ISSN No. 0976-5697

Volume 5, No. 4, April 2014 (Special Issue)

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Sensor Based Techniques for Searching Dimension Incomplete Databases

Sneha Arjun Dhargalkar
Department of Computer Engineering
Goa College of Engineering
Farmagudi, Ponda-Goa
snehadhargalkar@gmail.com

A.U. Bapat
Department of Computer Engineering
Goa College of Engineering
Farmagudi, Ponda-Goa
uab@gec.ac.in

Abstract: In recent years, wireless sensor networks (WSNs) are pervasively used in environment monitoring applications. It is paramount that data from these sensors be reliable since it could be used for critical decision making. However the data acquired cannot be used directly as it suffers from noise, missing data and incompleteness. When the dimensionality of the collected data is lower than its actual dimensionality, the correspondence relationship between dimensions and their associated values is lost resulting in dimension incomplete problem. Querying incomplete databases has gained substantial research interests. Many techniques are being proposed to deal with incomplete databases by estimating and replacing missing sensor values using a well-suited statistical imputation technique. Some of the methods which are applicable to impute missing data in sensor readings are WARM (Window Association Rule Mining), CARM (Closed Itemsets based Association Rule Mining), however these methods are used as avoidance methods, which detect incomplete data and impute the value for the missing value before storing the data into the database, to further avoid querying dimension incomplete databases. No substantial research has been focused to deal with missing values present in the existing databases. Querying such dimension incomplete databases could lead to obtaining incomplete results. Considering this limitation this paper proposes to incorporate the above avoidance methods as a part of searching dimension incomplete databases. The advantage of the proposed approach is that the result of the user query will always have complete data, hence avoiding incomplete results.

Keywords: Dimension Incomplete Database, Wireless Sensor Networks, Missing Data Imputation, Association Rule Mining, Window Association Rule Mining, Closed Itemsets based Association Rule Mining.

I. INTRODUCTION

Wireless Sensor Networks (WSNs) are deployed in various applications to acquire information about different anomaly in realtime, wherein, sensor nodes present in the network forward these raw sensor signals to each other. However, in sensor networks, the received data may be incomplete during wireless communication, due to synchronization problems, sensor faults, sensor power outages, communication malfunctions, packet collisions, signal strength fading, etc. Most autonomous learning techniques are not robust to incomplete data; therefore generate incorrect results (such as false positives) for missing data.

One solution is to use a reliable transport protocol, which requires nodes to re-transmit their data until successful. But since sensor nodes are usually battery-powered, data re-transmission costs significant additional energy. In addition, the re-transmission process could incur a delay in the decision making when the processing algorithms are embedded in the network. If additional nodes are added to create a dense WSN that relies on data redundancies, the extra hardware is more costly and adds to the computational cost of routing path discovery.

A better technique for many applications is to estimate and replace missing sensor values using a well-suited statistical imputation technique. Imputation is the process of calculating most probable values and replacing missing data with the calculated values. Since analyzing missing data could lead to inaccurate results, imputation is seen as a way to avoid pitfalls involved with list wise deletion of cases that have missing values, wherein when one or more values are missing for a case, most statistical packages default to discarding any case that has a missing value, which may introduce bias or affect the quality of the results. Imputation preserves all cases by replacing missing data with a probable value based on other available information. Once missing values have been imputed, the data set can then be analyzed using standard techniques.

There are various imputation methods available for imputing missing data; however most of these methods are not suitable for sensor readings. Some of the methods which are applicable to sensor readings are WARM (Window Association Rule Mining), CARM (Closed Itemsets based Association Rule Mining), but these methods are used as avoidance methods (detects the presence of incomplete data and imputes the value for missing value before storing the data into the database) to further avoid querying dimension incomplete databases.

This paper proposes to incorporate these avoidance methods, as a part of searching dimension incomplete database (recovery methods), so that any query on dimension incomplete database will always lead to the complete user query result.

The remainder of this paper is organized as follows: Section II describes the literature survey that was carried out. Section III explains the existing approaches. Proposed approach for searching dimension incomplete databases is provided in section IV and Section V concludes the paper.

