Towards an optimization model for outlier detection in IoT-enabled smart cities

Moulay Lakbir Tahiri Alaoui, Meryam Belhiah and Soumia Ziti

June 26, 2022

Towards an optimization model for outlier detection in IoT-enabled smart cities

Moulay Lakbir Tahiri Alaoui, Meryam Belhiah, Soumia Ziti

Mohamed V University in Rabat, Rabat, Morocco

lakbir.tahiri@um5r.ac.ma
m.belhiah@um5r.ac.ma
s.ziti@um5r.ac.ma

Abstract. In a connected world, the growing attention given to the IoT (Internet of Things) is driven by its economic, societal, and ecological impact among others as well as its vast applications and services. The decisions made by those applications and services are based on the data gathered from different networks of IoT sensors. A poor quality of data forwarded to control centers may lead to ill-informed decisions, inadequate services and impact adversely the business objectives.

In this paper, parameters that influence the levels of data quality (DQ) will be addressed. These problems may be due to errors in measurements or precision of the data collection devices, energy restrictions, intermittent connectivity, interference with other devices, sampling frequency, noisy environment, and data volume among others, DQ levels are evaluated against a set of dimensions. Herein, we will focus our research on the most important dimensions for end users, such as accuracy, completeness and timeliness.

As outlier detection (OD) is a major problem in both IoT and Data Quality, it will also be addressed as a sub-domain for DQ. OD is actually a complex matter since received data which seems to be abnormal can be a normal behavior. In this case, what is expected to be an outlier may be a valuable information that should not be discarded like in health diagnosis. Many techniques and methods are used for OD, each one is used for specific domains. Techniques for OD will be presented and classified. We will describe the most used techniques such as statistic-, distance-, density-clustering- and learning based methods. A technique to detect outliers in the field of IoT-enabled smart cities will be recommended

Keywords: Internet of Things, Data quality, Outlier Detection

1 Introduction

Data quality levels are evaluated against a set of dimensions. It is impacted by different IOT layers starting from physical layer to network and application layer.

Interest will be put in data gathered from different cars with IOT sensors having a GPS device to detect their geolocation. A model will be developed to treat this data to detect outliers. Once the model is validated it will be possible to treat other kind of IOT data. A detailed description of a technique to detect outliers in the field of IoT-enabled smart cities will be recommended.

A study of needed theorycal part will be treated in the first chapters. The most important dimensions for end users, such as accuracy, completeness and timeliness will be described. Many techniques and methods are used for OD, each one is used for specific domains. Techniques for OD will be presented and classified. We will describe the most used techniques such as statistic-, distance-, density- and clustering.

For each method advantages will be presented, drawbacks and also a method that's using this procedure and recommend field for witch the method is suitable. We will finally present the future research direction in the domain of IOT, smart cities and data quality.

