Research on Data Quality Assessment of Accuracy and Quality Control Strategy for Sensor Networks

Jin Zhang, Yajie Ma and Doudou Hong
College of Information Science and Engineering, Wuhan University of Science and Technology, Engineering Research Centre for Metallurgical Automation and Detecting Technology, Ministry of Education, Wuhan 430081, China.
Email: zhangjinfight@163.com

Abstract. In this paper, to make the data meeting the network application has a specific standards, it is necessary to apply a scale to measure the attributes of data in sensor network, for the concept of data-based sensor networks service quality is relatively vague. Based on the signal-to-noise ratio quality model proposed by Mini Mathew et al, the data quality evaluation standards are introduced. The definitions of accuracy and timeliness for sensor data are given with the understanding of data quality management. Then according to the maximum entropy algorithm of processing sensor data, the relationship between accuracy and network communication parameters is derived. Finally, the unity and limitations between data accuracy standard and signal-to-noise mode are discussed. The relationship between the sampling frequency and the data quality accuracy is analyzed as well.

1. Introduction

With the expansion of the sensor networks and the development of the Internet of Things, the data collected by the sensor network has exploded. In particular, for monitoring scenarios in large areas or with large data traffic, the amount of data collected by the network will. On the one hand, massive data (even big data) provides a large amount of data sources to analyze for user applications. While, it also requires more storage and leads to more onerous computational burden to the data processing. Other hand, under the influence of target node deployment location, sampling frequency, transmission channel signal-to-noise ratio, node storage space and initial energy, etc. Ultimately, the stored data could not be completely reliable with high-delay, incompleteness, the error caused by these factors makes it difficult to guarantee the data quality. A large amount of data does not adequately mean more information acquisition and more accurate data analysis results. In this case, how to assess the data quality of the sensor network, establish a relationship model between the data and network parameters affecting the data quality (signal-to-noise ratio, time delay, etc.) to predict and control the data quality is an important issue.

To assess or predict the quality of data in a sensor network, a standard that measures the quality of the data is essential. The standard can also provide a basis for further analysis of data for third-party data research and computing organizations. Based on the collected sensor network data, this paper mainly proposes a network data accuracy assessment standard for accuracy measurement in data quality, and studies the relationship between data quality and network parameters to facilitate data.

The main contributions of this paper are listed as follows:

1) The accuracy of the data is marked by communication parameters of the sensor network, such as signal-to-noise ratio and packet loss rate. A method for assessing the accuracy of the data of the
sensor network is proposed, and the connection between data accuracy and network property characteristics is constructed.

(2) This paper proposes a top-down approach to study how to control the accuracy of data in sensor networks. Then the impact of network parameters on data accuracy is qualitatively analyzed, which provides an analytical basis for subsequent prediction or control information accuracy.

The remainder of this paper is organized as follows. The Section 2 introduces the related work about this paper. In the Section 3, a sensor network data accuracy model is presented and discusses how to control data accuracy quality. The Section 4 is the experiment and Analyze and the Section 5 is the conclusion and further work.

2. Related Work on Data Quality Assessment
With the complexity of the sensor network structure, the diversification of technologies and the wide requirements of applications, many researchers have independently developed the top-level information quality framework for comprehensive consideration of sensor network performance. There are also many studies in the level of satisfaction with specific measurement of information and data. In the literature [3], the sensor network QoI framework is also standardized according to the current development of the sensor network, and at the same time various evaluation metrics of QoI are proposed. Specifically, Marjanovi [4] proposed a MCS framework of quality-driven sensing management based on the weighted calculation method of multiple properties indicators to assess the comprehensive quality of monitoring objects in the literature. It provides an assessment idea of the application but does not give a specific quantitative approach. In the literature [5], Hao proposed a semantic-difference method based on QoI and information fusion theory to measure the accuracy of sensor network information and use time-scale technique to assess information timeliness according to different delay-sensitive networks. However, the research works contain much subjective components, or information quality is only defined but quality control issues with sensor networks do not discuss.

