An RFID Data Cleaning Strategy Based on Maximum Entropy Feature Selection

Yunheng LIU[1], Y.Z LIU[2*], H. ZHANG[2], T. LI[3]
[1]Information Technology Department
Nanjing Forest Police College, Nanjing, China
[2]School of Computer Science and Engineering
Nanjing University of Science and Technology, Nanjing, China
[3]School of Computer Science
Florida International University
11200 SW 8th Street Miami, FL 33199, U.S.A.
new025@126.com

Journal of Digital Information Management

ABSTRACT: *Data cleaning is an essential step of RFID data streams processing. Since the original RFID data have the characteristics of large scale, high speed, and uncertainty, the efficiency and the accuracy is crucial to the RFID data cleaning. By analyzing of the characteristics of RFID data streams, this article proposed an RFID data streams cleaning strategy based on maximum entropy feature selection. By introducing the MEFS to data cleaning, our strategy was able to classify the characteristics of RFID tuples, compute the cleaning costs by analyzing the time consumption and the error, and choose the best cleaning method. The simulation result showed that our strategy improved both the efficiency and accuracy of uncertainty RFID data streams cleaning.*

Categories and Subject Descriptors
H.2.8[Database Applications]: Data Mining; **[I.2.9 Robotics]:** Sensors

General Terms: Data mining, RFID Data Processing

Keywords: RFID data streams, Cleaning Strategy, Cleaning Costs, Maximum Entropy Feature Selection (MEFS)

Received: 19 December 2014, Revised 28 January 2015, Accepted 4 February 2015

1. Introduction

RFID technology provides stronger ability to perceive, understand and manage the world of the Internet of Things (IoT). RFID can collect data quickly in the form of non-line-of-sight, identify object accurately with a unique identifier, and can be widely used in applications such as identification, locating, tracking and monitoring the physical objects in the IoT [1]. In the RFID system, the RFID signal is launched by the antenna. The read/write devices would receive and process the information, and then respond correspondingly. The process of signal transmission would inevitably be mixed with noise, which could greatly affect the accuracy of the RFID data streams processing [2].

Since the RFID data streams are uncertain, the data need to be pre-processed before formal data mining modelling and complex events detection. The pre-process is mainly based on RFID data cleaning technology. Because the RFID data streams contain spatial-temporal information, today's industry and academia defined RFID data as spatial-temporal streams, of which the characteristic adaptive cleaning solution still needs further research.

In recent years, many researchers have carried out extensive and in-depth researches on the RFID data cleaning method, and obtained a certain results [1-3]. Those studies mainly focused on the cleaning method, but not on the efficiency. While the cleaning strategy research is about different cleaning methods based on different RFID physical dirty data types, in the meantime, it should aim to improve the efficiency of RFID data cleaning, with the premise of guaranteeing the cleaning effect.

The RFID system has considerable wide applications. For example, there are a library usually has thousands of books. Such a library can deploy hundreds of RFID readers and thousands of RFID tags. Without giving full

consideration to the costs, the online cleaning for such an amount of RFID data streams would consume a lot of time and energy in the process of data pre-processing, and it is hardly possible to process RFID data streams in time. Therefore, the trade-off of accuracy and efficiency must be considered in cleaning strategy. Most RFID data cleaning researches focus on the cleaning method[1-3], though the efficiency problem is greater in the practical use. RFID data streams cleaning strategy is the important guarantee of the success of the RFID application, and the strategy can be adjusted according to the actual RFID application to deal with different situations. Thus, the cleaning strategy should be more in line with the actual needs.

2. Related Works

2.1 RFID Data Streams Cleaning Strategy

Gu Yu et al. [4] proposed a comprehensive data cleansing strategy based on RFID application. This mechanism is composed of local and global filters. The local filter processes the received data for single reader, sorts RFID data according to the timestamp, and sets different constraints to delete the redundant received data according to the distribution of RFID data streams. The global filter handles the received data for multiple readers, fills the missing data and deletes the redundant data based on the spatial-temporal relationship of tags data, then sets the constraints to delete the redundant received data, which can realize the correction for all kinds of dirty data.

