一种不确定 RFID 数据流清洗策略

刘云恒[1] 刘耀宗[2] 张 宏[2]
(南京森林警察学院信息系 南京 210023)[1] (南京理工大学计算机学院 南京 210094)[2]

摘 要 原始 RFID 数据流含有大量噪声且具有不确定性,必须在使用之前对其进行数据清洗,而清洗策略是清洗质量的保证。提出一种适合不确定 RFID 数据流的清洗策略。该清洗策略引入了最大熵原理,对待清洗的 RFID 元组的特征属性进行权重选择,并根据清洗节点的时间消耗以及误差进行清洗成本分析,决策出最佳的清洗方法。仿真实验结果表明,该清洗策略提高了不确定 RFID 数据流的清洗效率与精度。

关键词 RFID 数据流,不确定性,清洗策略,清洗成本,最大熵特征选择
中图法分类号 TP181 **文献标识码** A

Uncertain RFID Data Stream Cleaning Strategy

LIU Yun-heng[1] LIU Yao-zong[2] ZHANG Hong[2]
(Department of Information, Nanjing Forest Police College, Nanjing 210023, China)[1]
(School of Computer, Nanjing University of Science and Technology, Nanjing 210094, China)[2]

Abstract The original RFID data stream contains a lot of noise and uncertainty, so the data must be cleaned before using and the cleaning strategy is the guarantee of the quality of the cleaning. In this paper, a new method for cleaning the RFID data stream was proposed. The maximum entropy principle is introduced in the cleaning strategy, and this treat cleaning RFID tuple attributes to select weights. the cleaning cost analysis is performed according to the cleaning node time-consuming and error to decide the best cleaning method. Simulation experiment results show that this cleaning strategy improves the cleaning efficiency and accuracy of the RFID data stream.

Keywords RFID data stream, Uncertainty, Cleaning strategy, Cleaning costs, Max-entropy feature selection

RFID 技术为人们提供了强大的感知、理解及管理物联网世界的能力,可广泛地用于识别、定位、跟踪、监控物联网中的物理对象[1]。在 RFID 系统中,射频信号经天线发射出去,读写器接收信息并进行处理以及做出相应反应,在信号传递的过程中不可避免地会引入噪声,这些噪声大大影响了 RFID 数据处理的准确性[1]。

由于 RFID 数据本身的不确定性,在进行正式的数据挖掘建模、复杂事件的探测前需要对数据进行预处理,这种预处理主要基于 RFID 数据清洗相关技术。数据预处理占据数据挖掘过程中的大部分时间,目前对数据流进行预处理的研究重点是预处理与数据挖掘相结合[2],必须针对 RFID 数据流的特点作相对应的清洗。RFID 数据受各种因素影响,具有不确定性(Uncertainty),目前工业界和学术界将 RFID 视为不确定数据流(Uncertainty Data Stream),需要研究更适应其特点的清洗方案。

近几年来,众多研究机构和工业界对 RFID 数据清洗方法进行了广泛而深入地研究,并取得了一定的成果[3-5]。这些研究主要侧重于清洗方法,并不注重清洗效率,而清洗策略主要研究的问题是根据 RFID 物理脏数据的类型而采取不同的清洗方法,在保证清洗效果的前提下,提高 RFID 数据清洗的效率。

有的实际 RFID 应用的规模相当巨大,如 RFID 图书馆有数十万册图书,可以同时布署近千个的阅读器和数十万个标签。在对海量的 RFID 数据流进行在线清洗时,如果不充分考虑清洗的成本,那么在清洗过程中大量的时间与精力将消耗在数据预处理过程中,而无法及时对 RFID 数据作进一步的处理,所以清洗算法必须要考虑精度与效率的折衷问题。

目前已有的 RFID 清洗研究侧重于清洗方法,而在实际应用中更需要考虑清洗的效率问题,RFID 清洗策略是 RFID 应用成功的重要保障,可以根据实际 RFID 应用的情况采取相应的清洗策略,从具体应用出发的清洗策略应该与应用直接相关,从而更加符合 RFID 的应用实际需求。

