Researh on Qualitative Data Cleaning for IoT Streming Data

Chen Binger

1. Research Content

This Research focus on qualitative data cleaning method for IoT(Internet of Things) streaming data, especially the characteristics of time series and the multidimensional data in WSN(Wireless Sensor Networks). Considering the big data scale of IoT, the method is further improved by incremental learning[1].

2. Background

IoT is dramatically changing our lives. The large amount of data produced contains valuable information. WSN is a typical IoT instance. It is used for real-time detection of real-world environments or events through multi-sensor data collection and communication and widely applied in smart homes, smart cars, smart cities, smart health and so on. Data mining for IoT can promote more intelligent services. The IoT data is usually streaming time series data which is one of the most studied data types. A recent survey by KDnuggets found that 48% of analysts analyzed time series data in the past, second only to table data[2]. This abundant real-time data with temporal characteristics can provide significant research values.
Due to the low cost of sensors, limited resources, and unstable IoT environment, theres many dirty data such as data missing, data anomalies[3][4]. Its very common in time series. For example, the abnormality of sensor readings in the GPS track[5]. Direct use of these incomplete, noisy, inconsistent data in real world can lead to erroneous decisions and unreliable analysis in applications such as pattern mining[6] or classification[7]. Improving data quality to support following applications is necessary[8]. Data cleaning is a key area of big data management. In addition, due to the big scale of IoT streaming data, improvements should be made in original data cleaning methods considering the time and space costs.
The most common errors in time series are spike errors and consecutive errors (including missing values)[9]. The data cleaning methods can be either quantitative or qualitative. Quantitative data cleaning often involve statistical methods to identify abnormal behaviors and errors[10][11], while qualitative techniques are based on a series of data quality rules like integrity constraints[12]. This process usually leverage data mining or expert knowledge. There are few studies on qualitative data cleaning for IoT streaming data which will be addressed in this research.

3. Related Works

There have been a lot of research on time series data cleaning, but they have some limitations. Although IoT streaming data cleaning has become a popular research area, most of the previous research focused on the quantitative level, such as outlier detection[13][14].
Existing data cleaning methods are mostly used for spatial data and not suitable for time series[15]. As for the research on data cleaning of time series, the general method can only be used for offline data rather than online streaming data. Even if some studies aimed at online cleaning, they still didnt make full use of the temporal characteristics of time series for rule mining[9].
Many methods regard dirty data as noise and discard them after detecting[16]. However, such incomplete time series can lead to the unreliability of following research. Repairing the abnormal data to approach true value can improve the results of subsequent data processing[17][7].
Moreover, many methods can only solve one kind of data error, some only dealing with spike errors[18]. [9] can cover both spike errors and consecutive errors, but it cannot solve the problem of missing value.
Many other studies of IoT streaming data cleaning can only clean one-dimensional data[12][19][20]. However, in WSN therere multiple sensors working together to obtain effective information. Some research has studied multidimensional data, but they assumed only one sensor reading is faulty at a time[9]. This research believes that the problem can be solved by feature extraction in deep neural networks. Similar attempts have been made before, but these attempts have neither considered qualitative data cleaning[21], nor have they considered the drawbacks of neural networks[22].
This research is committed to solving these problems.

4. Proposed Solution

4.1 System Overview

The architecture of the proposed solution is shown in Figure 1. In order to perform data cleaning, speed constraints, rule mining on every timestamp[20], and LSTM(Long Short-Term Memory) modeling of time series are deployed on WSN historical data to analyze data features in multiple aspects. Then active learning will be applied on the corrected data to select the appropriate amount for human labeling. These data will be used to update the data features through incremental learning.