XIA Xiu-feng et al. [5] proposed a triage mechanism of RFID data cleaning strategy through the analysis of uncertain characteristics of RFID data. The cleaning strategy adopted the concept of cleaning queue. While the proportions of three kinds of dirty data are different in the real RFID data streams, the best cleaning route can be chosen by the judgment of the conditions of cleaning nodes, without the need of traversing all cleaning nodes in the system, which can save a lot of time for data transmission and cleaning waiting. Experiments showed that the strategy alleviated the pressure of data transmission, and improved the efficiency of RFID data cleaning.

3.2 The RFID Data Streams Cleaning Strategies Based on Machine Learning

The existing RFID data streams cleaning algorithm is mainly measured by accuracy, which is the proportion of accurate data from the data after cleaning. But for the RFID applications which have large layout scale, such as applications having large-scale readers and tags, the measurement of algorithm can consider not only the accuracy of the data, but also the time cost of the algorithm.

The RFID data cleaning algorithm based on machine learning proposed a solution for the issue above. Hector Gonzalea et al. [6] proposed a cleaning method considering the cost, and trying to achieve cost minimization by putting forward new cleaning rules. Moreover, they also proposed the cleaning method based on dynamic Bayesian Network to estimate the probability of the next possible tags through the history of the reader observations. Since the algorithm used the historical data, the quality of the historical data would directly determine the accuracy of estimate results. It firstly proposed a cleaning framework for large-scale RFID data and an RFID data cleaning strategies, also analyzed the cleaning cost for various corresponding strategies, then proposed accuracy optimized algorithm which can adjust cleaning overhead cost strategy. The cost includes three parts, the training expense of each tuple in the machine learning, the storage cost and operation cost, and the modification cost during the error classification.

RFID data streams cleaning can also be regarded as a classification problem [6]. The data streams formed by RFID data tuples can be classified online. The RFID data model is defined as *D (EPC, Reader, timestamp, location, detected, other)*, where the *EPC* and *Reader* refer to the unique code for tags and readers respectively; the *location* is the location of tag when the tag be detected by the reader; the *detected* indicate whether the tag be detected; the *other* contains some other relevant information, such as the object characteristics, geographic conditions and the label agreement on the tag. That information can be used as training data set, summed up the relevant rules by machine learning and be used to select the optimal cleaning method. The characteristics information could be different in different application backgrounds and environments, which is mainly related to RFID application types.

Definition 1. RFID Data Cleaning Strategy Based on Machine Learning

The RFID data cleaning model is based on machine learning according to the different features of RFID data streams block to optimize the choice of cleaning strategy. This could reduce the cleaning cost, improve the efficiency of cleaning and achieve the cost optimization. The cleaning method uses the RFID data blocks which need to be cleaned, in a tag instance form ($<Tag, t>, <f_1,\ldots, f_i,\ldots, f_k>$) as a method classifier *C*. The f_i refers to the attributes used to describe the labels in the environment.

There are four kinds of features for RFID data (number of features is not fixed):

1) **Tag Features:** Describe the tag attributes, such as communication protocols and historical data;

2) **Reader Features:** Describe the reader attributes, such as the antenna number, communication protocols;

3) **Location Features:** Describe the tags position when they been read;

4) **Item Features:** Describe the tags material (such as metal or plastic).

All these features can be used as standards for classification learning, while the final selection still needs

to be learned from the initial training data set. Feature selection process affects the optimal cleaning strategy selection according to the data characteristics in different environments, to achieve the optimal overall efficiency. The specific cleaning methods can be chosen by the users by using the traditional decision tree or Bayesian method.

The RFID data cleaning strategy based on machine learning could uses the decision tree classification algorithm and Bayesian classification algorithm on the different data to process the optimal cleaning strategy selection, so as to achieve the minimum overall cost and the highest efficiency. Cleaning rules can specify the classification conditions, then according to these conditions, the strategy model could classify each influent RFID data tuples and find its cleaning method, thus to achieve the cleaning strategy with the minimum costs.

