对非随机缺失中的缺失依赖关系研究

郑奇斌 刁兴春 曹建军

（解放军理工大学 指挥信息系统学院，江苏 南京 210007）

摘　要：数据缺失是一种十分重要而又很常见的数据质量问题。对缺失数据的常见处理方法为估计缺失值或者直接删除缺失记录。这两种方法都只用到了未缺失数据中的信息，而隐含在缺失记录中的信息则被舍弃了。在非随机缺失中各个缺失属性并不是独立的，其中存在着依赖关系。本文使用关联规则挖掘的技术，主要是关联规则挖掘，从发生非随机缺失的数据集中发现属性间的缺失依赖关系。利用发现的依赖关系，可以对数据分析或者信息采集改进提供帮助。通过在一个真实数据集上的实验，证明本文的方法可以有效的发现缺失中的依赖关系。

关键词：数据质量 非随机缺失 缺失依赖 关联规则

Research on Missing Dependencies in NMAR

Zheng Qi-bin, Diao Xing-chun, Cao Jian-jun

(College of Command Information Systems, PLA University of Science and Technology, Nanjing 210007, China)

Abstract: Data missing is an important and common data quality problem. The methods for dealing with missing data primarily attempting to infer the missing value or delete the record with missing data. Both of the methods only used the information of non-missing record. There are dependences among missing attributes. In this context, we use data mining techniques, mainly association rules mining, to find dependences among missing attributes in the dataset with NMAR. The dependence relationship can be useful to data analysis and information collection promotion. by applying our approach on a real dataset, we demonstrate the validity of it. The experiment result shows that the approach was able to find the dependence relationship of missing.

Keywords: Data Quality; NMAR; Missing Dependencies; Association Rule

1　引言

近年来，随着云计算和大数据概念和技术的不断发展，社会产生的数据总量正在飞速增长。这些产生的数据反过来又极大的促进了社会的发展，数据对社会活动的影响越来越大。然而由于信息供应链本质上存在的缺陷，我们在数据分析中所使用的数据集经常伴随着一些数据质量问题。这些数据质量问题常常会导致分析结果的失准，严重的影响了依靠分析结果作出的决策，导致表现不佳甚至错误的决策。

数据缺失是一种在日常生活、科研以及商业应用中都十分常见的数据质量问题。由于数据使用者往往不是数据的生产者，甚至完全不知道数据生产者的是谁，在使用数据前我们不能期望得到完整的数据集。面对存在缺失的数据集，分析人员如果忽略数据缺失的存在，通常会造成统计分析和推断的结果的偏差甚至错误[6] 。为了得到更为贴近

基金项目：国家自然科学基金资助项目（61371196）；中国博士后科学基金特别资助项目（201003797）；解放军理工大学预研基金项目（20110604）

作者简介：郑奇斌，男，1990 年生，硕士研究生，主要研究方向为数据工程；刁兴春，男，研究员，博士生导师；曹建军，男，博士。

联系方式：13255260915，E-mail：zhengqibin1990@163.com。

事实的分析结果，如何处理数据集中的缺失成为一个关键问题。

从缺失发生的原因来看，数据缺失分为完全随机缺失（Missing Completely At Random，MCAR）、随机缺失（Missing At Random，MAR）和非随机缺失（Not Missing At Random, NMAR）[5] 。由于在 MCAR 和 MAR 的情况中，基于似然估计的估计量是无偏的，所以 MCAR 和 MAR 被认为是可以忽略的。NMAR 中估计量常常是有偏的，这样的缺失被认为是不可忽略的，所以 NMAR 有时也被称作不可忽略机制（NI）[6] [7] 。

根据缺失值的所属属性是否为同一属性，数据缺失分为单值缺失和任意缺失。单值缺失中，所有缺失记录中的缺失数据都属于同一属性。在任意缺失中，缺失记录的缺失数据分布在不同的属性中。