3. Data Quality Assessment Model
In this section, we consider defining and analyzing sensor network data in the perspective of data quality management and focusing on the data itself. To explain the physical quantities, this paper focuses on the accuracy measurement of data, and gives corresponding inferences to derive the relationship between data accuracy and communication parameters based on literature [6] in the meantime. As for data quality control issues for the combination of data timeliness and sensor networks will continue to be studied detailedly in future work. Some symbols and parameters used to define the accuracy and timeliness are described in Table 1.

Table 1. Symbols and parameters description.

| Symbols | Meaning | Symbols | Meaning |
|---------|---------|---------|---------|
| $f_s$   | Sampling frequency | $N$     | Number of samples in a period $\Delta t$ |
| $T_0$   | The moment when the query message is sent to $D_i$ | $N_s$   | The number of non-null sample values over a period of $\Delta t$ |
| $T_s$   | The time when the first non-null sample is received in a period $\Delta t$ | $N_E$   | The amount of sample values that is considered wrong in this paper |
| $T_e$   | The time when the last non-null sample is received in a period $\Delta t$ | $l$     | Packet loss rate within $\Delta t$ from $D_i$ to Sink |
| $D_i$   | Target node for data query | $l_B$   | Node-level packet loss rate to Sink |
|         |         | $l_E$   | Link-level packet loss rate |

3.1. Sensor Data Quality Definition
In data quality projects, the evaluation and analysis of raw data has important practical significance. The assessment of data quality is a key step in data quality management. For different applications, there are mainly three evaluation techniques: Data Quality Survey Method (DQS), Data Quality Index Quantification Method (QDQM), and Comprehensive Data Quality Management (TDQM).
For data processing and analysis modelling in sensor networks, data needs to be quantified. Therefore, the QDQM evaluation system for data quality index quantification is selected. This paper combines the Cod database integrity constraint [7] with component-quality-dimension. The data quality measurement standards in [7] are introduced into the sensor network data management, and the main measurements of the sensor network data quality assessment include accuracy, integrity, consistency, and timeliness metrics and so on. In this paper, in order to improve the data quality requirements of the sensor network applications, more accurate and credible data need be provided for further analyze. A preliminary assessment of the sampling data of the sensor network for subsequent further processing has a quality standard reference is given as follow:

\[
q_a = 1 - \frac{N_E}{N} \tag{1}
\]

Equation (1) shows the ratio of the number of erroneous data units stored in sensor network database to the total number of data units, where, the accuracy is defined as \( q_a \). The non-null discrete sample values sequence within \( \Delta t \) is \( X_i = \{ x_1, x_2, ..., x_{N_i} \} \) by a single sensor target node, \( N_E/N_i \) is the rate of sample values that is considered wrong in this paper. In the network, node-level queued packet loss rate \( I_B \) [8] due to limited node buffer queues, and the channel collisions, transmission error and other link-level packet loss rate \( I_E \) [9] result in incompleteness of data in the database. Then it can be described as follows:

\[
\begin{cases}
N = f_s \cdot \Delta t \\
I = I_B + I_E
\end{cases} \tag{2}
\]

Under the premise that the lost data occupies the cells of the database, Equation (1) can be summarized as follows:

\[
q_a = 1 - P(\text{er}) \cdot (1-I) \tag{3}
\]

In Equation (3), \( P(\text{er}) \) is the rate considered as incorrect data in non-null data set \( X_i \). The timeliness can be described as follows:

\[
q_t = 1 - \left( \frac{\sum_{i=1}^{N_s} \tau_o(i)}{T_c - T_0} \right)^s
\tag{4}
\]

where, \( N_s = f_s \cdot (T_c - T_i)(1-I) \), \( \sum_{i=1}^{N_s} \tau_o(i) \) is the total time of the queue delay of all data in the non-null data set within \( \Delta t \) from the target sensor \( D_i \). \( s \) is the adjustment factor for better observation in wave motion of \( q_t \), with network parameters. Further discussion and research will be conducted in the future. Equation (3) shows that the accuracy of data quality is directly related to the network packet loss rate and the amount of sample values that is considered wrong. However, for the sample data of the sensor network, there is no absolute scale to measure the correctness. It can be considered that the reliability of data carrying environmental information will be different, relatively erroneous data can also be called large-errors data which is off-centered in the distribution of data due to sensor hardware, network communication environment, noise and other factors. Obviously, too much contaminated data will result in worse decision-making. Although the elimination of coarse data leads to the loss of partial information, it provides more accurate results for the upper analysis decision.