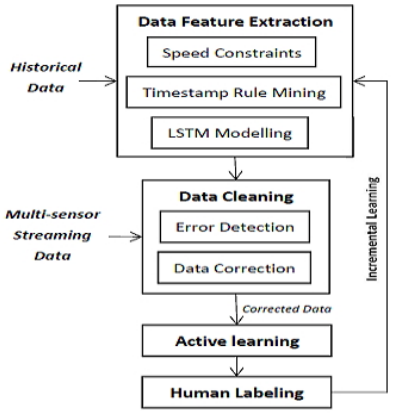

Figure 1. System Overview

4.2 Multi-sensor Time Series Data Cleaning

(1)Speed Constraints
It is used to control the velocity of data changing. For sensor i, the values at two adjacent timestamps t_{k-1}, t_k are s_i^k, s_i^{k-1}. The speed at t_k is then $v_i^k = \left| \frac{s_i^k - s_i^{k-1}}{t_k - t_{k-1}} \right|$. The maximum value $v_{i,max}$ is set as speed constraints.
(2)Timestamp Rule Mining
This process is to mine association rules between multi-sensor data at every timestamp. In order to generate more general rules, the data is vectorized. For example, the value range of a sensor can be divided into [0, 1), [1, 2), [2, 3]. When reading 1.5, it can be denoted as [0, 1, 0]. In this form we can generate a series of rules based on historical data. Given three sensors A, B, and C, when A values [0, 1, 0] and B values [1, 0, 0], C has 60% possibility values [1, 0 , 0] and 40% [0, 1, 0]. Therefore, if A[0,1,0,], B[1,0,0], C[0,0,1] appear at same time, it is judged that data C is abnormal, and it has confidence of 0.6 to value [1,0,0].
(3)LSTM Modeling
In order to further study the temporal characteristics of data changes, this research adopts LSTM, which is dedicated to mining the long-term dependencies of sequences and very suitable for time series data cleaning. Moreover, the feature extraction mechanism of deep neural network can also effectively solve the problem of multi-dimensional data-detect and correct the dirty data of multi-sensor. It is also universal for variable length time series.
The model is shown in Figure 2(X-Sensor Data(2 sensors), M-Mask Vector, t-Timestamp, Δ-Time Interval). A variable length sensor data time series is used as input. In addition, to solve the problem of missing values, an mask vector(setting 0 for sensor without value, 1 for sensor with value) and time interval are also input, which indicate the position and duration of missing values. The objective function is softmax loss. This model is able to fully extract the comprehensive features of time series multi-sensor data association. Once new data is generated, the characteristics of current sequence is analyzed and the rationale of current data is judged.

$$X = \begin{bmatrix} 2 & N/A & 6 \\ N/A & 1 & 3 \end{bmatrix} \quad M = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$
$$t = [0 \quad 2 \quad 3] \qquad \Delta = [0 \quad 2 \quad 1]$$

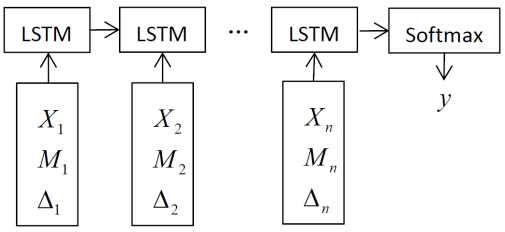

Figure 2. LSTM Modeling

4.3 Method optimization for IoT Big Data

In previous studies, historical data or expert knowledge were often leveraged to establish fixed rules for data cleaning. However, as enormous informative data are generated, it is essential to update the original rules and models. This study introduces incremental learning to leverage new data. Nevertheless, there is a risk in using the data obtained by completely automatic data cleaning. Appropriate amount of corrected data are judged by human labeling in this research, and only the reasonable correction can participate in the model update. At the big data scale, manual intervention of every outcome is impossible taking into account time and space costs. An active learning approach that allows the system to automatically select most significant new information for model update is applied.

5. Experiment

5.1 Datasets

In addition to the popular open source datasets and synthetic data, such as Gesture[23], ILD data with synthetic errors[24], this research also plans to experiment on multi-sensor data from automobiles to solve emerging real-world problems. The sensors from automobile can detect the current state of the car and the surroundings, and make decisions on driving behavior, which is of great significance for the development of autonomous driving technology. These data are planned to be collected through real-user participation or industry-university collaboration.

5.2 Evaluation

(1) Verify the accuracy of error detection and error correction respectively, then compare with baseline.
(2) Verify results for single sensor and multi-sensor time series respectively and explore how the error produced by different sensors can change the results.
(3) Increase the error amount to observe the results changing.
(4) Change the missing values amount to examine the methods effectiveness.
(5) Compare the rule-based method and method improved by LSTM to verify the LSTMs function.
(6) Compare the effectiveness of model with incremental learning and model without updating.
(7) Verify the impact of human intervention.
(8) Compare the processing time before and after introducing active learning.