1 RFID 数据流清洗策略研究进展

1.1 RFID 数据流的清洗策略机制研究

文献[3]基于 RFID 应用提出一种综合性的数据清洗策略,该机制由局部过滤器和全局过滤器组成,局部过滤器处理单个阅读器接收的数据,通过时间延迟对 RFID 数据按时间戳进行排序,并根据 RFID 数据流的分布情况设置不同的约束从而删除多读数据;并通过全局过滤器处理多个阅读器接收的数据,通过标签数据的时空关联性填补漏读数据和删除多读数据,并设定约束条件删除冗余数据,可以实现对各种类

本文受中央高校基本科研业务(LGYB201602)资助。
刘云恒(1975-),女,硕士,主要研究方向为大数据分析,E-mail:new025@foxmail.com;刘耀宗(1975-),男,博士,主要研究方向为 RFID 数据流管理与挖掘;张 宏(1956-),男,教授,博士生导师,主要研究方向为数据挖掘与信息安全。

型的脏数据的修正。

文献[4]通过对不确定 RFID 数据特征进行分析，建立了一套分流机制下的 RFID 数据清洗策略。该清洗策略引入清洗队列的概念，真实 RFID 数据流中 3 种脏数据的比例是不同的，可以根据清洗节点的判断条件选择最佳的清洗路线，无需遍历清洗系统中的所有清洗节点，从而节省了大量的数据传输和清洗等待时间。实验表明，该策略很好地缓解了数据传输压力，有效地提高了 RFID 数据清洗的效率。

1.2 基于机器学习的 RFID 数据流清洗策略

已有的 RFID 数据清洗算法主要考虑的衡量标准是精确性，即清洗后的数据中准确数据所占比例。但是当某个 RFID 应用布置规模极大时，如涉及到数千个阅读器和数万个标签，这时算法的衡量标准就不能只考虑数据的精确性了，还要考虑到算法的时间开销问题。

以机器学习为背景的清洗算法中就此问题给出了一种解决方法。2007 年 Hector Gonzalea 等人[5]提出了基于代价考虑的清洗方法，通过提出新的清洗规则来实现代价的最小化，此外，该文还介绍了基于动态 Bayesian Network 的清洗方法，通过阅读器历史的观测结果来估算标签下一次可能出现的概率。由于该算法会用到历史数据，因此历史数据的质量直接决定了估算结果的精确性。它首次提出了一个针对大规模的 RFID 数据集的清洗框架和一系列的 RFID 数据清洗策略，并分析了各个策略对应的清洗开销，由此引出了一个适应性调整清洗成本开销策略的精确性总体优化算法。开销主要包括 3 个部分：1）机器学习中，每个元组的训练开销；2）存储开销和运行开销；3）分类错误时修改所需的开销。

RFID 数据清洗也可以看成是一个分类（classification）问题[5]。可以将 RFID 数据元组组成的数据流进行在线分类，RFID 数据的建模形式为 D（EPC，Reader，timestamp，location，detected，other），其中，EPC 和 Reader 分别指标签和 Reader 阅读器的唯一编码，location 指阅读器探测到标签时标签所处位置，detected 表示是否被阅读器检测到，other 包含一些相关信息，如标签所附物体特征、地理条件、标签协议等。这些信息可以被当作数据训练集，通过机器学习，总结出相关规则，用于选择最优的清洗方法。在不同的应用背景和环境下，所携带的特征信息可能会不同，这主要跟 RFID 具体应用类型有关。

定义 1（基于机器学习的 RFID 清洗策略） 基于机器学习的 RFID 数据清洗模型根据不同的 RFID 数据流块的特征进行最优清洗策略选择，从而降低清洗成本，提高清洗效率，实现总体开销最优化。清洗方法是将待清洗的 RFID 数据流块按标签实例（$\langle Tag,t\rangle$，$\langle f_1,\cdots,f_i,\cdots,f_k\rangle$）作为输入的一个方法分类器 C，f_i 是指用来描述标签的各个属性或所在的环境的特征。

