Hybrid Context-Aware Method for Quality Assessment of Data Streams

Mostafa Mirzaie(✉)

Ferdowsi University of Mashhad (FUM), Mashhad, Iran
mostafa.mirzaie@mail.um.ac.ir

Abstract. Data quality is one of the most important issues that if not taken into consideration appropriately, results in the low reliability of the knowledge extracted through big data analytics. Furthermore, the challenges with data quality management are even greater with streaming data. Most of the methods introduced in the literature for processing streaming data do not use contextual information for the purpose of addressing data quality issues, however, it is possible to improve the performance of these methods by considering the contextual information, especially those obtained from the external resources. Based on this point of view, our main objective in this thesis is to propose a hybrid multivariate context-aware approach for data quality assessment in streaming environments, such as smart city applications.

Keywords: Data quality assessment · Context awareness · Streaming data

1 Problem Statement and Contributions

According to Statista website[1] report, the total amount of data created in the world reaches 175 zettabytes by the year 2025. This data is generated from various sources including sensors in a smart city platform [1]. However, an important point is that the observations of a sensor might be of insufficient quality due to various constraints, like environmental conditions or hardware malfunctioning [2]. The poor quality data may result in wrong business decisions being made by organizations [3]. Therefore, it is important to find data quality issues and clean poor data before using for any knowledge extraction or decision making. Some researchers use context-aware methods that quality of data is determined not only through analysis of the local application-specific information, but also using information from a global context, which in turn enhances the performance of big data quality management [4]. Based on our recent

[1] https://www.statista.com/statistics/871513/worldwide-data-created/

Supervised by Behshid Behkamal, Samad Paydar (Ferdowsi University of Mashhad (FUM), Mashhad, Iran) and Mohammad Allahbakhsh (University of Zabol, Zabol, Iran).

H. Hacid et al. (Eds.): ICSOC 2020 Workshops, LNCS 12632, pp. 10–16, 2021.
https://doi.org/10.1007/978-3-030-76352-7_2

systematic literature review on big data quality[2], we have observed that although a number of techniques have been proposed in the literature to improve data quality in the big data field, only a few consider contextual information in the process of data quality assessment. Since no context model for big data quality assessment has been proposed, in another study[3] we reviewed context-aware studies to provide a context model for big data quality, according to which we found that in all studies, only internal contexts (available in the subject data set) are used and none of them has considered the external context (available from other data sources). In addition, none of the context-aware techniques have used the stored data to increase the accuracy of quality assessment. In what follows, some of the challenges are mentioned:

- **Variety of arrival rate:** Data values arrives at the different rate, so the evaluation algorithm should provide a mechanism for processing existing data before the arrival of the new incoming data.
- **Infinite:** In streaming data, data is continuously being received, and the evaluation process must be done online and without interruption of the main retrieval process.
- **Volatility:** In data stream, volatility is a significant challenge that data expires after a while and lose credibility, so data processing should be done before it expires.
- **Heterogeneous sources:** Data may be received from different sources, in which case it is necessary to integrate and extract the correlation of these data in order to obtain the appropriate context information.

Based on the discussion above, in this early stage proposal, we intend to present a novel hybrid context-aware method using environmental information to assess the quality of data stream. The novel contributions of our work can be summarized as follows:

- We use external context (related information extracted from other sources), in order to improve data stream quality assessment performance.
- We benefit historical data that enables tracking of data values over time which gives key insights, in order to increase detection precision.
- We propose a grid-based clustering to decrease execution time.

2 State of the Art

In this section, we discuss and compare the context-aware quality assessment studies in streaming data, based on our systematic literature review. Studies have been compared based on several criteria, including level of management, type of contextual information, processing type, variable quantity, and technique used. From the point of view of level of management, there are two approaches: sensor-level, in which all quality controls are performed by the sensors without any interference by users, and user-level, in which pre-processing phase is performed by the user after receiving data from the sensors. From another point of view, we have classified these methods into two groups based on the type

[2] State of the Art on the Quality of Big Data: A Systematic Literature Review and Classification Framework.

[3] Contextualization of Big Data Quality: A framework for comparison.

of contextual information they use for detecting quality deficiencies. The first group has used intra-network information, while the later exploited external resources information. In terms of processing type, stream data quality assessment studies are divided into two categories of online methods, which have evaluated data in a time frame i.e., window, and hybrid methods that have used stored data for the quality assessment of real-time data. We have also categorized the studies into two groups of univariate and multivariate in terms of the number of features used for analysis. The last feature for comparing studies is the technique used for quality evaluation. Some methods calculated the similarity or distance between all data values to consider data that is far from the threshold as poor quality data. These methods are classified as distance based methods. The other techniques that attempt to obtain normal distribution through time analysis and model building, are called model-based, and finally, studies that use clustering or classification techniques, are named as learning based techniques. The advantage of model-based methods is that they have less computational complexity than other methods and are more suitable for online processing. Distance-based and learning-based methods are not recommended for online processing [5, 6]. The summary of this feature based evaluation is presented in Table 1.