6. Research Limitations

(1) It may not be possible to obtain enough historical data. Model optimization is expected to reduce the impact of this problem.
(2) Historical data annotation and human checking are still needed. This is expected to be minimized by improving the selection approach of active learning.
(3) Its time-consuming to train large neural network with high dimensional data, which cannot fulfill real-time processing. Model compression approaches such as dimensionality reduction and pruning are required.
(4) How to minimize deviation between corrected value and true value is a tricky problem.
(5) The models scalability needs to be further verified.
(6) Semi-structured and unstructured data cleaning is another unsolved problem.

7. Significance

This research provides a reliable solution for the most common multi-sensor time series cleaning problem in real world. The method is universal and can be used in a variety of scenarios,

which compensates the drawbacks of previous studies. The corrected data are extremely close to true value. It has practical significance for the IoT, especially the smart car scenario.

REFERENCES

[1] et al. Tsai, Chun-Wei. Data mining for internet of things: A survey. *IEEE Communications Surveys and Tutorials*, 16(1):77–97, (2014).

[2] Piatetsky-Shapiro G. Url retrieved on april-2-2018. https://www.kdnuggets.com/polls/2014/data-typessources-analyzed.html, (2018).

[3] M Scannapieco C Batini. Data quality: concepts, methodologies and techniques. *Springer Publishing Company*, (2010).

[4] H Wang et al. D Ganesan, S Ratnasamy. Coping with irregular spatio-temporal sampling in sensor networks. *ACM Sigcomm Comput Communication Rev*, 34(1):125C130, (2004).

[5] P. J. Brockwell and R. A. Davis. Introduction to time series and forecasting. *Springer Science & Business Media*, (2006).

[6] Fabian Moerchen. Algorithms for time series knowledge mining. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, (2006).

[7] Jian Pei Xing, Zhengzheng and S. Yu Philip. Early classification on time series. *Knowledge and information systems*, 31(1):105–127, (2012).

[8] et al. Karkouch, Aimad. Data quality in internet of things: A state-of-the-art survey. *Journal of Network and Computer Applications*, 73:57–81, (2016).

[9] et al. Nemati, Hassan. Stream data cleaning for dynamic line rating application. *Energies*, 11(8), (2018).

[10] Austin J Hodge VJ. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85C126, (2004).

[11] Aggarwal CC. Outlier analysis. *Springer*, (2013).

[12] Papotti P Chu X, Ilyas IF. Holistic data cleaning: Putting violations into context. In *In 2013 IEEE 29th International Conference on Data Engineering (ICDE)*, (2013).

[13] S. Madden. Database abstractions for managing sensor network data. *Proceedings of the IEEE*, 98(11):1879C1886, (2010).

[14] V. Vassalos V. Stoumpos A. Deligiannakis, Y. Kotidis and A. Delis. Another outlier bites the dust: Computing meaningful aggregates in sensor networks. In *IEEE International Conference on Data Engineering*, (2009).

[15] J. Szlichta M. Volkovs, F. Chiang and R. J. Miller. Continuous data cleaning. In *IEEE International Conference on Data Engineering*, (2014).

[16] Pei J Han J, Kamber M. Data mining: Concepts and techniques. *Elsevier: New York*, (2011).

[17] C. Li S. Song and X. Zhang. Turn waste into wealth: On simultaneous clustering and cleaning over dirty data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM*, (2015).

[18] et al. Zhang, Aoqian. Time series data cleaning: From anomaly detection to anomaly repairing. *Proceedings of the VLDB Endowment*, 10(10):1046–1057, (2017).

[19] Flaster M Rastogi R Bohannon P, Fan W. A cost-based model and effective heuristic for repairing constraints by value modification. In *In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, (2005).

[20] Wang J Yu P.S Song S, Zhang A. Screen: Stream data cleaning under speed constraints. In *In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, (2015).

[21] Bin Yang Kieu, Tung and Christian S. Jensen. Outlier detection for multidimensional time series using deep neural networks. In *2018 19th IEEE International Conference on Mobile Data Management (MDM)*, (2018).

[22] et al. Krishnan, Sanjay. Boostclean: Automated error detection and repair for machine learning. In *arXiv preprint arXiv:1711.01299*, (2017).

[23] Lima C. A. & Peres S. M. Madeo, R. C. Gesture unit segmentation using support vector machines: segmenting gestures from rest positions. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing. ACM*, (2013).

[24] http://db.csail.mit.edu/labdata/labdata.html, ild.