	e	d
6	b	b	f	?
7	b	b	g	d
8	b	c	h	d
9	?	d	i	e
10	?	f	j	f

表中，行称为元组(记录)，列称为属性。如表 1 所列，4 个属性为(A1,A2,A3,A4)，其中 A1 可能取值为 a,b,c,d；A2 可能取值为 a,b,c,d,f；A3 可能取值为 a,b,c,d,e,f,g,h,i,j；A4 可能取值 c,d,e,f。表 S' 中含有缺失值，如表 2 所列，其中?代表缺失值。一个属性-值对称为一个项目，D 中所有项目的集合称为全总项目集，记为 $I=\{i_1,i_2,\cdots,i_k\}$，每项都是 I 的一个子集，表 1 的 I 表示成：$I=\{A1=a,A1=b,A1=c,A1=d,A2=a,A2=b,A2=c,A2=d,A2=f,A3=a,\cdots,A3=j,A4=c,A4=d,A4=e,A4=f\}$。$W=\{w_1,w_2,\cdots,w_k\}$ 是对应 I 的权重集，其中 w_j 表示项目 i_j 的权重，且 $0\leqslant w_j\leqslant 1,j=\{1,2,\cdots,k\}$。设 $X=\{x_1,x_2,\cdots,x_p\},Y=\{y_1,y_2,\cdots,y_q\}$ 是 I 的子集且 $X\cap Y=\emptyset$，数据集 D 中项目集 X 的支持率和置信度记为 $Support(X)$ 和 $Confidence(X)$。参考文献[26]，作如下概念描述。

定义 1 项集加权的支持度(weigthed support)为：

$$(\sum_{i_j} w_j)\times support(x) \tag{3}$$

定义 2 加权关联规则 $X\Rightarrow Y$ 的支持度为：

$$Wsup=(\sum_{i_j\in X\cup Y} w_{i_j})\times support(X\cup Y) \tag{4}$$

定义 3 加权关联规则 $X\Rightarrow Y$ 的可信度为：

$$Wconf=\frac{support(X\cup Y)}{support(X)} \tag{5}$$

定义 4(加权关联规则) 同时满足最小加权支持度(Wminsup)和最小加权可信度(Wminconf)的称为加权关联规则。

3.2.2 加权关联规则的数据填补方法

在表 2 中，假设用户的查询条件为 $A1=a$，则执行如图 2 所示的填补过程。

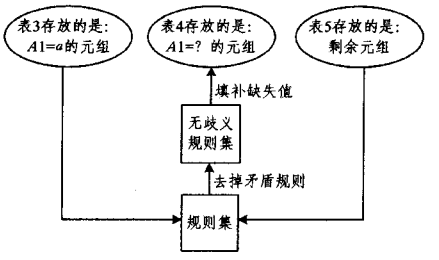

图 2 基于查询条件分表

根据查询条件将表 2 分成如表 3－表 5 所列的元组。

表 3 $A1=a$ 的元组

ID	A1	A2	A3	A4
2	a	a	b	?
3	a	b	c	c
4	a	b	d	c

表 4 $A1=?$ 的元组

ID	A1	A2	A3	A4
1	?	a	a	c
5	?	b	e	d
9	?	d	i	e
10	?	f	j	f

表 5 剩余元组

ID	A1	A2	A3	A4
6	b	b	f	?
7	b	b	g	d
8	b	c	h	d

表 3 中存放满足 $A1=a$ 的元组，表 4 中存放 $A1=?$ 的元组，表 5 中存放剩余元组。在表 3 和表 5 中分别进行关联规则查找，生成初始规则集，挖掘出一对矛盾规则 $A2=b\Rightarrow A1=a$ 和 $A2=b\Rightarrow A1=b$，这时不知道表 4 中当 $A2=b$ 时 A1 填补哪个值，所以应用加权关联规则来进行更精确的填补。权值的分配根据查询相关项目频率的高低给定，如表 6 所列。

表 6 项目权值分配表

Item	Weight
$A1=a$	0.6
$A1=b$	0.3
$A2=a$	0.1
$A2=b$	0.2
$A4=c$	0.3

根据式(4)、式(5)计算出加权置信度，$A2=b\Rightarrow A1=a$ 这条规则更高，那么用这条规则填补图 2 中表 4 的缺失值并加入无歧义规则集中。

在挖掘出可供利用的关联规则之后，将其存入此类查询规则库中，待后续查询过程中填补缺失值时使用。一般而言，可将上述规则转成规则库中的 SQL 语法，用下列规则式表示：IF 前提(antecedent) THEN 结论(consequent)规则[27]。其意义为：若前提描述的条件满足，则执行结论所指定的动作。

