A Summary of Data Cleaning Methods for Wireless Sensor Networks

Mrs. Suvarna. S. Patil
Assistant Professor
Departmen of E&CE
RYMEC
Ballari,India

Dr. B. M.Vidyavathi
Professor Department of
CSE BITM
Ballari, India

Abstract — **Over the years, wireless sensor networks are deployed in many monitoring applications to collect huge amount of raw sensed data. However, due to sensing failures of sensors and environmental effects, the collected raw sensed data contains missing and noisy values. Unnecessary and extra information processing causes a lot of energy utilization. Hence, appropriate data cleansing techniques to be used to eliminate missing and noisy data before data mining. Moreover the key task of data mining is to gain knowledge from raw data to help various decision support systems. This paper explores a survey on available data cleaning methods, which provide the different data cleaning methods used to clean missing and noisy data of wireless sensor networks.**

Keywords— Data cleaning, wireless sensor networks, data mining.

I. INTRODUCTION

An ample variety of WSNs have been installed in buildings and in natural environment [1,2] for observing various parameters like health, traffic, status of machine, weather forecast, pollution, surveillance, military etc [3,4], due to the technological development of wireless devices operating at low-power. The collected data happens to be erroneous and inaccurate due to battery power exhaustion [5], noise and malicious attacks on the network and damage of the devices. Therefore appropriate data cleaning techniques have to be used to ensure the reliability and usability of data. Also interference and machine malfunctioning reduces quality of data, which leads to missing and incorrect values [6] of Wireless Sensor Networks (WSNs). This loss in quality of data in turn impacts the decision support system performance. Hence, it is important to clean the data before applying knowledge discovery techniques.

Further it is noticed that without physical contact an attacker can exploit different mechanisms of sensor nodes by spreading malicious code over the whole network [7, 8]. This affects the data collection process of senor nodes and adds incorrect and missing readings. Generally, sensors are placed in remote areas, making replacement of their batteries impractical. Thus the other type of error is fading away of sensors energy levels [9].Usually sensors (such as automatic weather stations)use solar energy to recharge their battery, so the batteries of such sensors may become flat because of a cloudy day. This reduction in energy levels of sensor nodes may deteriorate the sensing capabilities of the sensors. The data produced by such low energy sensors may contain error or loss of data. The existence of missing and/or erroneous values in a data set seriously impacts the performance of decision support systems. Hence we need to cleanse the data before knowledge extraction.

Traditionally, a field expert manually reviewed the collected data to identify any uncommon incidents, which were slow, expensive and tedious. Manual review is not scalable for huge amount of dataset [10]. Therefore, it was necessary to have an automated data cleansing process. The data cleansing process is as shown in figure 1.