2 IOT data quality domain

2.1 IoT Paradigm: layers and enabling technologies

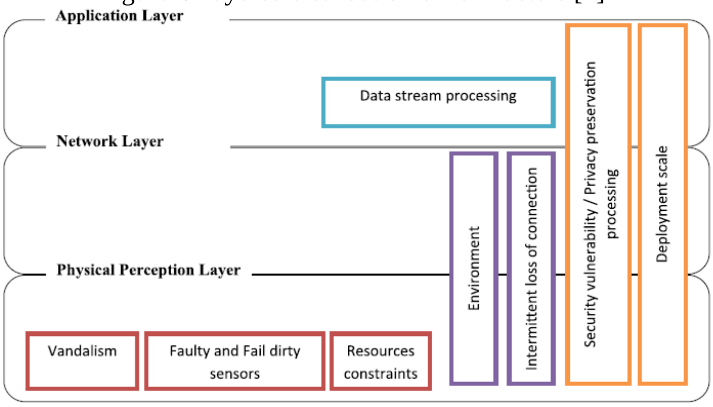
The IOT concept was coined by a member of the Radio Frequency Identification (RFID) development community in 1999, and it has recently become more relevant to the practical world largely because of the growth of mobile devices, embedded and ubiquitous communication, cloud computing and data analytics[1]. Internet of things can be defined as: Internet of things (IOT) is a network of physical objects. The internet is not only a network of computers, but it has evolved into a network of device of all type and sizes , vehicles, smart phones, home appliances, toys, cameras, medical instruments and industrial systems, animals, people, buildings, all connected ,all communicating & sharing information based on stipulated protocols in order to achieve smart reorganizations, positioning, tracing, safe & control & even personal real time online monitoring , online upgrade, process control & administration[2].
Internet of Things (IoT) is a concept and a paradigm that considers pervasive presence in the environment of a variety of things/objects that through wireless and wired connections and unique addressing schemes are able to interact with each other and cooperate with other things/objects to create new applications/services and reach common goals. In this context the research and development challenges to create a smart world are enormous. A world where the real, digital and the virtual are converging to create smart environments that make energy, transport, cities and many other areas more intelligent. [2].
The goal of the Internet of Things is to enable things to be connected anytime, anyplace, with anything and anyone ideally using any path/network and any service.

2.2 Data quality definition *and issues*

"In the OMB (Office of Management and Budget) guidelines data quality is defined as an encompassing term comprising utility, objectivity, and integrity." [3]
Data Quality may be defined also as "data quality is more than simply data accuracy. Other significant dimensions such as completeness, consistency, and currency are necessary in order to fully characterize the quality of data."[3]

Data harvested from different and heterogeneous IOT objects spread all over the world suffers from different kind of problems which impact its quality. Usually, sensors are placed in pairs to work as master/slave to control data quality and to offer a continuous data generator in case of a device failure.
In this chapter most common problems impacting data will be addressed, in general the issues may be related to the 3 IOT layers shown in Figure1, Physical Layer, Network Layer or Application Layer.

Figure1: Layered distribution of IoT factors [4]


Application Layer
Network Layer
Physical Perception Layer
Data stream processing
Security vulnerability / Privacy preservation processing
Deployment scale
Environment
Intermittent loss of connection
Vandalism
Faulty and Fail dirty sensors
Resources constraints

- Problems related to physical layer:
  - Maintenance: the sensors may be placed in inaccessible terrains like mountains or Sahara to measure needed insight, (reach out, replacing and maintaining such devices is difficult and costy).
  - Vandalism: devices are located in public and accessible spaces, thus are not in secured places and are an easy prey for all kinds of vandalism.
  - Faulty devices: A faulty sensor may still generating unsound data which will impact the whole process of proving services and generating business insights.
  - Sensors constraints: being low cost, sensors are not generally of the best quality and their capacities are limited: this regards all capacities as low battery, connectivity, calibration errors.
- Problems related to network layer:
  - Noise: climatic conditions are impacting signal frames received by the sensors
  - Security: ciphering gathered data increases the amount of transmitted packets and computational process/time for encrypting and decrypting transmitted frames which may delay real time insight.
  - Intermittent connectivity: unreachable devices during data connection frames impacts the continuity of insight changes
  - Interference: with the huge number of heterogeneous devices transmitting in the same time signals are interfering which reduces received signal quality.
- Problems related to application layer:
  - Heterogeneity: tremendous amounts of data are gathered from different types of devices in different places of the world which affect the interpretation of the frames.

2.3 *Data quality Dimensions*

Data of low quality deeply influences the quality of business processes [3], measurement of DQ is done using a number of dimensions, below the most important ones:

Accuracy: is the most important in IOT [5], it is used to measure how can an observation reflects the real world. It’s difficult to detect faults related to accuracy since data control.

Timeliness: The IoT data is considered timely when an observation of an object was updated at a desired time of interest [6].