4 基于元组相似度的填补方法

如上文所述，采用基于加权关联规则可以填补常规型缺失(MIN)；但对于异常缺失(MIA)，用这种方法进行填补可能会降低准确率，因此本文采用数据推荐策略来计算元组相似度并实现数据填补。

4.1 元组相似度的计算

计算元组相似度之前，首先要找到相似元组集合，相似元组的查找根据目标元组的项目集合，即找到与目标元组具有

相同项目的元组，其中这些元组中可以包含缺失值，最后利用公式建立相似矩阵。本节主要从以下两个方面进行：1）找到相似元组集合；2）计算相似矩阵，找到与目标元组相似性最大的元组，用此元组的完整属性值填补目标元组的缺失值。

本文用余弦相似度计算两个元组之间的相似度。设 $N(u)$ 代表记录 u 所有的非空项集合，$N(v)$ 代表记录 v 所有的非空项集合，那么 u,v 的相似度计算如下。

余弦相似度：

$$C_{uv}=\frac{|N(u)\cap N(v)|}{\sqrt{|N(u)|\times|N(v)|}} \tag{6}$$

例如，填补表 2 中的 ID_9 的 $A1$ 属性值，首先需要从总数据表中找到与 ID_9 具有相同项目的元组，如表 7－表 9 所列。

表 7　满足条件的元组 1

ID	A1	A2	A3	A4
12	d	d	i	?

表 8　满足条件的元组 2

ID	A1	A2	A3	A4
14	c	d	c	e

表 9　满足条件的元组 3

ID	A1	A2	A3	A4
18	c	?	i	f

根据表 7－表 9 建立一个项目-元组的排列表，如表 10 所列。

表 10　项目-元组的排列表

项目	对应元组
$A2=d$	ID_9, ID_{12}, ID_{14}
$A3=i$	ID_9, ID_{12}, ID_{18}
$A4=e$	ID_9, ID_{14}
$A1=c$	ID_{14}, ID_{18}

根据表 10 建立一个 4＊4 的矩阵，如图 3 所示，其中主对角线表示元组自身的关联，本文均设置为 0。对于具有相同项目的元组，若两两之间相同则加 1。例如 $A2=d$ 的元组有 ID_9, ID_{12}, ID_{14}，那么在矩阵中它们两两加 1，如图 3 所示。

	ID_9	ID_{12}	ID_{14}	ID_{18}
ID_9	0	2	2	1
ID_{12}	2	0	1	1
ID_{14}	2	1	0	1
ID_{18}	1	1	1	0

图 3　初始矩阵

应用式(6)生成相似矩阵，图 3 生成的初始矩阵代表式(6)中的分子部分，继续进一步的计算得到如图 4 所示的相似矩阵。

	ID_9	ID_{12}	ID_{14}	ID_{18}
ID_9	0	$\frac{2}{\sqrt{3\times3}}$	$\frac{2}{\sqrt{3\times4}}$	$\frac{1}{\sqrt{3\times3}}$
ID_{12}	$\frac{2}{\sqrt{3\times3}}$	0	$\frac{1}{\sqrt{3\times2}}$	$\frac{1}{\sqrt{3\times2}}$
ID_{14}	$\frac{2}{\sqrt{3\times4}}$	$\frac{1}{\sqrt{3\times2}}$	0	$\frac{1}{\sqrt{3\times3}}$
ID_{18}	$\frac{1}{\sqrt{3\times3}}$	$\frac{1}{\sqrt{3\times2}}$	$\frac{1}{\sqrt{3\times3}}$	0

图 4　相似矩阵

由图 4 得出，与 ID_9 相似度最大的元组是 ID_{12}，然后判断 ID_{12} 中 $A1$ 属性值是否缺失，若缺失，再判断 ID_{14} 的 $A1$ 属性值，以此类推。

4.2　基于元组相似度的数据填补算法

如上文所示，基于元组相似度的数据填补方法的关键是建立相似矩阵，如算法 1 所示。

算法 1　SimTuplesImp

```
Input:不完整数据集 D
Output:完整数据集 D'
1. Step1:生成项目-元组排列表 S
2. Initialize:相同项目数组 A
3.   For each Attribute in D
4.    If Attribute not empty
5. For each Tuple in database
6.      If Tuple. Attribute==Attribute
7.      Tuple ID->A
8.   A->S
9. Step2:生成初始矩阵
10. Initialize:空初始矩阵 M
11. For each Attribute in S
12.    If Id in S M[x][y]++
13. Step3:生成相似矩阵
14.    For each item in D
15. For each item in M
16.      //T(ID(x)) T(ID(y))非缺失属性个数
17.      M[x][y]=M[x][y]/T(ID(x))×T(ID(y))
18. Step4:填补缺失值
19. For each item in M
20.    If M[x][y] is max and item. Attribute not empty
21.    D. attribute=item. Attribute
```