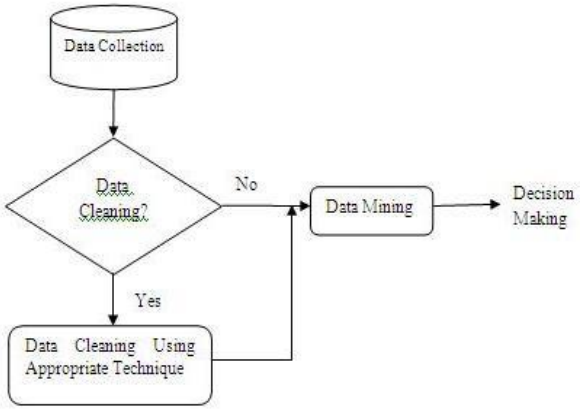

Fig 1: Data Cleansing Process

A data cleaning method was proposed by [10] to clean sensor data automatically. Initially a single sensor history was used and then a probable model of the sensor's behavior was derived. Further, the sensors future readings quality is assessed by computing their probability over the model. The sensor status decides good or bad depending on the analysis made for the future readings. If the status is bad, such data is ignored. Instead, data is accepted if the sensor is examined as good. It was found that the performance of this method [10] heavily depended on the probability sample which was built from the data history. Usually, there will be many sensors surrounding a sensor in a so formed sensor network. The model can be made stronger by using the data produced by these neighboring sensors. However, redundant data gets collected at the sensor which monitors overlapping and neighbouring region. The more accurate predictions about the interdependencies of the sensors can be achieved by using redundant data of neighbouring sensors without requiring long term historical records [11]. This proposed method [11] uses the data collected from surrounding sensors to assess the data of a given sensor. The algorithms used to predict the sensor data are, artificial neural network (ANN), k-nearest neighbours (KNN), and locally weighted regression (LWR).

Predicted data was compared with the sensor reading to determine erroneous values. If the read data happens to be erroneous, a new value is calculated based on actual sensed and predicted values, to replace erroneous value. It was observed that the learning process had shortcomings. For example, the training time taken by ANN algorithm is more. LWR deals with entire training data whenever it computes produced instance. Due to the dissimilarity in the patterns of current and past data, the learning algorithms had a serious issue in identifying noisy values and imputing missing values. Therefore, some cleansing techniques [12] performed clustering of data in the first stage to make groups of similar items as a single entity. The pre-processing step required for sensor data analysis uses some offline cleaning methods. It was a common practice to have a static dataset to analyse the data. However recent studies have shown the usefulness of statistical methods like machine learning and data mining for automatic data cleaning in various applications.

In this paper we present various researches conducted to address issues in data cleaning used in wireless sensor networks, which have changed the previous data cleaning techniques and projected new data cleaning algorithms considering the sensor network limitations, such as data generated by sensor networks is dirty in nature because of missing and erroneous data. The classification of different data cleaning techniques is as given below.

II. CLASSIFICATION OF DATA CLEANING TECHNIQUES FOR WSNS

In this section, we have provided a classification of cleaning methods designed for WSNs data. At the top, classification of general data cleaning techniques like cleaning data at base station and in-network (sensor side) are listed. Data cleaning at base station have adapted the traditional regression, neuro-fuzzy regression and high-level declarative queries to detect and clean noisy data of WSNs. In-network data cleaning classes used regression (related attributes), clustering, relational based, neuro-fuzzy regression, CAIRAD and FIMUS based upon the data collected in relaying sensor nodes.

The next classification level gives various approaches capability of cleaning noisy and/or missing data. Main idea behind in-network processing is to restrict the messages and communicating energy before sending to server. This improves the WSNs lifetime and can cleanse data to a maximum extent, whereas, in data cleansing at base station, the central server gets entire network data for cleansing. The central server has rich resources; hence there are no limitations for selecting the algorithm. Cleansing at central server always discouraged researchers because of large flow of data and communication which creates queue and loss of communication bandwidth. They play a key role in cleaning. So a suitable method is selected for cleaning. The hierarchy structure for data cleaning is as shown in figure 2.

Wireless Sensor Networks (WSNs) are mainly deployed in many monitoring applications, like weather, battlefield, healthcare, etc. However, the gathered data may be inaccurate and unreliable because of inclusion of noise, energy exhaustion and various other reasons. A lot of extra power was consumed while transmitting redundant and erroneous data. Below we list a survey of existing researches on data cleaning at base station and in-network or sensor side for imputing missing values and eliminating noisy data.