RFID 数据的特征属性主要分为 4 种情况（特征数目并不固定）：

1）标签特征（Tag Features）：描述标签属性，如通信协议、历史数据等；

2）阅读器特征（Reader Features）：描述阅读器属性，如天线数量、通信协议；

3）位置特征（Location Features）：描述标签被读出时所处的位置，如书架；

4）标签所附物体特征（Item Features）：如标签的物质材料（金属或塑料）。

以上这些特征都可以用来作为分类学习的标准，最终的选择还是需要从初始训练数据集中学习得到。特征选择指针对不同环境的数据特征进行最优清洗策略选择，从而实现总体效率最优。具体清洗方法由用户自己选择，可以采用传统的决策树或贝叶斯方法。基于机器学习机制的 RFID 清洗策略可以采用决策树分类算法或贝叶斯分类算法，根据不同数据特征进行最优清洗策略选择，从而达到整体开销最小、效率最高的原则。清洗规则可以具体指明清洗分类的条件，策略模型根据这些条件对每个流入的 RFID 数据元组进行分类并找到适合它的清洗方法，从而达到总开销最小。

文献[5]将布置 RFID 系统的环境特征、标签所附在的物体特征等作为策略分类的标准，由于 RFID 数据受环境影响较大，因此这些特征很大程度影响了 RFID 数据的准确性，但原始数据训练集如何选择及哪些特征属性才能有效覆盖 RFID 应用还有待进一步研究；其次，特征属性与清洗结果的相关性是什么及如何根据特征属性选择合适的清洗方法也需进一步研究，由于 RFID 数据具有不确定性，如何度量 RFID 特征的不确定性是影响 RFID 清洗效率与精度的重要因素。

本文利用信息熵（entropy）度量 RFID 原始数据的不确定性，对影响清洗效果的特征属性进行分析，引入最大熵原理，按特征属性的信息熵对待清洗的元组进行排序，并采用决策树分类算法选择合适的清洗方案。对比实验证明该方法对海量不确定 RFID 数据清洗有比较高的效率。

2 基于 MEFS 的 RFID 数据流清洗策略

2.1 最大熵特征选择

RFID 数据流的数据不确定性可细分为元组级不确定性和属性级不确定性[3]。由于 RFID 不确定数据的类型不同，无法采用统一的清洗方法来实现数据的一次性清洗。因此，需要从实际应用的角度出发，根据不确定数据特点，实现对数据的分类清洗。RFID 数据流包含大量的无关信息和冗余信息，这些信息可能极大地降低分类算法的性能。挖掘属性之间的相关性，对数据进行特征选择，大量研究实践证明，特征选择能够有效地消除无关和冗余特征，提高分类任务的效率，改善预测精确性等学习性能。

信息熵（entropy）是衡量随机变量的信息含量的测度，将其物理意义应用至特征选择的范畴中能够帮助选出具有最高信息含量的特征，是一种比较好的全局特征测度手段。文献[6]基于信息熵理论，提出了一种高效的不确定性数据清理方法。文献[7]用最大熵（max entropy）度量数据流分类的优劣。文中总结出一种好的数据清洗方法应该让 RFID 系统尽量减少不确定性，即熵值尽量小，反过来说清洗策略应该优先选择清洗熵值较大的对象。采用熵作为不确性的一种度量，小的熵意味着变量更加确定，即熵值很小时，就可认为该变量是确定的，并可将返回概率最大对应值作为该不确定变量的值。

定义 2（最大熵原理，Principle of Maximum Entropy，POME） E. T. Jaynes[8]于 1957 年提出了最大熵估计原理，文中指出：在信息不足和概率空间不完备的情况下进行统计推断时，应充分利用现有信息，选择具有熵最大的那一种概率分布作为统计推断的结果。最大熵原理的基本思想是：给定训练样本，选择一个与训练样本一致的模型。最大熵模型应

选择与这些观察相一致的概率分布，而对于除此之外的情况，模型赋予均匀的概率分布。

定义 3（最大熵模型，Max-Entropy Model） 数据 X 的某个属性为 x，设 $x\in X$，c 是 x 的子串，c 对 $y\in Y$ 有表征作用，则称 (c,y) 为模型的一个特征。假设存在 n 个特征 $f_i(i=1,2,\cdots n)$，设 $p(f)$ 为特征 f 对于模型确实的概率 $p(x,y)$ 的数学期望，而 $p(x,y)=p(x)p(y|x)$，条件概率 $p(y|x)$ 均匀下的具有最大熵的模型定义为：

$$H(p)=-\sum_{x,y}p(x)p(y|x)\log nP(y|x) \tag{1}$$

在允许的概率分布 C 中选择模型，具有最大熵：

$$p^*=\arg\max_{p\in C}H(p) \tag{2}$$

定义 4（最大熵特征选择，Max-Entropy Feature Selection，MEFS） 设特征选择的分类属性构成随机过程，所有输出值为 Y，对于每个值 $y\in Y$，已知与 Y 相关的所有决策属性值组成的集合为 X，给定的所有属性 $x\in X$，计算输出为 $y\in Y$ 的条件概率，即对 $p(y|x)$ 进行估计，特征选择的目标就是从所有决策属性中选择出对分类属性最具有表征作用的属性。