5　实验与分析

为了便于验证本文所提出的 IATS 方法的正确性，实验使用原无缺失数据的完整数据集，对其部分属性值进行人工添加缺失，使之成为符合实验要求的数据集。实验所用数据集、最小支持度和最小置信度的设定参照 RAR 算法的实验[20]，数据集是用 Gabor Melli 创建的 The datgen Dataset Generator(Version 3.1)生成的 500 条数据，每条数据包含 10 个属性，每个属性包含 5 种不同值，并将两种缺失类型按不同比率同时添加到数据集中，最小支持度设为 10%，最小置信度设为 80%。实验环境是 Intel 酷睿 2. 2GHz 的 CPU，4GB 内存，操作系统为 Windows 7 64 位。本文从填补率、准确率和时效性 3 方面进行实验对比分析，并对实验中数据的两种缺失类型出现的比率进行设定，如表 11 所列。

表 11　两种缺失类型的比率/%

缺失类型	缺失比率		
	MIN>MIA	MIN<MIA	MIN≈MIA
MIN	5～15	1～5	10～15
MIA	1～5	5～15	10～15

5.1　填补率分析

填补率是填补工作必须满足的基本条件之一。设 n 是缺失值的总数，a 是填补缺失值的总数，则填补率计算公式如下：

$$p=a/n$$

其中，p 介于 0～1 之间，当 p 等于 1 时代表完整填补，结果如图 5 所示。

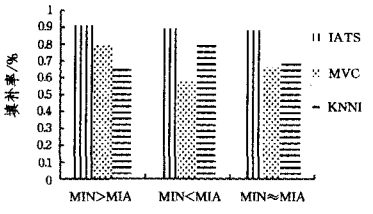

图 5 填补率分析

如图 5 所示，随着数据缺失率的增加，MVC 和 KNNI 算法的填补率都有一定程度的降低，且这两种算法无法实现异常缺失填补，同时随常规缺失与异常缺失所占比重的不同而呈现出不同效果；而本文提出的 IATS 策略在填补率方面表现得较为稳定。

5.2 准确率分析

准确率是填补工作必须满足的基本条件之一。设 n 是缺失值总数，m 是正确填补缺失值的总数，则准确率计算公式如下：

$$c=m/n$$

其中，c 介于 0～1 之间，当 c 等于 1 时代表完美填补，结果如图 6 所示。

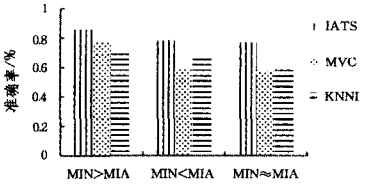

图 6 准确率分析

如图 6 所示，随着数据缺失率的增加，MVC 和 KNNI 算法的准确率都会有一定程度的降低，同时随着常规缺失与异常缺失所占的比重的不同而呈现出不同的效果；而本文提出的 IATS 策略在准确率方面表现得较为稳定。

5.3 时效性分析

算法时间效率指标选用平均时间，即为平均每次运行时间消耗。设时间消耗总计为 T，实验运行次数为 x，则平均时间公式如下：$AT=T/x$。实验结果如图 7 所示。

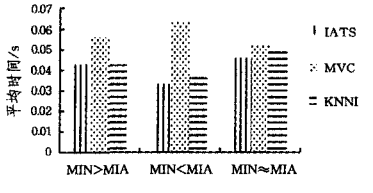

图 7 时效性分析

如图 7 所示，随着数据缺失率的增加，不同填补方法在时间消耗上都会有一定程度的升高，同时随着常规缺失与异常缺失所占比重的不同而呈现出异常的波动；而本文提出的 IATS 策略在时效性方面表现得较为稳定。

综上所述，无论缺失数据类型如何变化、缺失率如何增减，本文提出的 IATS 策略在填补率和准确率方面都优于 MVC 和 KNNI 算法，且缺失数据填补的质量也比较高，时间消耗相对较小。

结束语 本文在大数据环境下，针对数据挖掘在实际应用中往往面临的不完备数据集，提出了一种基于元组相似度的不完备数据填补方法 IATS。该方法主要解决不同数据缺失类型同时出现在不完备数据中的复杂情况，采用加权关联规则对常规缺失数据进行填补，采用推荐筛选策略计算元组相似度对缺失数据进行相应填补，提高了数据的有效利用率和用户查询结果质量。实验结果证明，与基于关联规则集的缺失值填补方法和基于 k 最近邻算法的缺失值填补方法相比，本文给出的基于 IATS 填补策略使数据填补工作能更加快速有效地实现。

参 考 文 献

[1] CHEN M, MAO S, LIU Y. Big Data: A Survey[J]. Mobile Networks & Applications, 2014, 19(2): 171-209.