在现实的应用中，对于缺失数据的处理主要有两种：缺失值估计和删除（也就是忽略缺失值）。但是对于非随机缺失（NMAR）的数据，因为数据缺失发生在数据采集的源头，基于似然估计的估计量通常是有偏的。同样如果忽略缺失数据（也就是直接删除存在缺失的记录），不但不能改变整体数据的偏离，而且会丢失存在于非缺失属性中的信息。这两种处理办法对于 NMAR 处理效果都不甚理想。

例 1 在一项涉及宗教信仰的调查中，如果一个人生活的地区中大部分人对于他的信仰的宗教有偏见，那么他在调查中很可能会隐瞒自己的宗教信仰（即不去填写涉及宗教信仰的条目）。

通过例 1 中我们可以看到，NMAR 的发生经常是由人的主观原因或者历史局限造成的，所以其中往往隐含着关于缺失原因信息。也就是说，在 NMAR 中发生缺失的记录往往具有一些可观测的特征，具有这些特征的记录发生缺失的概率更大。Monica Chiarini Tremblay 在[1] 对这种隐含在缺失中的倾向性模式进行了研究，提出了一种使用数据挖掘技术来发现单值缺失中存在的模式。

例 2 在对婚姻状况的调查中，如果一个人出于隐私的目的不回答关于是否婚配的问题，那么在对其配偶姓名以及子女数量的问题中他很有可能也会保持沉默。

但是在实际的应用中遇到的大部分都是任意缺失的情况，通过例 2 可以看到，由于每条记录中的不同属性实际上并不是完全独立的，一个属性的缺失往往会伴随着其他属性的缺失。为此本文尝试提出一种方法，利用知识挖掘的技术来发现任意缺失中存在的规律。

本文的结构如下：第 2 节中介绍了文章相关的一些概念；第 3 节全面的描述了文章采取的方法；第 4 节通过实验初步验证了方法的有效性；第五节是对文章研究结果的总结，并讨论了下一步研究的方向。

2 相关概念

2.1 数据缺失机制

数据缺失机制描述的是缺失数据同数据集中变量值的关系，从本质上说明了缺失发生的原因[6] 。缺失机制的概念是由 Rubin 在 1976 年首次提出，Rubin 将数据缺失按照原因划分为三类：完全随机缺失、随机缺失和非随机缺失[5] 。

记 D 为完整的数据集，D_{ij} 表示第 i 条记录的第 j 个属性的值，$D_{missing}$ 表示缺失的数据，D_{obs} 表示可观测数据。记 M 为缺失指示矩阵，$M_{ij}=1$ 当且仅当 D_{ij} 缺失，否则 $M_{ij}=0$。为研究缺失机制，考虑以下条件概率[4] ：

$$P(M \mid D,\Phi)$$

其中 Φ 表示与缺失机制有关的某个未知参数。

1）缺失类型属于 MCAR，如果

$$P(M \mid D,\Phi) = P(M \mid \Phi) \qquad (1)$$

MCAR 表明数据缺失发生的概率同数据集中所有数据值都没有关系。

2）缺失类型属于 MAR，如果

$$P(M \mid D,\Phi) = P(M \mid D_{obs},\Phi) \qquad (2)$$

MAR 表明数据缺失发生的概率同缺失属性的值自身无关，只依赖于数据集 D 中其他可观测值 $D_{missing}$。

3）缺失类型属于 NMAR，如果

$$P(M \mid D,\Phi) = P(M \mid D_{Missing},\Phi)\quad(3)$$

或

$$P(M \mid D,\Phi) = P(M \mid D_{Missing},D_{obs},\Phi)\quad(4)$$

NMAR 表明数据缺失发生的概率依赖于缺失属性本身的值，或者既依赖于发生缺失的值也依赖于可观测的值，也就是依赖于数据集整体。

举例说明，对于有两个属性的数据集：属性 *A* 是“年龄”，属性 *B* 是“收入”，发生缺失的属性是 *B*。

如果收入属性发生缺失的概率与年龄和收入的属性值都无关，那么它属于 MCAR。如果对于不同年龄，收入发生缺失的概率不同，而对应于同样年龄的不同收入，收入的缺失概率相同，那么它属于 MAR。如果对于同样年龄不同的收入，表现出不同的缺失概率，那么它属于 NMAR。