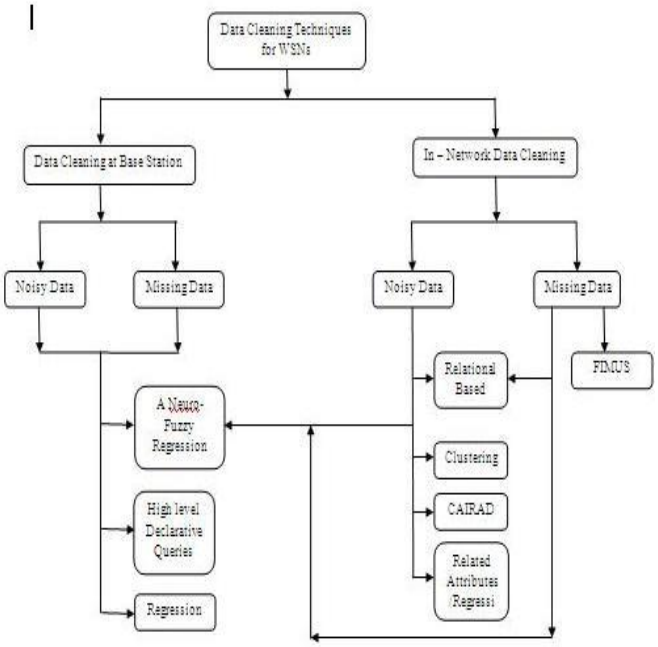

Fig 2: Hierarchy of Data Cleansing Techniques for WSNs

Haeyoung Bae, Ying Xia, Haiyang Bi, Jun Huang & Jianjun Lei [13] proposed a data cleaning method for performing cleaning at sensor-side. Here, a comparison between the attributes of data being local and data being spatial were made to determine the things that are very different from others. This outlier determination was made by using the designed sensor side processing architecture. It is the first method to be used to clean noisy data of WSNs. Here, cleansing operation is performed by four separated tasks; the first step involves the algorithm implementation in the sensor side for detecting outliers. Entire collected data of the sensor is checked by the algorithm to find outliers. Then, along with abnormal data, complete dataset is transferred to sensor's relaying neighbour node. Neighbour node differentiates event outliers from remaining outliers found at stage 1 using implemented algorithm. From neighbour node data is sent to the sink node. The sink node runs smoothing algorithm to handle the data which may be lost while transmitting to the sink. The tool used for algorithm implementation is MATLAB. New ideas on variable-relationship and spatial-relationship are used to recognize faulty readings and event outliers. This technique is evaluated taking real-time measurements of Intel Lab. Due to packet loss the dataset contains missing values. These lost values are replaced within sliding windows by the mean of points which exist before the lost points. The experiment was conducted taking temperature measurements as training samples and recorded data of node 10 to predict 30 upcoming humidity values by univariate linear regression model and quadratic regression model. Further, taking node 9 as the broadcasting neighbour at stage 2, it is shown that, data faults are separated from abnormal things detected in stage 1. The in-network solution has many advantages over the traditional server-side counterparts. (a)

Gives assurance of time for various applications. (b) Data damage due to faulty data is prevented by reducing transmission of erroneous data and reduces communication overhead. (c) Computation at sensors is more ascendable and adaptable than traditional centralized approaches. Experiments show that the implemented algorithms are capable of cleaning the data precisely and are more power-efficient in resource-limited wireless sensor networks.

Nadeem Iftikhar [14] proposed a relational-based sensor data cleaning to process data at the sensor device. Due to limited processing and storage capabilities of sensors, various works done on cleaning noisy values and imputing missing data, like classification and prediction techniques, reasoning engine and fuzzy/approximate matching algorithms can't be used for computing sensors' data. The relational based sensor data cleaning approach makes use of the deployed technology on sensors, such as a light-weight RDBMS and the relational technologies, like restrictions, causes and aggregations are used to clean and validate the sensor data locally. Experiment is conducted on the data issues in a farm project, LandIT [15], where diverse farm spray related activities are collected by the deployed wireless physical sensors on tractors. The data cleaning process involves of two phases, data validation and data aggregation. Validation process consists of removing wrong, partial and redundant rows, whereas data aggregation process manages missing and delayed values. The proposed work is simple but efficient, effective and adaptable, which can be used to cleanse data at resource-constraint devices.