根据最大熵特征选择的方法，将待清洗的 RFID 数据块按属性划分为优特征集数据和劣特征集数据，再按顺序流入清洗节点寻找适合的清洗方法，可以提高清洗效率并且减少错误，特别适合海量的不确定 RFID 数据流清洗[9]。

文献[10]中给出了该定义的详细证明与推论，本文不作详细介绍。本文将最大熵原理引入到 RFID 数据清洗策略，针对文献[5]中方法的不足，提出了基于最大熵特征选择的 RFID 数据清洗策略。由于篇幅有限，对于清洗的判定条件及清洗队列采用的清洗算法，本文不再赘述，可以参考文献[5]提供的清洗方法。

2.2 相关定义与分析

RFID 数据清洗策略从本质上可以看成是分类问题[5]，本文引入数据流分类的概念，并采用基于最大熵特征选择的方法对待分类 RFID 数据流块进行优化，提高了分类效率。

定义 5（清洗规则，Cleaning Rule） 清洗规则可以具体指明分类条件，根据这些条件，每个元组实例会找到适合它的清洗方法，从而达到总开销最小。一种简单的建模清洗规则的方法是使用决策树。在此基础上，对需要清洗的元组实例进行最优策略选择即可。元组实例一般不能直接获得，直接获得的是一个三元组，再根据已知的存储信息即可得到元组实例。

定义 6（清洗队列，Cleaning Sequence） 清洗队列为进入清洗节点前在缓冲区等待被清洗的 RFID 数据流的元组。在本文提出的清洗策略中，清洗队列中数据根据最大熵分类法则分为优特征集数据和劣特征集数据。

定义 7（清洗计划，Cleaning Plan） 在待清洗队列中，流入的数据块元组为 $\{D_1,\cdots,D_i,\cdots,D_n\}$，清洗方法的集合为 $M=\{M_1,\cdots,M_j,\cdots M_m\}$，每个数据块及采取的清洗方案形成的清洗队列为 $S_{D,M}=S_1,\cdots,S_n=Ms1\rightarrow Ms2\rightarrow\cdots\rightarrow Msk$，清洗队列按数据块对应的清洗方法形成的顺序，决策树分类的任务就是将按数据块的特征选择合适的清洗方法，从而达到开销最小的目的。

定义 8（优化清洗成本，Expected Cost Reduction） 数据集为 D，清洗方法为 M，数据集 D 使用特征 f 分割成子集 $D_1,\cdots,D_i,\cdots,D_{|f|}$，特征选择后的优化清洗成本如式(3)所示：

$$C(S_{D,M})=\sum_{i=1}^{|f|}\frac{|D_i|}{|D|}\times C(S_{D,M}) \tag{3}$$

$S_{D,M}$ 为数据集 D 采用清洗方法 M 的清洗队列。

改进算法：引入清洗成本概念，重新计算清洗代价，对待清洗的元组的特征按信息熵大小排序。极大熵理论中熵的极大化使得各概率分量尽可能地均等。

定义 9（清洗代价，Cleaning Costs） 为了决策出最优的清洗计划，对每个等待清洗的 RFID 原始元组数据形成的清洗队列进行清洗成本的计算。清洗成本除包括前面介绍的开销外，还要考虑清洗方法分类的误差成本。计算公式如式(4)所示：

$$C(S)=\alpha\cdot t\cdot C(S_{D,M})+\beta\cdot E_{D,M} \tag{4}$$

其中，$C(S)$ 为清洗代价；$C(S_{D,M})$ 为每个数据块的清洗开销；$E_{D,M}$ 为分类错误代价；α 与 β 是权重系数，根据分类器的错误分类开销与每个数据流块清洗开销调整不同的权重；t 是所有的数据流块。文献[5]机械地考虑每个数据流块的清洗代价，没有充分考虑分类错误代价，这样的清洗成本计算只能处于理想状态。