本文中研究的对象是属于 NMAR 的任意缺失问题，主要考虑的是缺失属性同其他缺失属性之间的关系，即 $P(M|D,\Phi)=P(M|D_{Missing},\Phi)$的情况。在实际情况中，由于领域知识的缺乏或者缺失分布本质上就不可知，很难知道具体的缺失机制。所以本文并不计算具体的缺失概率分布，而是试图通过数据挖掘的方法找出情况（3）NMAR 中缺失属性具体依赖的 $D_{Missing}$ 是哪些属性。

3 研究方法

本章首先针对缺失模式的特点，使用关联规则来描述缺失模式。然后使用知识挖掘方法来挖掘缺失中的这种关联规则，整个过程包括数据准备，规则挖掘，规则约减等步骤。

3.1 缺失模式描述

要研究多个属性缺失模式，首先要解决的问题是如何表示存在于多个属性缺失中的倾向性规律。

由于我们关注的偏向模式本质上表示的是缺失值和其他值（非缺失值以及其他缺失值）之间的相关关系，所以使用包含“与”、“或”、“蕴含”等连接词的逻辑表达式，即规则来描述这种模式是一种比较恰当的方式。为了更好的表述规则，也为了提高挖掘效率，需要对规则的形式进行约束。

为此将规则拆为规则右侧（RHS）和规则左侧（LHS）两个部分分别进行分析：

规则右侧通常表示结果，一般是是用户最为关注的问题。在不同的应用背景下用户关注的问题可能是不一样的，具体来说就是用户数据集中的各个属性关注度是不同的。比如：在一个关于毕业生就业情况的调查中，是否就业就是用户最为关注的问题（当然这样的属性可能有多个）。所以本文提出的方法中，规则右侧的属性就是用户关注的出现缺失的属性，本文称之为目标属性。

规则左侧表示的是一种条件，更进一步说就是导致了用户所关注结果的条件。用户对这里会出现哪些属性是不知道的，所以规则左侧中的备选属性集为（经过数据降维后）全部属性。

3.2 挖掘方法

关联规则挖掘（Association Rules Mining）即营销领域中的购物篮分析，是一项广泛使用的数据挖掘技术。购物篮分析中的目标是在大规模交易数据库中找到频繁出现的规则，类似的在本文中我们要挖掘的是缺失属性之间频繁出现的规则，如“属性 A 缺失，属性 B 缺失 → 属性 C 缺失”。

关联规则是形如“$A \rightarrow B$”的蕴含式，对关联规则的挖掘实际上是对频繁项集的挖掘。规则 $A \rightarrow B$ 具有支持度 $support(A \rightarrow B)$，即支持规则的记录占整个数据集 D 的百分比，实质上也就是项集 A∪B 出现的概率 $P(A \cup B)$。规则 $(A \rightarrow B)$ 具有置信度 confidence(A→B),即包含 A 的记录中，同时包含 B 的记录所占的百分比，实质上就是条件概率 P（B|A）。即；

$$\text{support}(A \rightarrow B)=P(A \cup B)\quad(5)$$

$$\text{confidence}(A \rightarrow B)=P(B \mid A)=\frac{\text{support}(A \rightarrow B)}{\text{support}(A)}\quad(6)$$

支持度和置信度是对规则强度的一种度量，通过这两个指标保证了规则的可信

度。对于频繁项集的挖掘有许多方法，其中常用的是 Apriori[3] 算法，这是一种基于频繁项集性质先验知识的逐层搜索方法。本文研究的是关于目标属性的规则，通过使用关联规则类（Class Association Rule，CAR）方法[10]避免了无用规则的产生。

3.3 数据预处理

在正式进行规则挖掘之前要对数据进行预处理，主要包括两个步骤：抽样和特征选择。抽样和特征选择在不丢失总体数据集中用户感兴趣的特征的前提下，降低了规则挖掘的复杂度。另外由于挖掘算法的特殊要求和数据集的不规范（如缺失标识混乱），所以还要对数据进行一些转换。