Chris Mayfield, Jennifer Neville and Sunil Prabhakar [16] presented ERACER, a statistical framework to maintain relational documents quality. For cleaning databases the proposed approach uses incomplete and erroneous data, correlated attributes and high-level dependencies among the attributes. This statistical method is able to learn dependencies between diverse attributes within the sensor, and also association between the neighbouring sensors measurement due to spatial position. Observed dependencies are then used to deduce missing values and/or corrupt values. As examples, the two databases considered for experimentation are; (01) Reasoning genealogical databases missed birth and death years of persons. (02) Deducing missing data in sensor networks. Here a statistical framework called ERACER was developed to jointly perform data cleaning and replacing missing data with substituted values. The proposed technique shows the implementation of deduction and maintenance processes powerfully at the record point. Further, the information available in the database is used to detect errors precisely and rectified data values improved the quality of supposition for the missing values. Algorithm is implemented using SQL and other functions. The framework achieved accuracy compared to a Bayesian networks with the same deduction. But, this framework has various applications than Bayesian network baseline, due to its capability to understand complex and cyclic relational dependencies.

Jeffer, S.R., Alonso, G., Franklin, M.J., Hong, W., & Widom, J, [17] presented an Extensible receptor stream processing (ESP), a framework designed for cleaning sensor data using high level declarative queries. It is an infrastructural approach to sensor data cleaning, where cleaning is performed between the sensor device and the application. The cleaning process used here is less complex and adaptive. ESP is designed in such a way that the application gets protected due to changing characteristics of errors in devices or environmental changes. Overall, the deployment cost is reduced allowing various applications to use the same cleansed data. The concepts of spatial and temporal granules are used to drive ESP's cleaning mechanisms. ESP uses these ideas to clean streaming sensor data through a pipeline processing stages. It is shown that the deployment and evolution of ESP infrastructure is easier because of the listed features: Declarative- the logic of cleaning in ESP infrastructure is easy through the use of high-level declarative queries. The queries are efficiently executed utilizing relational query processing. Pipeline- ESP cleaning operations consist of independently programmed segmented pipeline of cleaning operations. Cleaning framework- different cleaning logics are designed to handle sensor's error characteristics.

Alfredo Petrosino and Antonino Staiano, [18] proposed a neuro-fuzzy regression method. For ambiguity reduction and hence finding exact estimation of reading the ANFIS model is developed. Here, the learning algorithms are qualified based on earlier period to build a regression model. Training is performed at the base station level. The behavioural model of the sensors is described by the regression model. This derived behavioural model is used for correcting readings in two ways: (a) The missing reading is replaced by regression function itself. (b) Based on the prediction of regression model, the new readings are corrected. This process is done at central server or at sensor. The proposed Adaptive Neuro Fuzzy Inference System (ANFIS) was examined on 54 sensors measurements installed in the Intel Berkeley Research lab [19]. This method was compared with a polynomial regression kernel. But kernels' are learning methods dependent on memory; hence ANFIS approach is preferred because of computational lightweight. The cleaning results showed that ANFIS performance is better than kernel methods. It is also demonstrated that the ANFIS model is more effective for performing cleaning at sensor's level rather than at central-server.

Asmaa Fawzy Hoda M.O. Mokhtar, Osman Hegazy [20] proposed an outlier detection method, also called deviation detection or data cleaning, an important preliminary data mining practice for data analysis. Outliers are those whose characteristics will be different from that of the normal profile. Outliers exist due to mechanical faults, system changes, deceptive nature, human errors, and instrumental errors or simply through natural changes in population. In this approach, algorithm based on clustering and nearest neighbour outlier detection methods are combined for efficient clustering and outlier detection. An outlier detection technique is developed, which takes more number of steps. The first stage uses cluster-based approach to cluster the known information into normal and isolated clusters for saving energy and communication overhead. Next stage considers only detached clusters to find its source by testing correlations spatially and temporally; such that the readings from neighbouring nodes gives outlier readings if the time space between these values is small. It means an event exists which

remains for longer time and changes maximum sample of observed data. Else the algorithm treats the data to be noisy or error because of low power. The proposed method is tested with CollECT [21][23] for its precision and scalability on both real dataset obtained from Intel Berkeley research lab and synthetic dataset. Thus this approach obtained higher accuracy rate for identifying outliers, events and errors.