[2] AZADEH A, ASADZADEH S M, JAFARI-MARANDI R, et al. Optimum estimation of missing values in randomized complete block design by genetic algorithm[J]. Knowledge-Based Systems, 2013, 37(2): 37-47.

[3] AHMED I, AZIZ A. Dynamic Approach for Data Scrubbing Process[J]. International Journal on Computer Science & Engineering, 2015, 2(2): 416-423.

[4] 韩家炜. 数据挖掘：概念与技术[M]. 北京：机械工业出版社，2006.

[5] PYLE D. Data preparation for data mining [M]. Academic Press, 1999.

[6] REN Y. Data preprocessing for data mining[D]. Turku University, 2013.

[7] CHENG K O, LAW N F, SIU W C. Iterativebicluster based least square framework for estimation of missing values in microarray gene expression data[J]. Pattern Recognit, 2012, 45(4): 1281-1289.

[8] XIANGCHAO G, ALAN W C L, HONG Y. Microarray missing data imputation based on a set theoretic framework and biological knowledge[C]// International Conference on Pattern Recognition. IEEE Computer Society, 2006: 1608-1619.

[9] JUNGER W L, LEON A P D. Imputation of missing data in time series for air pollutants[J]. Atmospheric Environment, 2015, 102: 96-104.

[10] ZHANG G, HUANG K C, ZHENG X, et al. Across-Platform Imputation of DNA Methylation Levels Incorporating Nonlocal Information Using Penalized Functional Regression[J]. Genetic Epidemiology, 2016, 40(4): 333-340.

[11] ALLISON P D. Missing Data[D]. Sage University Papers Series on Quantitative Applications in the Social Sciences, Thousand Oaks, Sage, CA, 2001: 7-136.

[12] LITTLE, RUBIN R J A, STATISTICAL D B. Analysis with Missing Data(seconded)[M]. John Wiley and sons, Hoboken, NJ, 2002.

[13] HULSE J V, KHOSHGOFTAAR T M. A comprehensive empirical evaluation of missing value imputation in noisy software measurement data[J]. Journal of Systems & Software, 2008, 81(5): 691-708.

[14] YANG K, LI J, WANG C. Missing Values Estimation in Microarray Data with Partial Least Squares Regression[M]// Computational Science-ICCS 2006. Springer Berlin Heidelberg, 2006: 662-669.

[15] SHAN Y, DENG G. Kernel PCA regression for missing data estimation in DNA microarray analysis[C]// 2009 IEEE International Symposium on Circuits and Systems. 2009: 1477-1480.

(下转第 106 页)

MAE,从而预测精度更高。本文算法既考虑到用户特征对预测评分的影响,又考虑到专家信任对预测评分的影响,并且在近邻 $k=30$ 时,能够获得最佳的推荐效果。

结束语 本文针对传统协作过滤算法面临的稀疏性问题提出了综合用户特征及专家信任的协作过滤算法。当新用户到达推荐系统时,利用用户填写的注册信息可以有效缓解推荐系统中的冷启动问题。同时该算法根据不同用户特征有不同的兴趣偏好,并结合专家意见,利用协作过滤算法进行预测推荐。在 MovieLens 数据集上的实验验证了本文提出的综合用户特征及专家信任的协作过滤算法的有效性。该算法提高了推荐精度,同时缓解了数据集的稀疏性和冷启动等问题。在今后的学习工作中,可以针对不同数据集进行分析,以得到最优参数值,从而进行更好的推荐。

参 考 文 献

[1] SHAFER,SEN S W,FRANKOWSKI,et al. Collaborative Filtering Recommender Systems[C]// International Conference on Intelligent Systems Design & Applications. IEEE,2015:438-443.

[2] YANG C,AI C C,JIANG B,et al. Demographic Attribute-based Collaborative Filtering Algorithm[J]. Journal of Chinese Computer Systems,2015,36(4):782-786. (in Chinese)
杨超,艾聪聪,蒋斌,等. 一种融合人口统计属性的协同过滤算法[J]. 小型微型计算机系统,2015,36(4):782-786.

[3] JANNACH D,ZANKER M,FELFERNIG A,et al. Recommender Systems:An Introduction[M]. Int. j. hum. comput. interaction,2010.

