一种含缺失数据的记录逻辑检测方法

高 科 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：针对统计数据集中属性值缺失普遍存在的情况，为了更加准确地进行记录的逻辑性检查，在找出问题数据的基础上尽可能多地保留数据的有效信息，提出先利用 Fellegi-Holt 算法对规则进行推演，再进行数据逻辑检测的方法。分析了算法的基本原理，结合具体数据实例进行规则推演及检测分析，从召回率和准确率两个指标上进一步比较了采用不同方法对问题数据进行检测的效果。实验结果表明，此方法能发现隐含规则，进一步提高问题数据的检测效果。

关键词：属性缺失；Fellegi-Holt；规则推演；召回率；准确率

Record Logical Detection in Condition of Field Value Missing

Gao Ke, Diao Xing-chun, Cao Jian-jun

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

Abstract　Field value missing is common among statics data sets.To carry out the logical detection of records more exactly, keep as much useful information as possible while finding out the error data. Put forward a logical detection method based on rules after resolution by using Fellegi-Holt method. After discussing the principle of the Fellegi-Holt method, show the general steps of resolution. Judge from the recall and accuracy rate of different methods of error data detection which connecting with the concrete examples. The experiment result shows that such method can discover the potential rules and perform a better detection.

Key words　field value missing; resolution; recall; accuracy rate

1　引言

在各种实用的数据库中，属性值缺失的情况经常发生甚至是不可避免的，这使得信息系统在大多数情况下是不完备的。造成数据缺失的原因很多，主要包括：1）部分信息暂时无法获取；2）信息遗漏；3）有些对象特定属性不可用；4）实时性要求达不到等等[1]。

数据缺失作为问题数据的一个重要产生原因，是数据预处理过程中重点关注的问题之一。如何在消除实例层数据质量问题的前提下尽可能地保存数据的真实信息，盲目否定或删除所有含缺失的数据显然是不明智的，目前没有一种通用且高效的缺失数据处理方法。那么，如何准确找出含有属性值缺失数据中的问题数据呢？

本文主要考虑了具有简单逻辑关系的数据，在属性值缺失情况下，如何利用已知的约束规则，以及采用 Fellegi-Holt 方法求出规则闭集情况下分别对问题数据进行检测。并进一步从召回率和准确率方面比较了不同方法之间的优劣。

2　数据的逻辑检测方法概述

逻辑检测是指利用逻辑表达式等方法对数据所具有的逻辑关系进行检查，以进行数据质量在相关维度上的评估。

逻辑关系检测法是以统计指标体系中各个统计指标之间存在的包含、恒等以及相关等内在逻辑关系为判断标准，实现对统计数据的可信度的粗略检测[2]。

按照检测所依据的逻辑关系，该方法又可以分为比较逻辑检测法和相关逻辑检测法。

比较逻辑检测法多用于对统计数据中单个属性取值与某个常量或者某两个属性取值之间的简单逻辑关系。相关逻辑检测则包括多个逻辑规则之间的复杂关系。

本文考虑在领域相关的统计数据集中，存在属性值缺失情况下进行的相关逻辑检测。数据缺失是统计数据集中普遍存在的问题，而属性值之间往往存在着诸多约束关系[3]，有些属性值的缺失不会影响到数据的准确性和可用性，而与之关联的属性值

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：高科，男，1989 年生，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，博士后。

联系方式：18752004008，**E-mail：**18752004008@163.com。

却可能因此违背约束关系。合理利用这些约束关系，不仅能够更加准确地检测出数据中的问题，还能为缺失数据的预测提供更加准确的依据。

3 Fellegi-Holt 方法简介

1976 年，Fellegi 等人提出一个严格的形式化数学模型 Fellegi-Holt 方法[4]，对于具体的应用，根据领域知识定义显式约束规则[5]，然后根据数学方法求出规则闭集。对于每条记录，自动判断是否违反规则约束。Fellegi-Holt 方法主要进行逻辑规则的发现。