2.3 最佳清洗策略的流程分析

定义 10（最优清洗策略） 最优清洗策略通过训练得出的分类器从 M 中选择出最佳的清洗方法，即在准确率相同的前提下使得清洗代价最小。

通过不同数据特征和决策树方法进行最优清洗策略选择，以达到总体开销最小化；为了度量待清洗的 RFID 数据中特征的重要程度，通过计算特征集合中各个特征与不同清洗方法的信息量，对特征的重要程度进行排序，重新构建特征子集。

2.4 基于 MEFS 的 RFID 流数据清洗过程

由于 RFID 原始数据受环境等因素的影响很大，通常具有不确定性，在保证清洗结果准确可靠的前提下，标签数据训练集如何选取及特征如何选择才能进一步提高清洗准确性还需要进一步研究，基于前文所述的基于最大熵的数据流特征选择算法，本节提出一种 RFID 数据流在线清洗策略方案，清洗策略框架如图 1 所示。

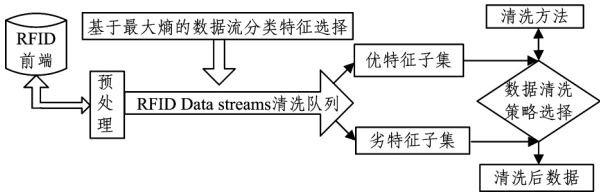


图 1 基于最大熵特征选择的 RFID 数据流清洗策略框架

基于最大熵特征选择机制(MEFS)的 RFID 数据流清洗策略的算法如下。

输入：$D_1,\cdots,D_n$ 流入的 RFID 数据元组；$M=\{M_1,\cdots,M_k\}$：表示清洗方法的集合；$C(M_1),\cdots,C(M_k)$ 表示对每个数据流元组中清洗的代价，$E_j(j=1\cdots k)$ 表示每个数据流元组误分类的代价

输出：根据不同特征选择不同清洗方法策略

步骤 1 RFID 前端捕捉到原始数据，将其送入预处理层，形成 RFID 数据流；

步骤 2 接收步骤 1 的流数据，采用文献[7]的算法，将待清洗的 RFID 数据流划分为两个特征子集(优特征子集与劣特征子集)；

步骤 3 对不同特征子集的 RFID 原始数据流按式(3)计算清洗代价；

步骤 4 采用 C4.5 分类器对进入清洗队列的 RFID 流数据进行清洗方案决策；

步骤 5 根据分类结果，针对不同 RFID 数据流中的元组数据采取不同的清洗方法；

步骤 6　上传清洗后的数据，根据式(4)计算清洗代价结果，调整分类器的工作策略。

分类器的工作策略调整主要包括 α 与 β 作为权重系数的调整以及清洗方法的调整，根据文献[7]提出的初始参数可以设为 $\alpha=0.8$，$\beta=0.2$。

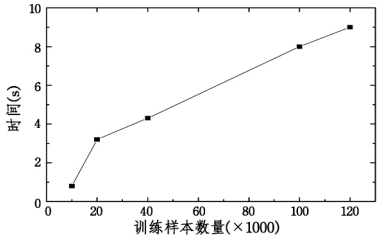
3　实验与分析

性能测试平台采用了文献[5]提供的 cleaning plan 数据集，并以数据流形式读取，所有实验均在 Intel CoreTM2 2.9 GHz、内存 8GB 的 PC 上完成，分类器采用 C4.5，系统环境为 cygwin1.50-1 和 gcc 3.4.4。

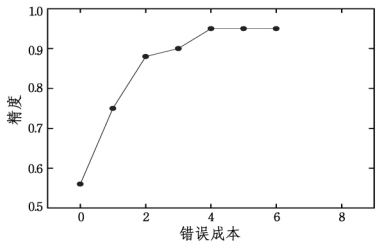
实验结果见表 1，其表明本文的基于 MEFS 的方法与文献[5]中的 ID3 决策树算法相比，时间复杂度更低。这是因为 ID3 采取信息增益的计算来度量节点的分裂。每进行一个分支，需要扫描一遍当前抽样集合的元组。基于 MEFS 的决策方法在待分裂之前对待分类的数据按熵值大小进行排序。只需进行一遍清洗方法的扫描便可将当前抽样元组含有的决策特征属性进行决策，从而可以映射到各个清洗方法所对应的特征属性，提高了决策效率。