Table 1. Feature based comparative evaluation

Ref	Year	Level of management	Contextual information	Processing type	Number of variables used	Technique
[7]	2012	Sensor level	Internal	Online	Univariate	Model based
[8]	2013	Sensor level	Internal	Online	Univariate	Distance based
[9]	2013	Sensor level	Internal	Online	Univariate	Learning based
[10]	2014	Sensor level	Internal	Online	Univariate	Distance based
[11]	2015	User level	–	Hybrid	Univariate	Model based
			Internal	Online	Multivariate	Model based
[12]	2018	Sensor level	Internal	Online	Univariate	Learning based
Proposed method	2020	User level	Internal and external	Hybrid	Multivariate	Learning based

As illustrated in Table 1, most of the methods control the quality of data at the sensor level. This means that the user has no role in data pre-processing. Given the limitations of processing and memory in sensors, this can affect the performance of such method.

In terms of contextual information, only one study [11] took the advantage of the stored data for quality assessment using two algorithms. In the first algorithm, the stored data are used to create the model without including any contextual information; while in the second algorithm, which benefitted contextual data, only window-based data are considered. Thus, it can be concluded that no hybrid processing method has been applied in these context-aware methods. Furthermore, most of the studies have not supported multivariate analysis. Advantages of multivariate analysis include an ability to glean a more realistic picture than looking at a single variable. On the other hand, hybrid processing takes advantage of historical data, which will increase the accuracy of the method.

In this project, we are going to use contextual information as well as historical data to increase the accuracy of quality assessment of streaming data.

3 Proposed Approach

In this proposal, we intend to present a hybrid multivariate context-aware approach to assess quality of data streams. The proposed method consists of two main steps, first the data is processed and the data distribution model is obtained in an offline manner, then using the model, the data quality assessment is performed in online mode. These phases are explained in detail as follows.

3.1 Constructing Model Using Contextual Information (Offline)

The aim of the first step in to extract the data distribution model using contextual information. Model-based techniques find poor quality data faster and are therefore suitable for online processing [6]. This phase consists of three steps including data preprocessing, grid-based clustering, and model construction using multiple regression.

Data Preprocessing: Firstly, each historical data value (historical data of data value that should be evaluated in terms of quality) and contextual information (historical data of contextual information that helps the quality assessment process) are stored in a data lake. After eliminating useless information, all the required data values is recorded in a log to be clustered. Moreover, in this step, a domain expert sets a predefined threshold in order to detect low quality data in online phase. Suppose that the threshold is defined as 5 and the input value is 20 while the correct value is 24. Since, the difference of input value and the correct value are less than the desired threshold, the input value is considered as an accurate data value.

Grid-Based Clustering: All historical data values on contextual information must then be clustered in terms of a common feature (e.g. location in geo-location data). For this purpose, this historical information is clustered with a grid-based clustering. The great advantage of grid-based clustering is its significant reduction of the computational complexity, especially for clustering very large data sets.

Model Construction using Multiple Regression: As soon as the clustering is complete, the data distribution model is extracted, for each cluster. In each cluster, information is modeled based on multiple regression which is an extension of simple linear

regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. This process continues until the data distribution model is obtained in all clusters. After this process, these models are used in the online quality evaluation step. In this proposal, we do not take concept drift, i.e. changes in the data distribution over time, into consideration. Consequently, it is not needed to update model during evaluation time and this will be considered as future work.

3.2 Assessing Quality of Streaming Data (Online)

After creating a distribution model for each cluster, by observing new data streams, the online quality assessment process begins, which has several steps including assigning data streams to the desired cluster, recording contextual information in the corresponding log, obtaining the distribution function and the predicted value, detecting and cleaning poor data quality.

Assign Data Streams to the Desired Cluster: At this stage, if the number of clusters is large, the allocation of data streams to the corresponding cluster will be time-consuming, and on the other hand, if the number of clusters is small, the accuracy of the distribution function will be reduced. As explained before, the proposed method will be able to make a trade-off by providing a grid-based clustering algorithm. In each of the first depth clusters, there is metadata for faster search and allocation of data to the desired cluster. Therefore, using these metadata, the data streams and their contextual information are quickly allocated to the desired cluster.