II. LITERATURE SURVEY

A. The Nature of Missing Data [2]:

a. Missing completely at random (MCAR):

Data may be missing due to various reasons. It includes malfunctioning of equipment, improper data entry; the

CONFERENCE PAPER
Two day National Conference on Innovation and Advancement in Computing
Organized by: Department of IT, GITAM UNIVERSITY Hyderabad (A.P.) India
Schedule: 28-29 March 2014

weather was terrible, etc. For data to be missing completely at random, the probability that an observation (Xi) is missing is unrelated to the value of Xi or to the value of any other variables involved in the analysis [2]. One prominent feature in this case is that the analysis of data remains unbiased. Design may lose power, but the estimated parameters are not biased by the absence of data [2].

b. Missing at random (MAR):

For example, people who are depressed might be less inclined to report their income, and thus reported income will be related to depression. Depressed people might also have a lower income in general, and thus when we have a high rate of missing data among depressed individuals, the existing mean income might be lower than it would be without missing data. However, if, within depressed patients the probability of reported income was unrelated to income level, then the data would be considered MAR, though not MCAR [2].

c. Missing Not at Random (MNAR):

If data is neither MCAR nor MAR then it is said to belong to the category of, Missing Not at Random (MNAR). For example, if we are studying annual improvement in an academics and student who have scored very less marks are less likely than others to reveal their marks, the data are not missing at random. Clearly the mean score for the available data will not be an unbiased estimate of the mean that we would have obtained with complete data [2].

Although missing data is not well-studied in the area of WSN'S, many missing data imputation methods have been developed outside the WSN research area, some of this methods are described below:-

B. Traditional approaches for handling missing data

a. The simplest approach--listwise deletion:

The most common approach to missing data [3, 4] is to simply omit cases having the missing data and run the analyses only on the complete cases. This approach is also known as complete case analysis, since it takes only complete cases into consideration. If the data is missing completely at random, it leads to unbiased parameter estimates. Along with the advantages it also has many disadvantages like when the data are not MCAR, it gives bias results. Proper decision making or knowledge discovery in large data sets cannot be made because of this problem.

b. A poor approach--pairwise deletion [4]:

This method is also known as "unwise" deletion. In this approach each element of the inter correlation matrix is estimated using all available data. If one applicant during the survey reports his weight and gender, but not his age, he is included in the correlation of weight and gender, but not in the correlations involving age. The disadvantage of this approach is that the parameters of the model will be based on different sets of data, different standard errors and with different sample sizes. It is even quite possible that the generated inter-correlation matrix be not positive definite, which is likely to bring your entire analysis to halt.

c. Hot deck imputation:

[4, 5] discusses this method. In the 1940's and 50's most people seemed to feel that they had a responsibility to fill out surveys, and, as a result, most of the questions were answered leaving out only few missing data. In this, the replacement data was randomly selected from a collection of similar participants (based on data). This method is not much used anymore.

d. Mean substitution:

The basic idea of this approach is to substitute a mean for the missing data [3]. For example, if you don't know age of the particular person, then just substitute the mean age in place of missing value and continue. This is a very simple and efficient method but it comes with various problems like, no novel information is added, the overall mean, with or without replacing missing data will be the same, this process leads to an underestimate of error etc.

e. Regression substitution:

In this approach previously collected data is analyzed to make a prediction, and then substitute that predicted value as if it were an actual obtained value [6]. This approach has one advantage over mean substitution. At least the imputed value is in some way conditional on other information we have about the person. With mean substitution, if we were missing a person's weight we assigned him the average weight. Put somewhat incorrectly, with regression substitution we would assign him the weight of males of around the same age. But the problem of error variance still persists.

f. Interpolation:

It is a method of constructing new points within the range of known data points. We are using Interpolation for treating missing data in datasets.