Md Zahidul Islam1, Quazi Mamun and Md. Geaur Rahman [6] proposed a scheme that uses data detection approach for identifying corrupt and missing values gathered from sensors. The existence of missing and incorrect values may drop data quality and also seriously influenced the decision support system performance. This is explained in a better way by conducting experiments on Intel lab data set accessible in the Intel Berkeley Research Lab [22]. The classification accuracy is computed by artificially creating missing values in the dataset and then using C4.5 classifier on existing and the lost data. It was observed that due to missing values the accuracy reduces. It shows the erroneous data obtained from the WSNs should be cleaned.

Table 1 shows a comparison of data cleaning techniques for wireless sensor networks.

Table 1: Comparison of data cleaning techniques for WSNs

Approach	Data Cleaning Method	Advantages	Limitations
Outlier detection and event outlier detection [13]	Related attributes	Energy and memory	Missing values were not considered
Relational based [14]	Data validation and data aggregation	Efficient, effective and adaptable	Not suitable for other application areas
ERASER [16]	Regression or Convolution	Efficient and scalable	Computational complexity
ESP [17]	High level declarative queries	Energy and adaptive	Computational complexity
ANFIS model [18]	A Neuro-Fuzzy regression	Time and memory	Less Accurate
Clustering [20]	Clustering and nearest neighbour outlier detection	Energy, communication and accuracy	Data dissemination and routing in WSNs
CAIRAD and FIMUS [6]	C4.5 Classifier	Prediction accuracy, communication	Nodes energy is ignored

III. CONCLUSION

In this paper we focus on the survey of data cleaning techniques for wireless sensor networks, to clean noisy and missing values of sensors. Conventional data cleaning methods cannot be used for WSNs because of typical features and resource restrictions of the networks. In this study we provide a detailed survey of the various data cleaning techniques used for WSNs. From examination it is found, different types of data cleaning algorithms are developed for different types of sensors' data. These algorithms individually can solve specific problems concerned to SNs data. After analysing the various developed methods, it is observed that network oriented cleaning techniques are more efficient, since it reduces communication overhead.

REFERENCES

[1] Amro, A., Elhajj, I.H., & Awad, M, "Energy-aware discrete probabilistic localization of wireless sensor networks", Intelligent Automation & Soft Computing, 19, 407–423, 2013.

[2] Huang, J., Meng, Y., Liu, Y., & Duan, " A novel deployment scheme for green internet of things", IEEE Internet of Things Journal, 1, 196–205, 2014.

[3] Jiang, Y.Q., Li, T., Zhang, M., Sha, S., & Ji, Y.H, "WSN-based control system of co2 concentration in greenhouse. Intelligent Automation & Soft Computing", 21, 285–294, 2015.

[4] Lei, J.J., Park, T., & Kwon, G.I, "A reliable data collection protocol based on erasure-resilient code in asymmetric wireless sensor networks", International Journal of Distributed Sensor Networks, 2013, 1–8, 2013.

[5] Subramaniam, S., Palpanas, T., Papadopoulos, D., Kalogeraki, V., & Gunopulos, D, "Online outlier detection in sensor data using non-parametric models", In Proceedings of the 32nd International Conference on Very Large Databases, Seoul, Korea, 2006. .

[6] Md Zahidul Islam1, Quazi Mamun and Md. Geaur Rahman, "Data Cleansing during Data Collection from Wireless Sensor Networks", Proceedings of the Twelfth Australasian Data Mining Conference (AusDM 2014), Brisbane, Australia, 2014.