3.2. Single Sensor Node Data Processing Method

There are many single-sensor raw data processing methods given in the research. The simple estimation of the tail-cutting method is relatively subjective for the actual cutting threshold value. The research [10] assumed that the sensor data distribution is Gaussian distribution. However, the distribution sensor data is not necessarily Gaussian distribution. Therefore, this paper combines the
maximum entropy [11] algorithm to discriminate and process affected sensor data. The MEM maximum entropy estimation algorithm fits the single sensor data. When the data information entropy is maximum, the most realistic distribution can be obtained, and it has been proved to have better accuracy. The biggest advantage is that if the probability distribution is not known in advance, it minimizes potential subjectivity. The threshold value \( \eta \) in the paper is used as the data reliability segmentation to eliminate the coarse data, and relatively reliable data can be multi-segmented and marked in the future work.

For the monitoring target with constant state within \( \Delta t \), the data collected from that sensor is regarded as sampling repeatedly. Maximum entropy constraint criterion can be described as follows:

\[
\begin{align*}
\text{Max} & \quad H(X) = -\sum_{i=1}^{N_x} P(x_i) \log P(x_i) \\
\text{s. t.} & \quad \sum_{i=1}^{N_x} P(x_i) = 1 \\
& \quad E(X) = \frac{\sum_{i=1}^{N_x} x_i}{N_x} \\
& \quad D(X) = \frac{\sum_{i=1}^{N_x} (x_i - E(X))^2}{N_x}
\end{align*}
\]

where, \( P(x_i) \) is the probability of occurrence of the sample value, \( H(X) \) is the information entropy for non-null data sequences above. In order to improve the conformability of the fitting curve. More moments should be included in the constraint \( \sum_{i=1}^{N_x} g_r(x_i)P(x_i) = \langle g_r(x_i) \rangle, r = 1, 2, 3 \ldots R \). Afterwards, Lagrangian Multiplier Method can be used to figure out the distribution of data value as follows:

\[
P(x_i) = \exp[-\rho_0 - \sum_{i=1}^{R} \rho_r g_r(x_i)]
\]

In the above equation, \( \rho \) is the Lagrangian multiplier. So, the accuracy segmentation threshold can be given as follows:

\[
\eta = [\sum_{i=1}^{N_x} (x_i - E(X))^2 \cdot P(x_i)]^{1/2}
\]

The reliable data interval is: \( S = [E(X) - \eta, E(X) + \eta] \).

At this point, rough data can be found from non-null sample sequence \( X_i \). According to the database language, data that is not in this confidence interval can be filtered, so

\[
P(\text{er}) = \frac{\sum(x_i \in X, x_i \notin S)}{N_S}
\]

Or: \( P(\text{er}) = \int_{x \notin S} P(x_i)dx \)

However, the method of Equation (8) could be used in the case of a large amount of raw data in order to calculate more accurately. Therefore, the center isometric division method can be used to mark various levels of data. The farther to the ideal value the data is in the distribution, the less probability of occurrence there is. On the contrary, the lower degree of noise pollution there is, the higher accuracy and the greater probability there will be.
3.3. Data quality and Sensor Network System

3.3.1. Combination of Data Quality Accuracy and Signal to Noise Ratio Model. The following examples show how to format a number of different figure/caption combinations. The above section 3.2 has obtained the coarse data rate and can be graded. According to the sensor signal-to-noise ratio model and network packet loss rate model described in the literature [6]:

\[
\begin{align*}
SNR^a &= 10 \cdot \log \left( \frac{a^2}{n^2 + l \cdot a^2} \right) \\
SNR^m &= SNR^m - 10 \cdot \log (1 + l \cdot 10^{-10})
\end{align*}
\]