C. Modern Approaches--Maximum Likelihood and Multiple Imputation:

a. Maximum Likelihood:

The principle of maximum likelihood [7] is fairly simple, but the actual solution is computationally complex. In this method we obtain some parameter estimates, fix those parameters into a regression equation to produce "imputed" missing values, and then start through the cycle again getting new parameter estimates until our parameter estimates do not change noticeably from one replication to the next.

b. The EM Algorithm :

One of the most common way to obtain maximum likelihood estimators is Expectation-Maximization algorithm [7], abbreviated as the EM algorithm. Though the basic scheme is simple, the calculation requires more effort and it also leads to biased parameter estimates. This method begins with E-step. In this step we fist compute variances, covariances and means, perhaps from listwise deletion. We then use those estimates to create a regression equation to predict missing data. Following this, M-step uses those equations to substitute for the missing data. We don't really fill in the missing data, but we can calculate what the variances and covariances would be without actually going through that step. We again move towards the E-step with new means and covariances, where we again construct regression equations to "fill in" missing data. Then to an M-step, we continue this process until the system stabilizes. At this point we have the necessary means and covariances for any subsequent form of data analysis.

c. Multiple Imputation:

[8, 9, 10] discusses this approach. Here we generate imputed values on the basis of available data, just as we did with the EM algorithm. We initiate with an estimate of the means and the covariance matrix, often taken from an EM solution. Using those parameter estimates we substitute the missing values with values estimated from a series of multiple regression analyses. After creating imputed data sets which are plausible representations of the data. Chosen statistical analysis on each of these imputed data sets are performed and finally the results of these analyses are combined ("average" them) to produce one set of results Advantages:- (as indicated by Rubin (1986))

a) Standard complete-data methods are used to analyze each completed data set; moreover, the ability to utilize data collector's knowledge in handling the missing values is not only retained but actually enhanced.
b) Multiple imputations allow data collectors to reflect their uncertainty as to which values to impute.

Disadvantages:-

c) The time intensiveness imputing five to ten data sets, testing models for each data set separately, and recombining the model results into one summary.

d. K-Nearest Neighbour (k-NN):

The k-Nearest Neighbour (k-NN) method is a common hot deck method, in which k donors are selected from the neighbours such that they minimize some similarity measure. Only the complete cases can take part in the imputation process. In this method [11, 12], missing values in a case are imputed using values calculated from the k nearest neighbours, hence the name. The nearest, most similar, neighbours are found by minimising a distance function, usually the Euclidean distance.

$$E(a,b) = \sqrt{\sum_{i \in D}(x_{ai} - x_{bi})^2} \qquad (1)$$

Where

a) E(a, b) is the distance between the two cases a and b,
b) x_{ai} and x_{bi} are the values of attribute i in cases a and b, respectively, and
c) D is the set of attributes with non-missing values in both cases.

It is important to note that we should not permit already imputed cases to be donors in any of the strategies.

III. EXISTING APPROACHES

Number of problems arises when applying all the above methods mentioned in the literature survey to sensor networks applications. None of the existing statistical methods answers the question that is critical to data stream environments: how many rounds of information should we use in order to get the associated information for the missing data estimation? It is difficult to draw a pool of similar complete cases for a certain round of a certain sensor when it needs to perform the data estimation. Since the missing sensor data may or may not be related to all of the available information, using all of the available information to generate the result would consume unnecessary time. And the last but not the least, sensor data may or may not Miss At Random (MAR), which makes it unfavorable to use those statistical methods that require the MAR property.

Some of the methods applicable for wireless sensor networks are:-

A. WARM (Window Association Rule Mining):

[13] Discusses this power-aware technique. WARM method saves battery power on sensors, instead of requesting sensor nodes (MS) having missing reading, to resend their last readings; an estimation of the missing value(s) is performed by using the available values at the sensors relating to the MS through association rule mining.

The DSARM approach can be stated in the following way:- Let $I = \{i_1, i_2, \ldots. i_n\}$ be a set of sensors. Let the size of the sliding window be *w* rounds. Given is the set *D* of last *w* rounds of reported sensor states, where each round *T* consists of reported states for the sensors in *I*. Find all association rules of the form $X \rightarrow Y \mid s$, (pronounced *X* determines *Y* w.r.t. *s*) where *s* is a sensor state out of all possible sensor states, and *X* and *Y* are subsets of *I* of size one (i.e. each *X* and *Y* is an item of *I*) and $X \cap Y = \emptyset$.