3.1 方法适用对象要求

- 数据以记录形式存在，取值范围为有限集；
- 各字段存在清晰的逻辑规则。

3.2 方法解决的两个方面问题

- 如何根据已知规则挖掘出所有潜在的规则；
- 对给定的问题记录，找出最小的错误字段集合。

3.3 方法涉及到的基本符号及概念

- R $(R_1, \cdots\cdots, R_N)$：记录 R 包含 N 个字段；
- A_j：第 j 属性值的值域 $(R_j \in A_j)$；
- A_j^e：不满足规则 e 的第 j 属性值的值域；
- $edit$ (e)：不满足某条规则 e 的记录取值空间；
- 最大产生集 MGE：

对于起始规则集 E 以及任一字段 i，产生新规则的方法 FHG 如下：

$$\mathrm{FHG}(i,E)=\prod_{j=1}^{i-1}\bigcap_{e\in E}A_j^e\times\bigcup_{e\in E}A_i^e\times\prod_{j=i+1}^{N}\bigcap_{e\in E}A_j^e$$

重复上述操作，直到不再有新的 *edits* 产生，此过程结束后产生的集合称为最大产生集 MGE(E)。式中 A_j^e 表示属性 j 的取值空间，因为域空间有限，所以此过程收敛。

4 数据表设计与分析

以某单位设备信息数据为例，为进一步简化计算，选取其中相关的部分字段进行分析。

表 2 设备数据表字段信息

序号	字段名称	值域	含义
1	设备编号	000-999	
2	类型	1-9	
3	使用状态	{0, 1, 2}	{未启用, 使用中, 已停用}
4	使用时间（天）	0-99	
5	维修次数	0-10	

系统在数据统计录入时并未对错误数据做任何校验处理，因此存在部分属性值缺失的记录。在后续的处理中，操作人员制定了以下规则（格式 A→B 表示：如果记录满足条件 A，那么可推出其满足条件 B）对数据进行了进一步的检测并找出其中的问题数据（这里的问题数据指不满足以下规则的记录）：

规则 1：“使用状态”=0 → “使用时间”=0

规则 2：“使用状态”>0 → “使用时间”>0

规则 3：“使用时间”<=3 → “维修次数”=0

规则 4：“维修次数”>0 → “使用状态”>0

即设备处于未启用状态其使用时间为 0；设备一经启用，其使用时间必然不为 0；设备至少使用三天进行第一次维护，所以若设备使用时间不足 3 天则其维修次数必然为 0；维修次数大于 0 其必然已经使用。

5 规则推演与数据分析

为了进一步发现上述规则中隐含的规则，采用 Fellegi-Holt 方法求解已知规则的最大产生集。

5.1 规则的符号形式化

分别用 $A_{station}$、A_{time}、A_{repair} 表示设备使用状态、使用时间、维修次数的取值，则上述规则可以用如下表达式表示：

e_1= $\{A_{station}: 0\}\times\{A_{time}: 1, \cdots\cdots, 99\}\times A_{repair}$

e_2 =$\{A_{station}: 1, 2\}\times\{A_{time}: 0\}\times A_{repair}$

e_3 = $A_{station}\times\{A_{time}: 0, 1, 2\}\times\{A_{repair}: 1, \cdots\cdots, 9\}$

e_4 =$\{A_{repair}: 1, \cdots\cdots, 10\}\times\{A_{station}: 0\}\times A_{time}$

5.2 新规则计算

根据 FHG 方法，利用已知规则集逐一字段进行新规则计算并生成最大产生集。经过计算，产生的新规则对应的 edits 如下：

e_5= FHG(A_{time}，{ e_1，e_3})=$\{A_{station}: 0\}\times A_{time}\times\{A_{repair}: 1, \cdots\cdots, 9\}$

e_6= FHG($A_{station}$，{ e_2， e_4})=$\{A_{repair}: 1, \cdots\cdots, 10\}\times\{A_{time}: 0\}\times A_{station}$