(10)

where, \(SNR^a\) is the network-signal-to-noise ratio (\(n\) is the acronym of network) at the end user, \(SNR^m\) is the signal-to-noise ratio (\(m\) is the acronym of network) measured by the target sensor and \(l\) is the network packet loss rate. \(a^2\) is the signal power and \(n^2\) is the noise power. Combining with Equation (3), we can arrive at Equation (11) as follows:

\[
q_e = 1 - P(\text{er}) \cdot [1 - (10^{-0.1 \cdot SNR^m} - 10^{-0.1 \cdot SNR^a})]
\]

(11)

When the sensor block is homogeneous and the characteristics of the data sent to the sink node are equal, the output signal-to-noise ratio almost increases to the peak value with the growth of sampling frequency. There are two methods to improve the quantization signal-to-noise ratio:

1. Increase the quantized word length of sensor;
2. Increase the sampling frequency. When the noise is white noise, increase the sampling frequency, and the noise spectrum will be doubled in the distribution range. The noise single-edge Power Spectral Density (PSD) is twice the ratio of noise power to sampling frequency. When the signal bandwidth \(B < f_s/2\), the quantization noise is evenly distributed in \(f_s/2\), but the quantization noise in the useful signal spectrum becomes smaller. The higher the sampling frequency is, the more dispersed the quantization noise distribution, and the higher the measured \(SNR^m\) will be.

**Theorem 1:** To a certain extent, \(P(\text{er})\) decreases as the sampling frequency increases.

**Proof:** As the sampling frequency \(f_s\) increases, the quantization noise power spectral density decreases, and noise power \(n^2\) distributed in the useful signal spectrum decreases. The smaller the data fluctuation, the smaller the second-order moment \(D(X)\) of the data sequence \(X\). The data probability distribution curve will be more centralized. At the same time, in the quality model in the literature [6], according to definition of network-signal-to-noise \(SNR^a\), the increase of noise power is mainly reflected in the loss of network data.

With the increase of the sampling frequency, the precision resolution of the sensor is improved and the SNR will be increased. However, a large amount of energy consumption due to frequent sampling leads to the decrease of network-lifetime [12].

3.3.2. Unity and Limitation between Data Quality Accuracy and the SNR model. Based on the signal-to-noise ratio quality model, in this paper, quality assessment management method is introduced. We define the data accuracy metric and give the expression with network characteristic parameters. In that paper, the authors point out that optimizing the quality at the end user can improve the accuracy of the data but do not adequately analyze the relationship between the accuracy of the data and the quality of the sensor block \(q_e\), which is defined as the product of the measurement quality (directly related to measured-signal-to-noise ratio) and the network quality (directly related to network-signal-to-noise ratio). It can be described as \(q_e = \frac{SNR^e}{SNR^m}\), where, \(SNR^e\) is the SNR (\(e\) is the acronym of network) required from that sensor, according to authors’ description and assignment \(\beta = l \cdot 10^{-10}\):

\[
SNR^e = \begin{cases} 
SNR^m & \beta \leq 1 \\
-10 \cdot \log(l) & \beta \gg 1
\end{cases}
\]

(12)
When $\beta \leq 1$, the measured signal-to-noise ratio is constant, there is almost no network packet loss. When $\beta \gg 1$, it is almost only related to the packet loss rate $l$, and the $SNR^e$ received at the end user has little to do with the change of the measured signal-to-noise ratio $SNR^m$ of the sensor. A small amount of packet loss rate can make a significant change to $SNR^e$.

When the sensor block requires higher quality for application, set the sensor block, and the of each sensor should be higher. If the sampling signal power is constant, according to Equation (10), the input noise is small after sampling, that is, the sampling frequency should be increased. When the channel condition and the link-level packet loss rate is constant and the network is in dormant state then restarts [13] in smooth operation, there is almost no node-level packet loss. When there is data congestion in the busy network, the node-level packet loss rate increases rapidly due to buffer overflow and queue delay, as shown in Figure 1.