[7] Giannetsos, T., Dimitriou, T. & Prasad, N. R, "Self-propagating worms in wireless sensor networks", in 'Proceedings of the 5th international student workshop on Emerging networking experiments and technologies', ACM, pp. 31–32, 2009.

[8] Sharma, K. & Ghose, M, "Cross layer security framework for wireless sensor networks", International Journal of Security and Its Applications5(1), 39–52, 2011.

[9] Ni, K., Ramanathan, N., Chehade, M. N. H., Balzano, L., Nair, S., Zahedi, S., Kohler, E., Pottie, G., Hansen, M. & Srivastava, M., "Sensor network data fault types", ACM Transactions on Sensor Networks (TOSN) 5(3), 25, 2009.

[10] Dereszynski, E. W. & Dietterich, T. G, "Probabilistic models for anomaly detection in remote sensor data streams", in Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, (UAI2007) , 2007.

[11] Ramirez, G, "Assessing data quality in a sensor network for environmental monitoring", 2011.

[12] Cheng, K., Law, N. & Siu, W, "Iterative biclusterbased least square framework for estimation of missing values in microarray gene expression data", Pattern Recognition 45(4), 1281–1289, 2012.

[13] Jianjun Lei, Haiyang Bi, Ying Xia, Jun Huang & Haeyoung Bae, "An in-network data cleaning approach for wireless sensor networks", Intelligent Automation & Soft Computing. Vol 22, No. 4, 599-604, 2016.

[14] Iftikhar, N., & Nordbjerg, F. E, "Relational-Based Sensor Data Cleansing", In T. Morzy, P. Valduriez, & L.Bellatreche (Eds.), New Trends in Databases and Information Systems: 19th East-European Conference on Advances in Databases and Information Systems, ADBIS 2015 (pp. 108-118). (Communications in Computerand Information Science; Vol. 539, No. 10.1007/978-3-319-23201-0_13). Poitiers, France: Springer, 2015.

[15] LandIT, http://www.tekkva.dk/page326.aspx

[16] Chris Mayfield, Jennifer Neville and Sunil Prabhakar, "ERACER: A Database Approach for Statistical Inference and Data Cleaning", SIGMOD'10, June 6–11, Indianapolis, Indiana, USA. Copyright 2010 ACM 978-1-4503-0032-2/10/06, 2010.

[17] Jeffer, S.R., Alonso, G., Franklin, M.J., Hong, W., & Widom, J, "Declarative support for sensor data cleaning", Pervasive Computing,3968, 83–100, 2006.

[18] Alfredo Petrosino and Antonino Staiano, "A Neuro-fuzzy Approach forSensor Network Data Cleaning", B. Apolloni et al. (Eds.): KES 2007/ WIRN 2007, Part III, LNAI 4694, pp. 140–147, 2007._c Springer-Verlag Berlin Heidelberg 2007.

[19] Intel Berkeley Laboratory Data, http://berkeley.intel-research.net/labdata/.

[20] Asmaa Fawzy Hoda M.O. Mokhtar , Osman Hegazy,"Outliers detection and classification in wirelesssensor networks", Egyptian Informatics Journal (2013) 14, 157–164, Elsevier 2013.

[21] Wang K, Yang S, Chang P, Shih C, "CollECT: collaborative eventdetection and tracking in wireless heterogeneous sensor networks.", In: Proceedings of the 11th IEEE Symposium on Computers andCommunications (ISCC); p. 935–40, 2006.

[22] IBRL-Web[onlineavailable:http://db.lcs.mit.edu/labdata/labdata.html] .Accessed August 7, 2014. URL: http://db.lcs.mit.edu/labdata/ladata.html

[23] Parameshachari B D et. al"Design and Analysis of Band Pass Filter and Matching Network Using ADS", National Conference on Recent Trends & Applications in Electrical & Electronics Engineering, at KSIT, Bengaluru10-12 May '2017.