将上述符号表达式表示成对应的规则即：

规则 5：“使用状态”=0 → “维修次数”=0

规则 6：“维修次数”>0 → “使用时间”>0

即设备处于未启用状态其维修次数为 0；设备一旦被维修过，其使用时间必然不为 0。

5.3 数据分析

为了更加清楚地比较不同方法的检测效果，这里定义正常数据为各字段均符合规则的数据记录，问题数据则包括以下两类：①属性值之间不满足逻辑关系；②属性值缺失过多，不能提供有价值的数据信息。考虑使用状态、使用时间、维修次数这三个字段的不同取值情况，可以将数据类型分成以下三大类，如下表所示：

1） 正常数据

表 3 正常数据举例

序号	设备编号	类型	使用状态	使用时间（天）	维修次数
1	001	1	0	0	0
2	002	1	1	19	0
3	003	1	1	19	1
4	004	1	2	99	1
5	005	1	2	99	6

2） 不含缺失值的问题数据

表 4 不含缺失值的问题数据举例

序号	设备编号	类型	使用状态	使用时间（天）	维修次数	违反规则序号
1	001	1	0	***1***	0	1
2	002	1	2	***0***	6	2
8	003	1	0	0	***1***	3
4	004	1	***0***	1	1	4

3） 含缺失值的数据

表 5 含缺失值的数据举例

序号	设备编号	类型	使用状态	使用时间（天）	维修次数	违反规则序号
1	001	1	0	Null	***1***	5
2	002	1	Null	***0***	1	6
3	003	1	0	***1***	Null	1
4	004	1	Null	0	0	正常
5	005	1	***0***	Null	6	4
6	006	1	2	***0***	Null	2
7	007	1	Null	1	***2***	3
8	008	1	Null	Null	***2***	缺失过多
9	009	1	Null	Null	Null	缺失过多

对于含有属性值缺失的数据集，系统能够检测出所有包含缺失数据的记录。这些记录中，通过进一步检测，可以发现以下几种典型的特征记录，例如：

- 记录 3：“003，1，0，1，Null”。“使用状态”字段取值为“0”表示设备尚未启用，“使用时间”字段取值为“1”，违反了自定义规则 1：使用状态=0→使用时间=0。
- 记录 1：“001，1，0，Null，1”。“使用状态”字段取值为“0”表示设备尚未启用，“维修次数”字段取值为“1”，违反了新规则 5：使用状态=0→维修次数=0。
- 记录 4：“004，1，Null，3，1”。“使用状态”字段取值为空，其他字段取值未检测出不满足规则的情况，检测结果为正常数据。
- 记录 8：“009，1，Null，Null，2”。多个字段取值缺失，不能提供有价值的数据信息，检测结果为问题数据。

6 实验方法及结果分析

对于以上数据集，对使用状态、使用时间、维修次数三个字段的取值进行错误检测。分别采用以下三种检测依据：

1） 含空即错：只要记录属性值中包含空值即检测为问题数据；

2） 既定规则：使用预先定义好的四条规则对字段取值进行检测，不满足任一规则即检测为问题数据（规则涉及属性值缺失时，不进行检测）；

3） MGE：对既定规则采用 Fellegi-Holt 方法计算出规则的最大产生集，然后对字段取值进行检测，不满足任一规则即检测为问题数据（规则涉及属性值缺失时，不进行检测）。

对于检测结果，用召回率和准确率两个指标进行方法评价，属性值之间不满足逻辑关系的记录视为问题数据，具体计算方法如下：

- 召回率=A/(A+C)；
- 准确率=A/(A+B)；

其中各字母含义如下：

A：检测到的错误记录数；

B：检测到的非错误记录数；

C：未检测到的错误记录数。

以某单位设备信息统计数据表中的 850 条记录作为检测对象，构造实验所需的数据使得其中共有 100 条记录单个属性含有缺失值，25 条记录至少两个属性含有缺失值（此种类型记录因为不能提供任何有价值的信息，视为错误记录）。实验中系统能够检测出所有含缺失值的记录，这些记录中有 50 条记录属性值之间不满足逻辑关系。

在所有记录中，共有 200 条记录属性值之间取值不符合逻辑关系。其中，有 165 条问题记录可以根据既定的规则检测出。其余 35 条需要利用推演出的规则检出。具体实验结果记录如下表所示：