As the sampling frequency increases, the network is stable. When it continues to increase, the network packet loss rate rapidly increases to a peak value $a\%$. According to Equation (3), under the data quality assessment method of this paper, the accuracy of the sensor data $q_a$ will increase with the growth of sampling frequency. When the network packet loss rate does not reach the $f_L$, which is the threshold frequency that congestion could occur. Therefore, under the conditions of $\gamma = (0, f_L)$, the relationship between sensor block quality $q_i$ and accuracy $q_a$ can be expressed as follows:

$$q_a \propto q_i, \quad f \in \gamma$$

And the qualitative relationship between $q_a$ and $q_i$ is shown in Figure 2.

**Figure 1.** Packet loss rate and sampling frequency  \hspace{1cm} **Figure 2.** Data accuracy $q_a$ and $q_i$

**Theorem 2:** Increasing the sampling frequency $f_s$ will increase the data accuracy $q_a$.

**Proof:** According to theorem 1, when the sampling frequency $f_s < f_L$, the measured-signal-to-noise ratio $SNR^e$ increases, and the coarse data ratio $P(err)$ decreases. Meanwhile, the network packet loss rate $l$ increases. Combining with Equation (3), the data accuracy $q_a$ will be greater.

However, as the sampling frequency continues to increase until the network packet loss comes up significantly at the node level. In other words, when $f_s > f_L$, the network packet loss rate increases uncontrollably, the non-null factor $\alpha = 1 - l$ decreases linearly, and the accuracy $q_a$ continues to increase. But with the packet loss rate grows significantly, a small amount of packet loss can result in the decline of receiving signal-to-noise ratio, which increases the system communication load and the delay of data to the destination. Unfortunately, the data timeliness will also be cut down.

**4. Experiment**

In Section 3, the data quality accuracy is analyzed in accordance with the signal-to-noise ratio quality model based on data assessment in data quality management when the network state is stable. The property of $q_a$ with the sampling frequency $f_s$ is summarized. In this paper, agricultural environmental data got by the Lora IoT platform [14] is selected as the non-null data set $X$. The data is collected in the sensor sampling frequency interval where the network is almost non-congested. The specific data is the soil temperature of the crop as shown in Table 2, and given as shown in Figure 3.
Table 2. Experimental data set

| Environmental Parameter | Date       | Sampling Time Bucket | Sampling Frequency |
|-------------------------|------------|----------------------|--------------------|
| temperature             | 2017/7/23  | 6:00~8:50            | 0.1 /min           |

![Figure 3. Experimental data value distribution](image1)

![Figure 4. Comparison of probability distribution](image2)

Afterwards by the calculation in MATLAB 2016a, the probability distribution of raw data set can be expression as follows:

\[ P(x_i) = \exp(-9.4663 + 224.0711 - 1327.9) \]

Performing a two-point uniform extraction and generating random noise into the raw experimental data set, we can obtain the new sampling frequency of processed data: \( f = f_s / 2 \). Then, implement the algorithm above and the other new probability distribution of processed data is expressed as follows:

\[ P(x_i) = \exp(-7.5492 + 1782934 - 1054.8) \]

We can get the following figure for comparison in Figure 4. Curve 1 is the distribution of the raw data when the sampling frequency is high, and Curve 2 is the distribution with the noise interference and the sampling frequency is lower. When the measured-signal-to-noise ratio can increase with the sampling frequency as Theorem 1 and 2, the distribution curve is wider, and the error rate is larger. When the sampling frequency is larger, the distribution curve is more centralized. In the meantime, there is a larger percentage of the data in the central interval. The \( P(e) \) will be smaller. The experimental results are shown in Table 3.

Table 3. Comparison of experimental results.

| Curve   | \( E(X) \) | \( D(X) \) | \( \eta \) | \( P(e) \) |
|---------|------------|------------|----------|----------|
| Curve1  | 11.82      | 0.04       | 0.163    | 0.238    |
| Curve2  | 11.79      | 0.11       | 0.198    | 0.371    |