Completeness: The extent to which all expected data is provided by IoT services [6] (Availability, Missing data)

Below more definitions and details about those dimensions Table1:

Table1 Data quality dimensions description

Accuracy	Data are accurate when data values stored in the database correspond to real- world values [7,8].
	The extent to which data is correct, reliable and certified [9].
	Accuracy is a measure of the proximity of a data value, v, to some other value, v', that is considered correct [10,8]
	A measure of the correction of the data (which requires an authoritative source of reference to be identified and accessible [21].
Completness	The ability of an information system to represent every meaningful state of the represented real world system [7, 8].
	The extent to which data are of sufficient breadth, depth and scope for the task at hand [9].
	The degree to which values are present in a data collection [10, 8].
	Percentage of the real- world information entered in the sources and/or the data warehouse [11, 8]
	Information having all required parts of an entity's information present [12, 8].
Timeliness	The extent to which age of the data is appropriated for the task at hand [9].
	The delay between a change of a real world state and the resulting modification of the information system state [13, 8].
	Timeliness has two components: age and volatility. Age or currency is a measure of how old the information is, based on how long age it was recorded. Volatility is a measure of information instability the frequency of change of the value for an entity attribute [12, 8].

Trade-offs Between Dimensions

Data quality dimensions are not independent, with their correlation if a dimension is more important for a specific application, favoring it twill impact negatively the other dimensions,

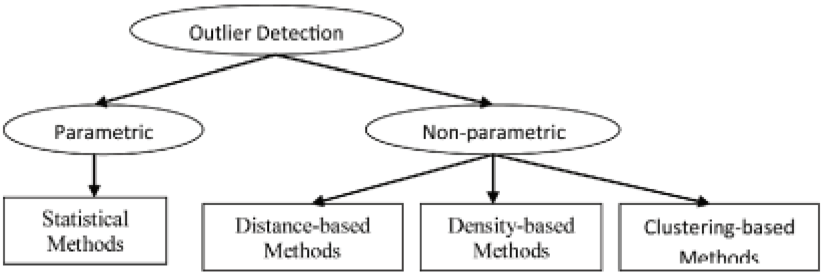
Data quality dimensions are not independent, i.e., correlations exist between them. If one dimension is considered more important than the others for a specific application, then the choice of favoring it may imply negative consequences for the other ones [3].

2.4 *Outlier detection: definition and techniques*

An outlier is a data that deviate too much from normal behavior. Unexpected values may be considered as measurement errors, sampling errors, but they also can correspond to true values, once the insight is not of a big interest, the outlier may be deleted, considered as a lost data value, an alarm or a warning may be generated but also the data maybe an interesting insight that should be exploited.

Boundaries between normal and abnormal values is thin, for that reason many studies were performed to detect outliers and many techniques were applied...
Many techniques that use different methods and algorithms were proposed to address OD (Figure 2), the different methods can be split to categories, for each category, we will present a description, challenges with the pros and cons.

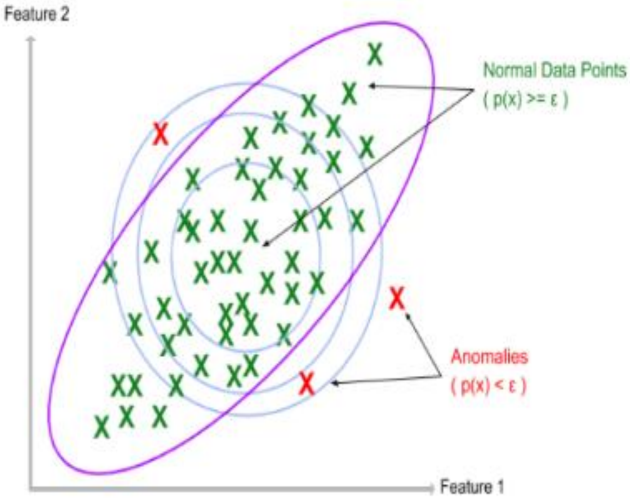
Figure 2: Outlier detection methods [14]



- **Statistical-based method**: used to detect sensors faults and outliers once the gathered data follow a known model in case of parametric method. In this case the measurement values are immediately compared with the model and are evaluated as outlier once the data is different than expected values (Figure 3). For some statistic methods the model is not known we talk about non-parametric [14]