An association rule $X \rightarrow Y \mid s$ is said to hold in the set of the currently stored rounds *D* with the *actual confidence actConf* if *actConf* percent of the rounds that report *s* for *X* also report *s* for *Y*, and with the actual support *actSup* if *actSup* percent of the currently stored rounds in *D* report *s* for both *X* and *Y*.

The task of mining association rules then is to find all the association rules between pairs of sensors w.r.t. all possible sensor states which satisfy both the userdefined *minimum support minSup* and *minimum confidence minConf*.

The WARM data model for storing the rounds of sensor readings at the server consists of three major data structures – the Buffer, the Cube, and the Counter. Three algorithms are developed to work with this data model – checkBuffer(), update(), and estimateValue().

a) The checkBuffer() Algorithm :- This algorithm checks the Buffer at a predefined time interval for a presence of missing sensors readings in the current round. If missing values were found, it invokes the estimateValue() algorithm, else it invokes the update() algorithm.

b) update() Algorithm:- The purpose of this algorithm is to update the Cube and the Counter every time a new round (without missing values) of sensor readings is stored in the Buffer.

c) estimateValue(missingSensorID) Algorithm:- This algorithm is invoked by the checkBuffer() algorithm when a missing value is detected. The purpose of this algorithm is to estimate the missing value(s), to store it in the Buffer, and when the estimation is completed, to call the update() algorithm. Determine all eligible states for MS and create StateSets for them.

a. Advantages:-

a) By generating only the sets of all 1- and 2-frequent itemsets, the time needed for extraction of all applicable association rules, as well as the overall time for estimating the missing value, will be significantly reduced.
b) The use of the data structures containing the metadata about all possibly existing 1- and 2-frequent itemsets is now feasible, and this will lead to an additional decrease of the time needed for generating all

applicable association rules and of the overall time for estimating the missing value.

b. Disadvantages:-

a) It is based on 2-frequent itemsets association rule mining and ignores the cases where multiple sensors are associated with missing values.
b) It finds only those relationships when both sensors report the same value and ignores the cases reporting different values.

B. CARM (Closed Itemsets based Association Rule Mining):

[14] Presents an online data estimation technique called CARM, based on a closed frequent itemsets mining algorithm in data streams, called the CFI-Stream. It uses CFI-Stream data structure called DIrect Update (DIU) tree [14]. A lexicographical ordered direct update tree is used to maintain the current closed itemsets. Each node in the DIU tree represents a closed itemset. There are k levels in the DIU tree, where each level i stores the closed i-itemsets. The parameter k is the maximum length of the current closed itemsets. Each node in the DIU tree stores a closed itemset, its current support information, and the links to its immediate parent and children nodes as shown in Figure 1.

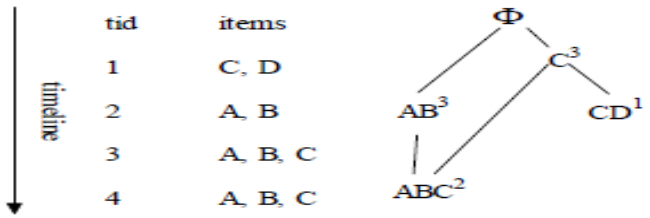

Figure1-The lexicographical ordered direct update tree

CARM proceeds in the following manner. First, it checks if there are missing values in the current round of sensor readings. If yes, it uses the current round of readings X that contains the missing items to find out its closure online. If the rules from X to its immediate upper level supersets satisfy the user specified support and confidence criteria, these upper level supersets are treated as starting points to explore more potential itemsets until CARM estimates all missing sensor data. Following this method, CARM continues to explore and find all closed itemsets that can generate association rules satisfying the users' specified support and confidence criteria. All these closed itemsets are the supersets of the exploration set and have the support and confidence along the path above or equal to the user's specified thresholds.