Hector Gonzalea et al. [6] chose the environment features of RFID system, the tags attached items features as the criteria for the classification of strategy. Since the RFID data are greatly influenced by the environment, these features influence the veracity of RFID data. However, how to choose the raw data training set, and which features can effectively cover the RFID application still need further research, followed by what is the relevance between the features and the cleaning result, and how to select the appropriate cleaning method according to the features. Since the RFID data streams are uncertain, how to measure the uncertainty of the RFID features is a major problem. Those problems are the important factors that could affect RFID cleaning efficiency and the accuracy.

Our study used the information entropy to measure the uncertainty of the RFID data streams, and analyzed the features that could influence cleaning effect. By introducing the maximum entropy principle, we sorted cleaning tuples according to the features, and used the decision tree classification algorithm to choose the appropriate cleaning solution. Experiments proved that our method could achieve high efficiency when cleaning massive uncertain RFID data.

3. The RFID Data Streams Feature Selection Based on Maximum Entropy

3.1 Information Entropy and the Uncertainty of RFID Data Streams

The uncertainty of RFID data streams can be subdivided into a tuple level uncertainty and feature level uncertainty [4]. The tuple level uncertainty describes the presence or absence of a tuple, and the feature level uncertainty does not involve the uncertainty of entire tuples. The query results returned in the form of data streams would be created with the same data but different probability value, which lead to duplicate calculations. Therefore, the dependence of uncertainty tuples on features must be taken into consideration to increase the corresponding weights of features and to clean data stream tuples with larger uncertainty preferentially.

Since there are different types of uncertain RFID data streams , it is not possible to one-time clean the data with a universal cleaning method. From the perspective of practical application, we need to classify cleaning the data according to the features of uncertain data. The RFID data streams would contain a large number of irrelevant information and redundant information, those information can greatly reduce the performance of classification algorithm. By mining the correlation between features, and processing the data feature selection, a lot of research have proved that the feature selection could eliminate the irrelevant and redundant features effectively, improve the efficiency of classification task, and improve the prediction accuracy.

Information entropy is a measurement of the information contents of a random variable. The application of its physical significance to the category of feature selection could help choose the feature with highest information content, which is a better measurement for global features. QIN Yuanxiang et al. [7] proposed an efficient method for uncertain data cleaning based on information entropy theory. LIU Yaozong et al. [8] measured data streams classification with maximum entropy, and proposed a data cleaning method that could reduce uncertainty for the RFID system, in other words reduce the entropy. On the other hand, the cleaning strategy should choose objects with larger entropy values. Entropy is used as a measurement for the uncertainty. Smaller entropy leads to larger certainty. When the entropy value is very small, we can think this variable is certain, and could return the corresponding value with the highest probability as the uncertain variable value.

Definition 2. The Maximum Entropy Principle E.T.Jaynes [9] proposed the principle of maximum entropy (POME) in 1957, their research pointed out that the statistical inference under the condition of the information shortage and incomplete probability space, should make full use of existing information, choose the entropy with the largest probability distribution, as the result of statistical inference. The basic idea of the maximum entropy principle is that given the training sample, choose a model consistent with training sample, and the maximum entropy model should choose the probability distribution consistent with these observations, while for other cases, the model chooses the uniform probability distribution.

Definition 3. The Maximum Entropy Model
Let a feature of data X be $x, x \in X$. Let c be the substring of x, if the c has the characterization for $y \in Y$, then the (c, y) is a feature of model. Assume there are n features f_i $(i = 1, 2, ...n)$, and $p(x, y)$ is the mathematical expectation for probability $p(f)$ of the feature f to the model, and $p = (x, y) = p(x)\, p(y|x)$, there exist a maximum entropy model under the uniform of conditional probability $p(y \mid x)$. We define the model as:

$$H(p) = -\sum_{x,y} p(x)\, p(y \mid x)\, log\, n\, P(y \mid x) \qquad (1)$$

The model is allowed to be chosen from the probability distribution, and have the largest entropy:

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

Definition 4. Maximum Entropy Feature Selection. (MEFS)
Assume the classification of features selection forms a random process, all the output values is Y. For each value $y \in Y$, we know the collection of all the decision feature value is related to Y as the set X, giving all the features $x \in X$, calculate conditional probability of the output which is $y \in Y$, namely to estimate the $p(y|x)$. The target of feature selection is to choose the decision features with the most characterization from all the classification properties[10].