3.3.1 属性转换

由于关联规则挖掘算法要求所有的属性都要是标称的，所以需要将数据集中的非标称属性转换为标称属性。另外由于数据集中的缺失值可能以多种形式存在，如“null”，“unknown”或者“not available”等等，为了便于进行规则挖掘，需要赋予所有缺失值一个统一的缺失标识。

3.3.2 抽样

在进行规则挖掘时，由于面对的数据集的规模往往十分巨大，将整个数据集作为挖掘样本是低效且没有必要的，而且如果不加约束的对整个数据集进行挖掘，有可能会产生很多无趣的规则（其中一种情况是与目标属性无关的规则）。通过进行合适的抽样，可以提高规则挖掘的效率，也能够过滤无用的规则[8] 。

例 3 一个极端的例子，如果过滤掉所有不符合条件的记录后只留下一条记录，挖掘算法产生以这条记录产生规则。这些规则可以通过支持度-置信度框架，但是其实可能只是一个偶然事件。

但是由于相对总体来看缺失的发生往往是稀有事件，使得数据集是一个不平衡数据集，也就是发生缺失的样本在总体中所占的比例过低。简单的随机抽样有可能会丢失许多目标属性发生缺失的样本，但这正是本文关注的样本。而如果只选取发生缺失的样本，又有可能发生例 3 的情况。为此，可以采用分层抽样的方法[2] 。

分层抽样又称分类抽样，其基本思想是按照某种特征将总体分为互不交叉的若干次级总体，然后在各次级总体中分别使用随机抽样来组成样本。Monica Chiarini Tremblay 在[1] 提出了一种针对缺失数据集的分层抽样策略：将总体样本按照目标属性是否缺失分为两个次级总体，分别从两个次级总体中随机抽样，使得来自目标属性出现缺失的次级总体的样本占最终样本一定的比例。例如，如果需要一个容量为 10000 的样本，要求目标属性出现缺失的样本占 50%，其他样本占 50%，那么就从目标属性缺失分组随机抽取 5000 个样本，从目标属性未缺失的分组中随机抽取 5000 个样本。

这样的分层抽样策略保证了提高缺失记录支持度，使得规则容易被发现；又保持了置信度不变，保证了发现的规则的客观性，减少了例 3 中的情况的发生。

3.3.3 特征选择

现代的数据分析任务面对的数据集往往具有较高的维度，而大部分的知识挖掘方法随着数据维度的增加，其性能往往会大幅度降低，这也就是维数灾难。而因为数据集中的有些属性同我们关注的目标属性相关性很弱，在数据分析时引入这些属性既降低了效率又有可能引起过拟合使分析结果失准。因此我们有必要通过特征选择筛选出同目标属性最为相关的属性，这样既提高了挖掘的效率又不会对分析结果产生太大的影响。

特征选择方法可以分为有监督和无监督的两种，因为缺失数据是有类标的（缺失属性和非缺失两类），所以应该使用有监督的特征选择方法并以目标属性作为类标。

3.4 关联规则挖掘实施

由于我们只关心同目标属性的缺失有关的规则，为了避免产生无用的规则我们使用关联规则类（CAR）方法[10] ，以目标属性作为类属性（Class Attribute）。

支持度和置信度的设置同结果规则的数量以及质量直接相关，如果设置的过低会产生大量的无用甚至虚假的规则，反之会漏掉一些可能十分重要的规则。对于支持度通常设置的比较低，比如 0.1；而置信度因为

关系到规则的可信度，可以设置的高一点，如 0.9。通过设置合适的支持度和置信度阈值，并选择目标属性为 CAR 方法的类属性，使用 Apriori 算法就可以产生我们需要的缺失依赖关联规则。

3.5 规则约减

关联规则挖掘经常会产生许多冗余的规则，特别是如果我们有多个目标属性，分别对多个目标属性进行关联规则挖掘会使总的规则集变的十分庞大。如果产生的规则集较大，为了减小结果规则集的规模以便数据分析者使用，可以通过规则聚合进行规则的约减。下面是一种常见的冗余情况，