CARM generates the estimated value based on the rules and selected closed itemsets, which contain item value(s) that are not present in the original readings X. It weights each rule by its confidence and calculates the summation of these weights multiplied with their associated item values as the final estimated result. These item values can be expected as the missing item values with the support and confidence values equal to or greater than the users' specified thresholds. In this way, CARM takes into consideration all the possible relationships between the sensor readings and weights each possible missing value by the strength (confidence) of each relationship (rule). This enables CARM to produce a final estimated result near the actual sensor value based on all of the previous sensor relationships information.

a. Advantages:-

a) CARM can discover the relationships between two or more sensors when they have the same or different values.
b) The derived association rules provide complete and non-redundant information; therefore they can improve the estimation accuracy and achieve both time and space efficiency.
c) CARM is an online and incremental algorithm, which is especially beneficial when users have different specified support thresholds in their online queries.

b. Disadvantages:-

Accuracy and the performance of this algorithm depends on the association rules support and confidence thresholds which need to be pre-specified by users. Since users are not familiar with the monitored environments usually and the vast raw data are difficult to be understood, users may not give the proper thresholds, which results that the accuracy and the performance of the algorithms decrease greatly.

This algorithm estimate the missing data according to the frequent patterns which are pre-computed based on the existing data. If such pattern doesn't appear in the existing data then, the missing data cannot be estimated.

IV. PROPOSED APPROACH

Methods discussed in existing approaches like WARM (Window Association Rule Mining), CARM (Closed Itemsets based Association Rule Mining), are currently used as avoidance methods (basically detects the presence of incomplete data and imputes the value for missing value before storing the data into the database) to further avoid querying dimension incomplete databases. But there is no such method which deals with missing values present in the existing databases. Querying such dimension incomplete databases will result in incomplete data (Query result retrieving missing values).Taking this disadvantage into consideration this paper proposes to incorporate the above avoidance methods as a part of searching dimension incomplete databases. The advantage of the proposed approach is that the result of the user query will always have complete data.

A. Proposed Algorithm:

Following are the steps to be carried out for Searching Dimension Incomplete Databases :-

Step 1: Query the dimension incomplete databases.
Step 2: Search the records as per the user query.
Step 3: Retrieve the result.
Step4: Iterate through the dimension under consideration within the retrieved result to check if any of its value is having "NA".
Step 5: If "NA" is not present in the retrieved value then it indicates that the dataset or the data required for a given query is complete so we can directly go to step7.
Step6: If "NA" is present in the retrieved result then it indicates that our retrieved resultset contains missing values. In that case we will have to follow the below steps:-

a. Find the T_Id and Timestamp associated with the record having attribute value as "NA". *T_Id have been included for simplicity as the timestamp format can vary from user to user.
b. Apply any of the data imputation method suitable forthe wireless sensor data. In our case we will be using WARM, CARM methods based on association rule mining which takes spatio-temporal relations into consideration as our data imputation method. After imputation, substitute the estimated value in place of "NA".
c. If any more "NA" is left to be imputed then go back to step 6.a, else go to step7.

Step7: Provide the requester with the required data as the output for the query.

Figure 2, shows the entire flow of the proposed system.