Feature selection based on maximum entropy method would divide the cleaning RFID data blocks into rich feature set data and poor feature set data according to the features, then the sequential inflow into cleaning nodes is processed to find a suitable cleaning method, which could improve the cleaning efficiency and reduce error, and is particularly suitable for the cleaning of massive uncertainty RFID data streams.

Since LIU Yaozong et al. [8] proposed the detailed proof and inference for this definition, we do not include redundant introduction here. This article will introduce the maximum entropy principle to the cleaning strategy for RFID data. Aiming at the shortcomings of the previous methods, we propose a RFID data cleaning strategy based on the maximum entropy feature selection. Due to space constraints, the decision condition for cleaning and algorithm for cleaning queue are not involved in this paper. Refer to previously published cleaning method for details.

3.2 Cleaning Strategy Based on Maximum Entropy Feature Selection

Hector Gonzalea et al. [6] treated the RFID data cleaning strategy as a classification problem. While in this paper we introduce the concept of data stream classification, and use the feature selection method based on the maximum entropy method to optimize the RFID data blocks, which improve the efficiency of classification.

Definition 5. Cleaning Rules
The cleaning rules can specify the classification conditions. According to these conditions, each tuple instance will find its suitable cleaning method, thus achieve a cleaning strategy with the minimum total costs. A simple method to model the cleaning rules is to use the decision tree. On this basis, the optimal strategy is chosen for the tuple instance that needs to be cleaned. Tuples instance generally cannot be directly received. While the directly received object is a triad, it is modified according to the known information, and then the tuples could achieve.

Definition 6. Cleaning Sequence
The cleaning sequence is RFID data streams tuples waiting in the buffer to be cleaned before the tuples enter the node. In the cleaning strategy proposed in this paper, the data in the cleaning sequence can be divided into rich feature set data and poor feature set data by using the maximum entropy classification rules.

Definition 7. Cleaning Plan. In the cleaning sequence, the inflow data block tuple is $\{D_1,\ldots,D_i,\ldots,D_n\}$, and the cleaning method set is $M = \{M_1,\ldots, M_j,\ldots M_m\}$. Each data block and cleaning solution form the cleaning sequence $S_{D,M} = S_1,\ldots,S_n = M_{s1} \rightarrow \ldots \rightarrow M_{sk},$. The cleaning sequence is sorted according to the cleaning method which corresponds to data block. The task of the decision tree classification is to choose the appropriate cleaning method according to the features of the data block, so as to achieve the aim of minimum costs.

Definition 8. Expected Cost Reduction

The data set is D, cleaning method is M. By using the feature f, data set D can be divided into subsets $D_1,\ldots, D_i,\ldots, D_{|f|}$, and the optimization of cleaning costs after feature selection is :

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

where the $S_{D,M}$ is the cleaning queue with cleaning method M for data set D.

Improved Algorithm: Introducing the concept of cleaning costs, recalculating the cleaning cost, and sorting features of the cleaning tuples by information entropy. The entropy maximization could equal the probability of each component as far as possible.

Definition 9. Cleaning Costs
In order to determine the optimal cleaning plan, we need to calculate the cost for each cleaning sequence formed by RFID data that is waiting for cleaning. The cleaning costs include all the costs described above. We also need to consider the error costs of the cleaning method classification.

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

where the $C(S)$ is the total cleaning costs; $C(S_{D,M})$ is the cleaning cost for each data block; $E_{D,M}$ is the classification error costs; α and β are the weight coefficients which could be adjusted according to the misclassification costs and each data streams piece cleaning cost; t is for all the data stream blocks. While Hector Gonzalea et al. [6] just considered the cleaning costs for each data streams blocks, the misclassification costs which are only applicable to the ideal state have to be considered in the proposed method.

3.3 The Analysis Process for the Best Cleaning Strategy

Definition 10. The Optimal Cleaning Strategy

The optimal cleaning strategy is the best cleaning method selected from M by using the classifier obtained through training, with the premise of the same accuracy, which makes the cleaning costs minimum.