For instance, Kernel Density Estimation is using this method to detect outliers:

Figure 3: Outlier statistic-based method. [15]

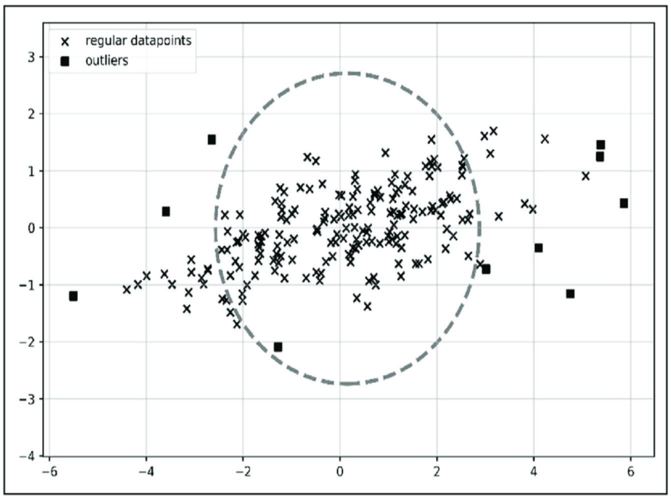


For a given distribution model, this is an efficient method and can be easily implemented, with a low computational cost (does not need many resources regarding the computational program, CPU, execution time...). In the real world, the distribution is not always known and the method cannot be applied.

- **Distance-based method**: The underlying principle of the distance-based detection algorithms (Figure 4) focuses on the distance computation between observations. [16], we can list the method based on distance threshold and also KNN k_nearest neighbor.[16]

This is a more realistic method since it does not depend on the distribution, it scales for large data comparing to statistical methods but it is expensive in multivariate and high dimensional data.

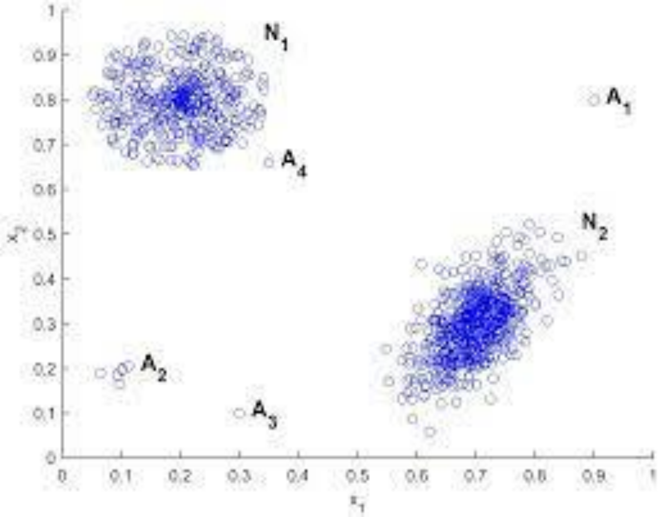
Figure 4: Outlier distance-based method [17]



- **Density-based method:** In this method an inlier is found in a high density region while an outlier is located in a low-density region (Figure 5). This method is effective but remain unsuitable for large data[18] ;

We can list LOF(local outlier Factor) as one of the known programs using this technique. [19]

Figure 5: Outlier density -based method [20]



There is other techniques for outlier detection like ensemble-based and learning based method. The choice of a method depends on many parameters like cost, computational complexity, number of dimensions to take into consideration, the volume of the data. Each method can be effective for a domain but not for the others.