When the sampling frequency does not exceed the frequency that could obviously causes network congestion, that is, \( f < f_s / 2 < f_L \), there is almost no network packet loss. The non-null factor \( \alpha = 1 - l \) does not change at this time. Assuming \( \alpha \) is equal to 99.999%, data accuracy has improved by 13.3%. Notably, in order to explore that the accuracy quality of the sensor data, we deliberately add noise to the raw data. The accuracy increment 13.3% does not have a high practical significance. When the amount of data is larger, the impact of network parameters on accuracy will be more obvious, calculation of \( P(e) \) will be more accurate. However, massive data calculation and processing requires higher performance on hardware and more time. When the sampling frequency of the sensor rises and is insufficient to allow congestion to occur, the data accuracy is almost determined by the bottom data acquisition layer. The above experiments show that when the sampling frequency is low, the data deviation is relatively large, and when the sampling frequency is large, the deviation is smaller.
certain growth of sampling frequency will reduce the data ratio of coarse data, which also shows that the signal-to-noise ratio of the sensor can improve the accuracy of the data.

5. Conclusion and Future Work
In this paper, the definition and measurement standard of sensor network data accuracy from the perspective of data quality management is discussed. Based on the quality model of signal-to-noise ratio, the relationship between sensor network data accuracy and network parameters is derived. Then the unity and limitations between data accuracy and quality model is analyzed. Finally, the property between the sampling frequency and the accuracy metric is studied. The effectiveness of the model is verified by experiments, which also show that the relationship derived in this paper is correct.

In further work, the data can be graded after wiping out coarse data. The closer to center value the higher data reliability level is. It is worth noting that increasing the sampling frequency optimizes the accuracy of the data, whereas enlarges the delay to sink node, which deteriorates the timeliness. For delay-sensitive network, the data accuracy and timeliness need to be considered synthetically. We will study in detail about the remainder attributes, such as timeliness, consistency, etc. with various communication parameters and more in-depth sensor network data quality control issues.

6. Acknowledgments
This work is supported by National Natural Science Foundation of Hubei Province (NO.2016CFB463) and Key Projects of Science Research Program of Hubei Provincial Ministry of Education (NO.D20151106).

7. References
[1] Zahedi S, Srivastava M B and Bisdikian C 2007 Proc. IEEE Mil. Commun. Conf. November 16-19 2008 San Diego California (Washington, DC: The Institute of Electrical and Electronics Engineers) pp 1–7
[2] Du P, Yang Q, Shen Z and Kwak K S 2016 IEEE Sens. J. 16 7278–86
[3] Bisdikian C, Kaplan L M and Srivastava M B 2013 ACM Trans. Sens. Netw. 9 48 (Preprint gr-qc)
[4] Marjanović M, Skorin-Kapov L, Pripužić K, Antonić A and Žarko I P 2016 J. Netw. Comput. Appl. 59 95–108
[5] Guo H, Pan Z M and Zhou J 2016 J. Natl. Univ. Def. Technol. 38 150–5
[6] Mathew M, Weng N and Vespa L J 2012 Proc. IEEE 31st Int. Performance Computing and Communications Conf. December 1-3 Austin Texas (Washington, DC: The Institute of Electrical and Electronics Engineers) pp 471–7
[7] Lee Y W, Pipino L L, Funk J D and Wang R Y 2009 Journey to Data Quality (Cambridge: MIT Press)
[8] Siregar S and Sani M I 2017 Adv. Sci. Lett. 23 3879–82
[9] Neuhold D, Schmidt J F, Klaue J, Schupke D and Bettstetter C 2017 Proc. 20th ACM Int. Conf. on Modelling, Analysis and Simulation of Wireless and Mobile Systems November 21-25 2017 Miami Florida (New York: ACM Press) pp 137–42
[10] Zheng J P, Han Q T and Zhang H 2013 J. Beijing Univ. Posts Telecommun. 36 110–5
[11] Zeng J F and Teng Z S 2012 J. Electr. Measur. Instr. 26 1096–9
[12] Aalamifar F and Lampe L 2018 IEEE Trans. Commun. (Preprint gr-qc/06656v1)
[13] Jung T, Li X Y, Wan Z and Wan M 2015 IEEE Trans. Inf. Forensic Secur. 10 190–9
[14] Ma Y, Jin J, Huang Q and Dan F 2018 Intelligent Computing Theories and Application (vol 10954) eds D S Huang E et al (Cham: Springer) pp 219–30