4. The RFID Data Cleaning Model Based on the Maximum Entropy Feature Selection

Since the RFID data streams is affected by the environmental factors, it is usually of high uncertainty. With the premise of guaranteeing the accuracy and reliability of cleaning results, further studies are still needed for selecting tag training data set and choosing the features to further improve the accuracy of the cleaning. By using the data stream feature selection algorithm based on maximum entropy, this section presents an online RFID data streams cleaning strategy plan. The process is shown in Figure 1.

The work process of RFID data streams cleaning strategy based on maximum entropy feature selection is as follows.

Inputs: $D_1,\ldots, D_n$ is the RFID data streams tuples; $M = \{M_1,\ldots, M_k\}$ is cleaning method set; $C(M_1),\ldots,C(M_k)$ is the costs for each data streams tuples cleaning; $E_J(j = 1\ldots k)$ is the misclassification costs for each data streams tuples.

Step 1: The RFID front-end captures the data streams, and uploads it to the pre-processing layer, then generates the RFID data streams;

Step 2: Receive the data streams from step 1 and, by using the algorithm in literature [8], divide the cleaning RFID data streams into two feature subsets (rich feature subset and poor feature subset);

Step 3: Calculate the cleaning costs for RFID data streams of different feature subsets according to equation (3);

Step 4: Decide the cleaning solution using C4.5 classifier for the RFID data which inflow the cleaning sequence;

Step 5: According to the classification result, take different cleaning methods for different tuples data in the RFID data streams;

Step 6: Upload the cleaned data, adjust work strategy of classifier according to the cost results from equation (4). The adjustment of classifier working strategy mainly includes the adjustment of weight coefficient α and β, and the adjustment of the cleaning methods. The initial parameters are set to $\alpha = 0.8, \beta = 0.2$.

5. Experimental Results

The performance test platform adopted the cleaning plan data sets from literature [6], and data were going throught the form of data streams. All experiments were carried out on a desktop computer with common hardware configuration: Intel CoreTM2 2.9GHz CPU, 4GB RAM. The classifier was C4.5. The program was developed using GCC 3.4.4 under cygwin1.50-1.

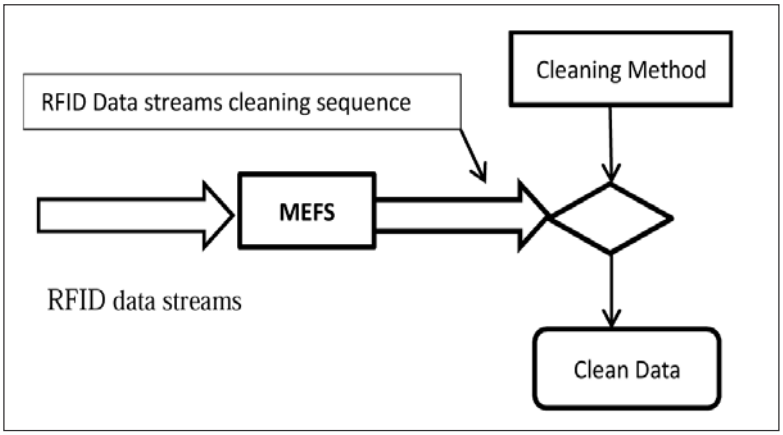

Figure 1. Mechanism of RFID data streams cleaning strategy based on maximum entropy feature selection

	C4.5 Time (s)	MEFS Time (s)	C4.5 Quality	MEFS Quality
1	798	465	89.2%	92.4%
2	789	456	87.6%	92.1%
3	768	453	88.2%	91.9%
4	787	467	87.8%	92.7%
5	778	446	88.4%	92.7%