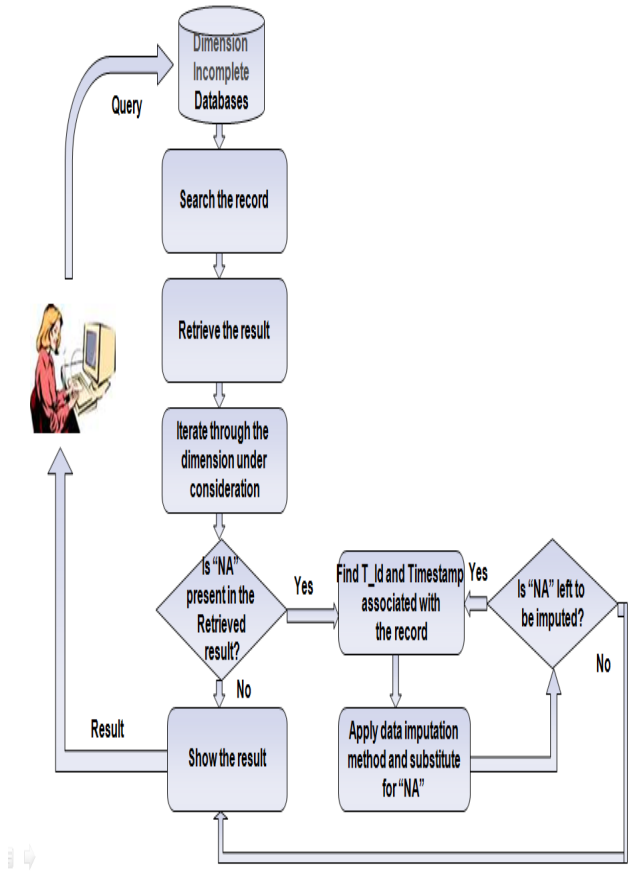

Figure 2- Proposed Design. "NA" –Not Available

V. CONCLUSION

It is extremely important that data from these sensors be reliable since actions are usually taken based on their readings. Dirty data can lead to unfavorable effects since it may be used in the activation of actuators. So for analysis of such data good data imputation method is required. There are various data imputation methods like WARM , CARM, etc but these methods act like avoidance methods which basically completes missing data before storing them to the databases. But there is no such method which deals with missing values present in the existing databases. Querying such dimension incomplete databases will result in incomplete data (Query result retrieving missing values).

This paper proposes to implement the above avoidance methods as a part of searching dimension incomplete databases so that the final query result will be the recovery version of the existing incomplete records (complete query result).

VI. ACKNOWLEDGMENT

This work was performed as part of a thesis in "Efficient Method for Searching Dimension Incomplete Databases". We want to acknowledge the contribution of our colleagues from Goa Engineering College for all the support that was provided.

VII. REFERENCES

[1]. Wei Cheng, Xiaoming Jin, Jian-Tao Sun, Xuemin Lin, Xiang Zhang, and Wei Wang, "Searching Dimension Incomplete Databases," in IEEE Transactions on Knowledge And Data Engineering, VOL. 6, No. 1, January 2013.

[2]. Barladi, A. N. &s; Enders, C. K. (2010), "An introduction to modern missing data analyses," in Journal of School Psychology, 48, 5-37.

[3]. Geeta Sikka, Arvinder Kaur Takkar, and Moin Uddin, "Comparison of Imputation Techniques for Efficient Prediction of Software Fault Proneness in Classes," World Academy of Science, Engineering and Technology 38 2010.

[4]. Gabriele B. Durrant, "Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review," ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute (S3RI), University of Southampton, June 2005.

[5]. Zeng, Yan, "A Study of Missing Data Imputation and Predictive Modeling of Strength properties of Wood Composites," "Master's Thesis, University of Tennessee, 2011.

[6]. Cohen, J. & Cohen, P., West, S. G. & Aiken, L. S. (2003), "Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences," in 3rd edition.

[7]. A. P. Dempster; N. M. Laird; D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," In Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. (1977), pp.1-38.

[8]. Schafer, J.L. & Olsden, M. K.. (1998), "Multiple imputation for multivariate missing-data problems: A data analyst's perspective," in Multivariate Behavioral Research, 33, 545-571.

[9]. Donald B. Rubin , "Expert Report on Multiple Imputation" .

[10]. Donald B. Rubin , "An overview of multiple imputation," Harvard University, One Oxford Street, Cambridge MA 02138.

[11]. Per Jönsson . Ronneby, Sweden, "An Evaluation of k-Nearest Neighbour Imputation Using Likert Data," School of Engineering, Blekinge Institute of Technology, PO-Box 520, SE-372 25, per.jonsson@bth.se, claes.wohlin@bth.se .

[12]. YuanYuan Li a,*, Lynne E. Parker b, "Nearest neighbor imputation using spatial–temporal correlations in wireless sensor networks," Information Fusion 15 (2014) 64–79.

[13]. Mihail Halatchev, Le Gruenwald, "Estimating Missing Values in Related Sensor Data Streams," The University of Oklahoma, School of Computer Science, Norman, Oklahoma 73019, U.S.A., 1-(405)-325-3498.

[14]. Nan Jiang and Le Gruenwald, "Estimating Missing Data in Data Streams∗," The University of Oklahoma, School of Computer Science, Norman, OK, 73019, USA.