Other techniques for OD exists like learning-based approaches which have been applied in many fields, especially in machine learning and deep learning, this method interact with the user and is efficient in detecting outliers despite its computational cost especially once data size grows. Following table2 presents a summary of different methods with their advantages and drawbacks

Table2: Outlier methods advantages and drawbacks

Based Method	Advantages	Drawbacks
Statistic	effective for given distribution models	cannot be applied when this distribution is unknown
Distance	does not depend on data distribution	expensive in multivariate and high dimensional data
Density	more effective	remains unsuitable for large data
Cluster	can handle data streams	needs too many parameters

3 Towards a model for outlier detection in IoT-enabled smart cities

The application Layer in IOT is growing very fast with a big number of applications related to smart living, smart energy, smart home, smart health, smart industry, smart transport…etc.

This work will concentrate on data gathered from IOT devices having a GPS (global positioning system) module. As a first step this data will be used to develop a model for vehicle geolocation in real time. Once the model is validated it will be generalized to other types of data.

3.1 Data quality for geolocation services

Data gathered from geolocation suffers from the same problems as other data coming from previous sensors, “exact localization may be impossible, e.g., due to nodes lacking Global Positioning System (GPS) access for reasons of cost, energy, or signal unavailability.” [22]

Noise and errors that may be present on geolocation data which needs to be filtered and corrected as a pretreatment phase. The computational program should include the identification of vehicles.

Data may also contain unsound positions which should be detected and corrected. Based on methods studied in the previous chapters, below a procedure that can control received data to detect outliers

3.2 Recommended method for Geolocation:

Suppose a vehicle/machine A is moving (being IN a moving state will require to have a minimum speed ($speed_{min}$).

Once a location is received about this machine being in (abscissa, ordinate axes and time) C1 (X1, Y1,t1).

The next position received in t2 (with a speed speed1) should necessary be in a position C2 (X2,Y2,t2) with a distance that cannot exceed speed*(t2-t1).

The speed being a continuous function and cannot be more than $Speed_{max}$ allowed, with a sample taken at a time intervals the distance C1C2 cannot exceed $Speed_{max}$ *(t2-t1) and also cannot exceed 1.5*speed1*(t2-t1).

Once a car is not in movement (for example, with a red light) its speed changes much with the green light. Which is different in case of a car in movement, if the speed at a time t_n is 60km/s the speed cannot change to 90km/s for the next sample especially if the sampling period is short.

To be efficient the parameter speed_new_value should not exceed 1.5*$speed_{old\_value}$ which is granted since the vehicle in movement has a speed greater than $speed_{min}$.

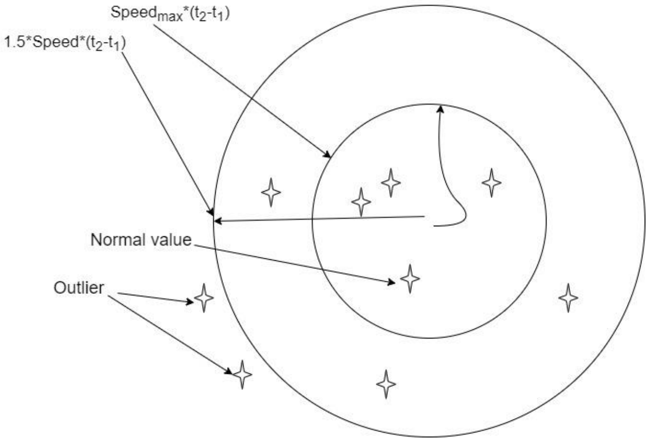
The direction may change during the movement.

With all this conditions the new location will necessarily be in a circle having C1 as a center and C1C2 as a radios.

We can conclude also that C1C2 < min ($Speed_{max}$ *(t2-t1), 1.5*speed1*(t2-t1)).

All values in the circle are considered as acceptable values, others are outliers as shown in figure 6.

Figure 6: Outlier detection in case of Geolocation



Conclusion

Data quality is still a fertile field for which much research is still ongoing, especially with the fast evolution in all telecommunication subdomains (sensors, transport speed, latency, computational programs, CPU...)
The choice of a method to detect outliers depend on many parameters which makes a suitable method for a domain not valid for a different field.