表 6 实验结果数据记录

依据	总记录数	总错误数	检出错误	实际错误	召回率	准确率
含空即错	850	250	125	75	30%	60%
既定规则	850	250	165	165	66%	100%
MGE	850	250	200	200	80%	100%

为了更加直观地感受不同方法在召回率和准确率指标上的性能差异性，作出如下统计图：

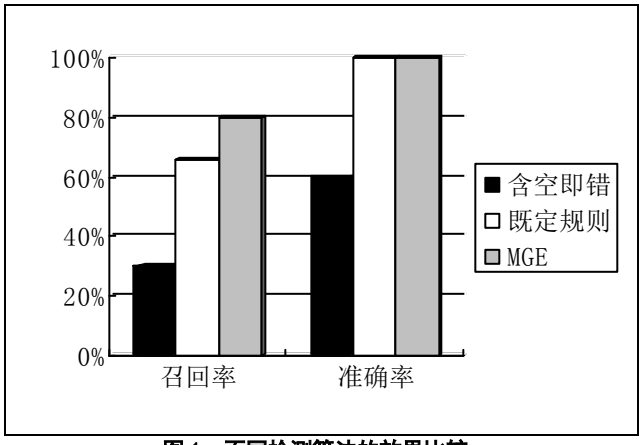

图 1　不同检测算法的效果比较

通过观察不难发现，盲目地排除含缺失属性值的数据其召回率和准确率都非常低下，极大地造成了原始数据信息的丢失。

利用 Fellegi-Holt 方法可以根据已知规则推演出所有的隐含规则，利用规则最大产生集能检测出所有不满足逻辑规则的问题数据，从召回率指标上来看，其检测结果优于仅仅利用既定规则进行的问题数据检测。但是对于缺失值过多的问题数据无法检出，因此准确率方面有所不足。如果将 MGE 方法结合空值检测，则能够进一步提高检测结果的召回率。

同时，利用得到的规则最大产生集还可以进一步为缺失数据的填充以及问题数据的纠正提供依据。例如，在{“使用状态”,“使用时间”,“维修次数”}取值为{0， Null，0}的记录中，根据规则“使用状态=0→使用时间=0”推断出使用时间取值为 0 来进行填充。

7　结束语

随着大数据时代的来临，种类繁多的数据必将面临更多的质量问题[6]，有关数据的缺失情况也将会更加普遍。如何从这些数据中尽可能多地获取我们所需要的有价值信息也越来越具有挑战性。当前，Fellegi-Holt 方法已经被广泛应用于众多政府企业的系统软件中，针对性地处理具有特定属性及逻辑关系的领域相关数据的质量问题[7]。

Fellegi-Holt 方法的有效性已经得到充分验证，如何进一步扩展该方法的适用范围，可以尝试将其进一步抽象用于领域无关的数据质量检查工具中，由用户根据待检测数据记录的属性值特征，个性化地对检查规则集进行自定义。本文所做的部分工作，可以作为上述扩展的简单尝试，为后续深入研究提供依据。

参考文献：

[1] (美)阿利森. 缺失数据[M]. 格致出版社, 2012.

[2] 许涤龙, 叶少波. 统计数据质量评估方法研究述评[J]. 计算机科学, 2004, 10:155-157.

[3] 梁吉胜, 李天阳, 王慧霞, 等. 基于约束的数据质量评估算法研究[J]. 统计与信息论坛, 2011, 7: 3-14.

[4] Agnes Boskovitz, Rajeev Gore, Markus Hegland. A Logical Formalization of the Fellegi-Holt Method of Data Cleaning[C]. Canberra ACT0200 Australia.

[5] 冀英伟. 基于谓词逻辑和包含集的分类规则约简方法[J]. 太原科技大学硕士学位论文, 2009.

[6] 曹建军, 刁兴春, 等. 大数据与大数据质量问题[J]. 现代军事通信, 2012, 4:45-48.

[7] William E. Winkler, Bor-Chung Chen. Extending the Fellegi-Holt Model of Statistical Data Editing[C]. Statistics2002 Washington D.C.