A=Missing，B=Missing → C=Missing

A=Missing → C=Missing

以上两条规则是一种常见的冗余情况，由于第二条规则比第一条更加一般，所以可以约减到第一条规则：A=Missing → C=Missing。

在结果规则集较大的情况下，应该采用自动的约减方法对规则进行约减，如文献[3]中所述的方法。在本文中产生的规则集较小，我们只是采用上面这种启发式原则人工对规则集进行规约。

4 实验

本章中，我们将第三章中描述的方法应用到一个真实的数据集上，并对实验结果进行分析。实验使用一台 PC 机上进行，CPU 为 Inter corei5,内存为 4GB，硬盘容量 500GB，操作系统为 Windows7 64 位，实验工具主要使用 WEKA（Waikato Environment for Knowledge Analysis）3.6.10。

4.1 数据来源及描述

实验中所使用的数据集 diabetic_data 来源于文献[9] 。该数据集是从 Health Facts database (Cerner Corporation, Kansas City, MO)中抽取出的，共有 101767 条记录，50 个属性，其中 13 个属性存在缺失，属于任意缺失。表 1 是缺失属性的具体情况。

表 1 缺失属性

属性	缺失数量（条）	所占比重（‰）
race	2273	22.34
gender	3	0.03
weight	98569	968.58
admission_type_id	10396	102.16
discharge_disposition_id	4680	45.99
admission_source_id	7067	69.44
payer_code	40256	395.57
medical_specialty	49949	490.82
diag_1	21	0.21
diag_2	358	3.52
diag_3	1423	13.98
max_glu_serum	96420	947.46
A1Cresult	84748	832.77

4.2 数据预处理

4.2.1 数据转换

如表 1 所示，该数据集中的缺失情况比较复杂，共有 13 个属性出现缺失，属于任意缺失。属性的缺失存在多种类型，分为“？”，“Unknown/Invalid”，“NULL”，“Not Mapped”，“Not Available”，“NONE”等。同一种缺失还可能有不同的标识，如 admission_source_id 属性中的“9”和“15”都表示“Not Available”。这样给我们分析缺失依赖规则带来了许多不必要的麻烦。为此在进行挖掘之前先将所有的缺失用同一种标识“Missing”来表示。虽然这样做看起来会混淆不同种类的缺失，但是对于挖掘缺失的依赖规则来说具体的缺失类型并不重要，事实上我们不关心“NULL”和“Not Available”的区别。另外，数据集中的 encounter_id，patient_nbr 等很多属性是连续属性，我们使用 Weka 中的 NumericToNominal Filte 对这些属性进行离散化。

4.2.2 分层抽样

目标属性的选择是根据具体应用来确定的，这里我们选取 admission_source_id 作为目标属性。根据 admission_source_id 是否缺失（即 admission_source_id=Missing）将数据集分为两个次级样本，其中 admission_source_id 缺失的为总体 A，容量 7067；admission_source_id 未缺失的为样本

B，容量为 94699。为了尽量保留缺失信息，我们对样本 A 进行 100%的抽样，即从中随机抽取 7067 条数据。另外从总体 B 中抽取 7067 条数据，使得最终样本中目标属性缺失和不缺失的样本各占 50%。

4.2.3 特征选择

使用 Weka 中的 CfsSubsetEval 作为评价函数，BestFirst 作为搜索函数，admission_source_id 作为类标。特征选择结果表明 race，admission_type_id，payer_code，medical_specialty，max_glu_serum 五个属性同目标属性 admission_source_id 相关性最高。

4.3 实验结果

利用 Weka 关联规则挖掘中的(Class Association Rule)可以产生规则右侧只为目标属性 admission_type_id 的关联规则。使用 Apriori 算法，设置最小支持度为 0.1，最小置信度为 0.9，并设置 class 为 admission_source_id。最终获得 17 条关联规则，因为我们只关注缺失之间的关系，所以去除掉不属于缺失依赖关系（即规则左侧或右侧出现“*Missing*”之外的属性值）的规则后得到如表 2 所示 6 条规则。