Now a days we can take benefits from 5G and 6G services like mMTC(massive machine type communications) In 5G and mMTC+ in 6G for smart building, smart healthcare, Smart services enabled by UAVs(unmanned aerial vehicle) and wide-range of IOT services.
It is important to underline the cost of such services and the reduced coverage for this new technology.

References

1. Patel, Keyur K., and Sunil M. Patel. "Internet of things-IOT: definition, characteristics, architecture, enabling technologies, application & future challenges." International journal of engineering science and computing 6.5 (2016).
2. Vermesan, Ovidiu, and Peter Friess, eds. Internet of things: converging technologies for smart environments and integrated ecosystems. River publishers, 2013.
3. Carey, M. J., et al. "Data-Centric Systems and Applications." Italy: Springer (2006).
4. Karkouch, Aimad, et al. "Data quality in internet of things: A state-of-the-art survey." Journal of Network and Computer Applications 73 (2016): 57-81.
5. Liu, Caihua, et al. "Data quality and the Internet of Things." Computing 102.2 (2020): 573-599.
6. Li F, Nastic S, Dustdar S (2012) Data quality observation in pervasive environments. In: Proc the 15th Int Conf Comput Sci Eng, IEEE, pp 602–609
7. Ballou D P, and Pazer H L (1985). Modeling data and process quality in multi-input, multi-output information systems, Management Science, vol 31(2), 150–162
8. Batini C, Cappiello C et al. (2009). Methodologies for data quality assessment and improvement, ACM Computing Surveys, vol 41(3), 1–52, doi:[10.1145/1541880.1541883].
9. Wang R Y, and Strong D M (1996). Beyond accuracy: What data quality means to data consumers, Journal of management information systems, vol 12(4), 5–33
10. Redman T C, and Blanton A (1997). Data quality for the information age, 1st Edn., ACM Digital Library, Artech House, Inc., Norwood, MA, USA, [ISBN:0890068836].
11. Jarke M, Lenzerini M et al. (2003). Fundamentals of Data Warehouses, SIGMOD record, Springer-Verlag, vol 32(2), 55–56. [ISBN: 3-540-42089-4].
12. Bovee M, Srivastava R P et al. (2003). A conceptual framework and belief-function approach to assessing overall information quality, International Journal of Intelligent Systems, vol 18(1), 51–74.
13. Wand Y W (1996). Anchoring data quality dimensions in ontological foundation, Communication of the ACM, vol 39(11), 86–95.
14. Suri, N. N. R. R., M. Murty, and G. Athithan. "Data mining techniques for outlier detection." Visual analytics and interactive technologies: data, text and web mining
15. Raju, Anuradha Samkham, Md Mamunur Rashid, and Fariza Sabrina. "Performance Enhancement of Intrusion Detection System Using Machine Learning Algorithms with Feature Selection." 2021 31st International Telecommunication Networks and Applications Conference (ITNAC). IEEE, 2021.
16. Wang, Hongzhi, Mohamed Jaward Bah, and Mohamed Hammad. "Progress in outlier detection techniques: A survey." Ieee Access 7 (2019): 107964-108000.
17. Erharter, Georg H., and Thomas Marcher. "MSAC: Towards data driven system behavior classification for TBM tunneling." Tunnelling and Underground Space Technology 103 (2020): 103466.
18. Smiti, Abir. "A critical overview of outlier detection methods." Computer Science Review 38 (2020): 100306
19. Wang, Hongzhi, Mohamed Jaward Bah, and Mohamed Hammad. "Progress in outlier detection techniques: A survey." Ieee Access 7 (2019): 107964-108000.
20. Chepenko, D. "A Density-based algorithm for outlier detection." (2018).
21. McGilvray D (2008). Executing data quality projects: Ten steps to quality data and trusted information. Morgan Kaufmann, Elsevier, Barlington, MA, USA. [ISBN: 978-0-12-374369-5].
22. Grey, Michael, et al. "Towards distributed geolocation by employing a delay-based optimization scheme." 2014 IEEE Symposium on Computers and Communications (ISCC). IEEE, 2014.