[4] SUN L F,HUANG M X. Collaborative filtering recommendation algorithm based on user characteristics and item attributes [J]. Application Research of Computers,2014,31(2):384-387. (in Chinese)
孙龙菲,黄梦醒. 综合用户特征和项目属性的协作过滤推荐算法[J]. 计算机应用研究,2014,31(2):384-387.

[5] SHUO L X,CHAI B F,ZHANG X D. Collaborative filtering algorithm based on improved nearest neighbors[J]. Computer Engineering & Applications,2015,51(5):137-141. (in Chinese)
硕良勋,柴变芳,张新东. 基于改进最近邻的协同过滤推荐算法[J]. 计算机工程与应用,2015,51(5):137-141.

[6] PENG D W,HU B. A Collaborative Filtering Recommendation Based on User Characteristics and Time Weight[J]. Journal of Wuhan University of Technology,2009,31(3):24-28. (in Chinese)
彭德巍,胡斌. 一种基于用户特征和时间的协同过滤算法[J]. 武汉理工大学学报,2009,31(3):24-28.

[7] CHOI K,SUH Y. A new similarity function for selecting neighbors for each target item in collaborative filtering[J]. Knowledge-Based Systems,2013,37(1):146-153.

[8] AMATRIAIN X,LATHIA N,PUJOL J M,et al. The Wisdom of the Few A Collaborative Filtering Approach Based on Expert Opinions from the Web[C]//Proceedings of International ACM SIGIR Conference on Research & Development in Information Retrieval. 2009:532-539.

[9] YUN L,YANG Y,WANG J,et al. Improving rating estimation in recommender using demographic data and expert opinions[C]// 2011 IEEE 2nd International Conference on Software Engineering and Service Science (ICSESS). IEEE,2011:120-123.

[10] BOζAλήÇ E,VOZALIS E,MARGARITIS K. Collaborative Filtering enhanced by Demographic Correlation[J]. Proceedings of the Aiai Symposium on Professional Practice in Ai Part of World Computer Congress,2004:293-402.

[11] LIU J G,ZHOU T,GUO Q,et al. Overview of the Evaluated Algorithms for the Personal Recommendation Systems[J]. Complex Systems & Complexity Science,2009,6(3):1-10. (in Chinese)
刘建国,周涛,郭强,等. 个性化推荐系统评价方法综述[J]. 复杂系统与复杂性科学,2009,6(3):1-10.

(上接第 102 页)

[16] LIU C C,DAI D Q,YAN H. The theoretic framework of local weighted approximation for microarray missing value estimation [J]. Pattern Recognition,2010,43(8):2993-3002.

[17] RAGEL A,CRÉMILLEUX B. MVC—a preprocessing method to deal with missing values[J]. Knowledge-Based Systems,1999,12(5/6):285-291.

[18] AGRAWAL R,SRIKANT R. Mining Quantitative Association Rules in Large Relational Tables[C]// ACM SIGMOD Conf. Management of Data. 1996:1-12.

[19] SCHNEIDER T. Analysis of Incomplete Climate Data:Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values[J]. Journal of Climate,2001,14(5):853-871.

[20] RAGEL A,CREMILLEUX B. Treatment of Missing Values for Association Rules[M]// Research and Development in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg,1998:258-270.

[21] GUSTAVO E,BATISTA A P A,Monard M C. An Analysis of Four Missing Data Treatment Methods for Supervised Learning [J]. Applied Artificial Intelligence,2003,17(5):519-533.

[22] RAHMAN M G,ISLAM M Z. Missing value imputation using decision trees and decision forests by splitting and merging records:Two novel techniques[J]. Knowledge-Based Systems,2013,53(9):51-65.

[23] RUBIN D B. Inference and missing data[J]. Biometrika,1976,63:581-592.

[24] LISTING J,SCHLITTGEN R. A Nonparametric Test for Random Dropouts[J]. Biometrical Journal,2003,45(1):113-127.

[25] PREISSER J S,WAGENKNECHT L E. Analysis of Smoking Trends with Incomplete Longitudinal Binary Responses[J]. Journal of the American Statistical Association,2000,95(452):1021-1031.

[26] WANG P,AN C,WANG L. An improved algorithm for Mining Association Rule in relational database[C]// 2014 International Conference on Machine Learning and Cybernetics (ICMLC). IEEE,2015:247-252.

[27] KONONENKO I,BRATKO I,ROSKAR E. Experiments in automatic learning of medical diagnostic rules[R]. Yugoslavia:Ljubljana,Lozef Institute,1984.