表 1　清洗方法的决策时间和质量的比较

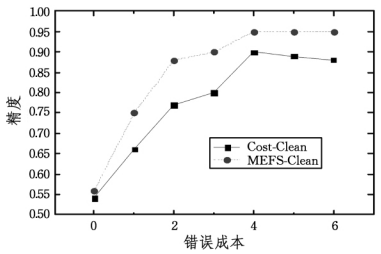
实验次数	决策时间(s)(ID3)	决策时间(s)(MEFS)	决策质量(%)(ID3)	决策质量(%)(MEFS)
1	798.2	465.2	89.2	92.4
2	789.1	456.7	87.6	92.1
3	768.6	453.4	88.2	91.9
4	777.2	466.3	87.9	91.7
5	776.8	466.2	88.8	92.3



(a)训练数据集与时间成本



(b)分类错误代价与精度



(c)两种清洗策略的清洗效果对比

图 2

为了测试该清洗策略的性能与效率，采用 RFID cleaning plan 数据集分别进行了训练数据集的时间成本与分类精度的测试，测试结果见图 2(a)与图 2(b)。由图 2(a)可以看出该算法的训练数据集样本与时间几乎呈线性关系，由图 2(b)可以看出该算法在分类误差达到允许范围的临界点后，分类的精确性维持在较高的水平(超过 0.9)。

采用了文献[5]提出的基于清洗成本的清洗策略(Cost-Clean)与本文提出的 MEFS-Clean 清洗策略在 RFID cleaning plan 数据集上进行清洗效果对比实验，结果如图 2(c)所示。由图可以看出 MEFS-Clean 清洗策略与 Cost-Clean 清洗策略在相同的清洗错误下，MEFS-Clean 的清洗精度要优于 Cost-Clean 清洗策略方法。

结束语　本文提出了一种基于最大熵特征选择的 RFID 数据流清洗策略模型，该模型采用了熵值度量 RFID 数据的不确定，引入了优和劣两种特征子集的清洗队列概念，有效地优化了不确定 RFID 数据流的清洗策略选择问题。与已有的 RFID 数据清洗策略进行性能对比实验，结果表明本文的策略具有良好的扩展性，在保证清洗成本的前提下大大提高了清洗决策的准确性，提高了不确定 RFID 数据清洗效率。

参 考 文 献

[1] Derakhshan R, Orlowska M E, Li Xue. RFID data management: challenges and opportunities [C]// Proceeding of IEEE International Conference on RFID. Dallas: IEEE Computer Society, 2007: 175-182

[2] 李战怀，聂艳明，等. RFID 数据管理的研究进展[J]. 中国计算机学会通讯，2007，2(3)：32-40

[3] 谷峪，于戈，等. 在于 RFID 应用的综合性数据清洗策略[J]. 东北大学学报(自然科学版)，2008，29(11)：1552-1555

[4] 夏秀峰，玄丽娟，李晓明. 分流机制下的 RFID 不确定数据清洗策略[J]. 计算机科学，38(10A)，2011：22-25

[5] Gonzalez H, Han J, Shen X. Cost-conscious cleaning of massive RFID data sets[C]// Proceedings of International Conference on Data Engineering(ICDE). Istanbul, Turkey, 2007: 1268-1272

[6] 覃远翔，段亮，岳昆. 基于信息熵的不确定性数据清理方法[J]. 计算机应用，2013，33(9)：2490-2492，2504

[7] Liu Yao-zong, Wang Yong-li, Wei Wei, et al. Feature Selection for Classifying Data Stream Based on Maximum Entropy[C]// Chinese Conference Pattern Recognition. 2009: 1-5

[8] Jaynes E T. Information theory and statistical mechanics[J]. Phys. Rev., 1957(106): 620-630

[9] 宋国杰，唐世渭，杨冬青，等. 基于最大熵原理的空间特征选择方法[J]. 软件学报. 2003，114(9)：1544-1550

[10] 刘华文. 基于信息熵的特征选择算法研究[D]. 长春：吉林大学，2010