Table 1. Decision time vs. decision quality for cleaning methods

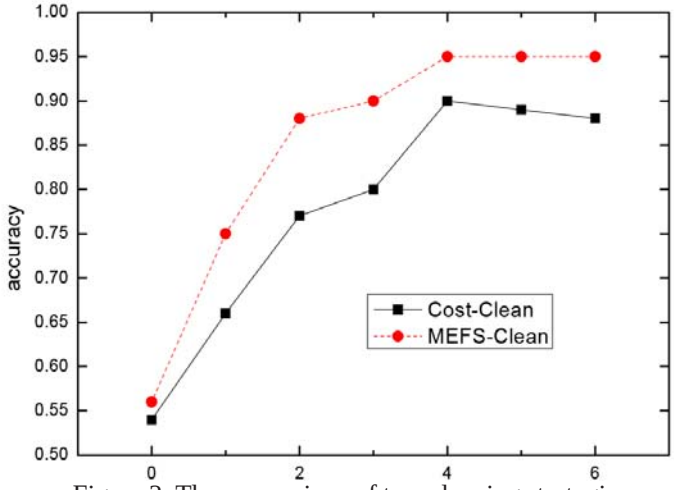

Figure 2. The comparison of two cleaning strategies

The experiments results showed that, our RFID data streams cleaning strategy method based on MEFS had lower time complexity compared to the algorithms that adopted ID3 decision tree in the literature [6]. This is because the process of determination of decision tree is to compare the current classification features. When processing each branch of the tree, it is also needed to scan current sample set again, and the scanning times is the depth of current path. However, the method based on MEFS just needs to scan once, and maps the decision features of current sampled tuples to corresponding features of each cleaning method.

We compared the experiment results of our **MEFS-Clean** strategy with the **Cost-Clean** method from literature [6], and plottd the results in figure 2. The comparison experiment showed that our MEFS-Clean strategy improved the cleaning quality effectively and reduced the error rate of the cleaning.

6. Conclusions and Future Work

We proposed a novel RFID data streams cleaning strategy based on the maximum entropy feature selection. The model uses entropy to measure the uncertainty of RFID data streams. By introducing the concept of rich and poor feature subsets, the model can effectively optimize cleaning strategy selection problem for the RFID data streams. The results of our study showed that, compared to the existing RFID data cleaning strategy, our strategy had a better scalability, greatly improved the accuracy of the cleaning decision-making with the premise of guaranteeing the cleaning time costs, and enhanced the efficiency of the massive RFID data streams cleaning.

Acknowledgement

This work was supported by the Fundamental Research Funds for the Central Universities, Nanjing Forest Police College (No.LGYB201506).

References

[1] Derakhshan, R, etc.al. (2007). RFID data management: Challenges and opportunities. Proceeding of IEEE International Conference on RFID. *Dallas: IEEE Computer Society*, p. 175-182.

[2] Shawn, R. J., Minos, G., Michael, J. F. (2006). Adaptive cleaning for RFID data streams. *Proceeding s of the 32nd International Conference on Very Large Data Bases.Seoul: VLDB* Endowment, p 167-174.

[3] Yaozong, Liu., Hong ZHANG., Fawang HAN., Jun TAN. (2015). An Efficient RFID Data Cleaning Method Based on Wavelet Density Estimation. *Journal of Digital Information Management*. 13 (1). 10-14.

[4] Yu, Gu., Li, Xiao-jing., Lv, Yan-fei., Yu, Ge. (2008). Integrated Data Cleaning Strategy Based on RFID Applications. *Journal of Northeastern University*, 2008, 29(11) 1552-1555.(In Chinese)

[5] XIA Xiu- feng., XUAN Li-juan., LI Xiao-ming. (2011). RFID Uncertain Data Cleaning Strategy under Shunting Mechanism. *Computer Science*, 38(10A),22-25.(In Chinese)

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

[7] QIN Yuanxiang., DUAN Liang., YUE Kun. (2013). Approach for cleaning uncertain data based on information entropy theory. *Journal of Computer Applications*.33 (9) : 2490-2492, 2504.

[8] Yao-zong Liu., Yong-li Wang., Wei Wei., et al. (2009). Feature Selection for Classifying Data Streams Based on Maximum Entropy//*Chinese Conference Pattern Recognition*, pages 1-5.

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

[10] SONG Guo-Jie., TANG Shi-Wei., YANG Dong-Qing., WANG Teng-Jiao. (2003). A Spatial Feature Selection Method Based on Maximum Entropy Theory. *Journal of Software*. 14 (09) 1544-1550.(In Chinese)