表 2 结果规则

序号	关联规则	支持度	置信度
5	admission_type_id=Missing payer_code=Missing medical_specialty=Missing → admission_source_id=Missing	0.10	0.98
8	admission_type_id=Missing payer_code=Missing → admission_source_id=Missing	0.32	0.96
9	admission_type_id=Missing payer_code=Missing max_glu_serum=Missing → admission_source_id=Missing	0.19	0.95
11	admission_type_id=Missing → admission_source_id=Missing	0.42	0.95
15	admission_type_id=Missing medical_specialty=Missing → admission_source_id=Missing	0.15	0.93
17	admission_type_id=Missing max_glu_serum=Missing → admission_source_id=Missing	0.19	0.91

观察得到的 6 条规则可以发现这些规则有一个共同点：所有规则左侧都出现了 admission_type_id=Missing。如果对这六条关联规则进行规则约减，则约减为第 11 条：

$$admission\_type\_id = Missing \rightarrow admission\_source\_id = Missing$$

这说明 admission_type_id 缺失同 admission_source_id 出现缺失是强相关的。事实上，我们单从两个属性的名称上就可以推断出二者关系十分紧密，正如例 2 中的婚姻状况同配偶姓名一样。而特征选择中的 race 属性虽然从特征选择的角度来说同 admission_source_id 属性相关性较强，但是 race 出现缺失同 admission_source_id 出现缺失之间的关系并无很强的相关性。

通过以上的实验结果表明，本文中的方法可以有效的发现任意缺失中存在的缺失依赖关系。领域专家利用这些依赖关系，可以对数据分析或者改进数据质量提供帮助。

5 结论

数据缺失是一种常见的数据质量问题，

但是数据的缺失并不意味着其中信息的完全丢失。在非随机缺失中，属性的缺失一般不会单独发生，多个缺失属性之间可能存在着依赖关系。如果对这种缺失数据进行估计将会产生很大偏差，而直接删除缺失数据则会丢失缺失数据隐含的信息。

由于任意缺失的缺失情况十分复杂，本文并没有研究具体的缺失机制，而是使用数据挖掘的技术，尝试从发生任意缺失的数据集中发现缺失属性之间的依赖关系。

通过在真实数据集上的实验，可以发现这种缺失属性之间的关联并不罕见，而通过本文提出的方法可以较高效的发现这样的关联规则。

文章中使用到的抽样，特征选择，规则约减方法都还有待探讨。在下一步的研究中将对整个挖掘过程中的一些细节技术在进行研究，以进一步提高挖掘的效率和准确率。

参考文献

[1] Monica Chiarini Tremblay, Kaushik Dutta,Debra Vandermeer.Using data mining techniques to discover bias pattern in data missing[J].ACM Journal of Data and Information Quality, 2010,Vol. 2, No. 1, Article 2.

[2] Han Jiawei. Micheline Kamber.Jian Pei. 范明, 孟小峰译. 数据挖掘概念与技术(Data Mining Concepts and Techniques)[M]. 北京：机械工业出版社 2012.

[3] Tobias Scheffer. Finding Association Rules That Trade Support Optimally against Confidence[C]. In Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery. PKDD 2001,LNAI 2168,pp.424-435,2001.

[4] Roderick J.A Little, Donald B. Rubin. Statistical Analysis with Missing Data[M].John Wiley & Sons, Inc. New Jersey.2002.

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

[6] 孙婕 金勇进 戴明峰.关于数据缺失机制的检验方法探讨[J].数学的实践与认识 2013.43（12）166-173

[7] 孙晓松.数据缺失机制及其检验[D].江苏：苏州大学，2007.16-24.

[8] Sandy Moens, Bart Goethals.[C]Randomly Sampling Maximal Itemsets. IDEA'13, August 11th, 2013, Chicago, IL, USA.

[9] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore[J]. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.

[10] LI, J., SHEN, H., AND TOPOR, R. W. 2001[C].Mining Optimal Class Association Rule Set. In Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining. PAKDD 2001,LNAI 2035,pp.364-375,2001.