Record Contextual Information in the Corresponding Log: When each of the given data values is assigned to the desired cluster, the final log for that time interval must be updated with the new data values. This will continue until the data value, which should be evaluated in terms of quality, is seen, the log record is closed.

Obtain the Distribution Function and the Prediction Value: After closing the data log, the distribution function obtained from the first step (offline processing) is calculated with the values of the last log record and the output of the function is considered as the prediction value.

Detect Poor Quality Data and Clean It: After calculating the function and the prediction value, this value is compared with the data value. The threshold value is specified by the expert, which determines the difference between the predicted value and the data value. If the data value is within the predicted value range, the data is normal, otherwise the data is of poor quality and must be improved and replaced with predicted value.

4 Evaluation Plan

To support our claim, we will theoretically and empirically evaluate our proposed approach. In theoretical evaluation, we will compare our method with previous methods

presented in Table 1 in terms of both accuracy and performance. For empirical evaluation, we develop an automated tool to measure the quality of data values of input datasets using both historical and contextual data. The details of experimental datasets are presented in Table 2. These datasets are available at Chicago city data portal[4] and all are related to past three years. A common feature in all datasets is geographic location, so we are able to find all contextual data and filter out low-quality data based on location.

Table 2. The details of the datasets used in our experiments

Datasets	No. of records	No. of features	Contextual features
Chicago traffic tracker	119M	22	Main dataset enriched with location
Chicago traffic crashes	417K	49	Date, time, location
Roadway construction events	17.5K	8	Start and end time, location
Public health department events	494	13	Time, location
Chicago weather	API	10	Weather type, visibility, wind speed

5 Conclusions

The goal of this research is to propose a hybrid multivariate context-aware data quality assessment method for data streams. Although many methods have been proposed to improve the quality of streaming data, none of them have used external contextual information. In this thesis, we are going to use external contextual information as well as historical data in order to improve the performance of the method. The proposed method will be evaluated on real datasets.

References

1. Perez-Castillo, R., et al.: DAQUA-MASS: an ISO 8000–61 based data quality management methodology for sensor data. Sensors **18**(9), 3105 (2018)
2. Bu, Y., Chen, L., Fu, A.W.-C., Liu, D.: Efficient anomaly monitoring over moving object trajectory streams. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 2009, p. 159 (2009)
3. Sidi, F., Panahy, P.H.S., Affendey, L.S., Jabar, M.A., Ibrahim, H., Mustapha, A.: Data quality: a survey of data quality dimensions. In: Proceedings of 2012 International Conference on Information Retrieval and Knowledge Management CAMP 2012, pp. 300–304, June 2014
4. Ardagna, D., Cappiello, C., Samá, W., Vitali, M.: Context-aware data quality assessment for big data. Future Gener. Comput. Syst. **89**, 548–562 (2018)

[4] https://data.cityofchicago.org/.

5. Anusha, A., Rao, I.S., Student, M.T.: A study on outlier detection for temporal data. Int. J. Eng. Sci. Comput. **8**(3), 16354–16356 (2018)
6. Chen, L., Gao, S., Cao, X.: Research on real-time outlier detection over big data streams. Int. J. Comput. Appl. **42**(8), 1–9 (2017)
7. Zhang, Y., Hamm, N.A.S., Meratnia, N., Stein, A., van de Voort, M., Havinga, P.J.M.: Statistics-based outlier detection for wireless sensor networks. Int. J. Geor. Inf. Sci. **26**(8), 1373–1392 (2012)
8. Iyer, V.: Ensemble Stream Model for Data-Cleaning in Sensor Networks (2013)
9. Zhang, Y., Meratnia, N., Havinga, P.J.M.: Distributed online outlier detection in wireless sensor networks using ellipsoidal support vector machine. Ad Hoc Netw. **11**(3), 1062–1074 (2013)
10. Zhang, Y., Szabo, C., Sheng, Q.: Cleaning environmental sensing data streams based on individual sensor reliability. In: Benatallah, B., Bestavros, A., Manolopoulos, Y., Vakali, A., Zhang, Y. (eds.) WISE 2014. LNCS, vol. 8787, pp. 405–414. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11746-1_29
11. Hayes, M.A., Capretz, M.A.M.: Contextual anomaly detection framework for big sensor data. J. Big Data **2**(1), 1–22 (2015). https://doi.org/10.1186/s40537-014-0011-y
12. Rassam, M.A., Maarof, M.A., Zainal, A.: A distributed anomaly detection model for wireless sensor networks based on the one-class principal component classifier. Int. J. Sens. Netw. **27**(3), 200 (2018)