DIFFICULTIES WITH MISSING DATA IN DIFFERENT APPLICATIONS

Vivek Thoutam[1]

[1]Software Developer, Otsuka Pharmaceutical Co., Ltd, Princeton, New Jersey, USA

ABSTRACT

In IoT we collect data from multiple sensors and process them using data processing workflows and transform these data as per the requirements of the application. In the process of transforming the data if any one of the sensors is unable to generate the data due to environmental / technical problem, how can we handle the continuation of data which is to be sent to applications without making the applications to wait for the recovery of sensor. That is, how can we minimize the time gap of retrieving the loss data from sensor. It's like in networks if a packet is lost in internet the destination system asks for retransmission of packet, which consumes some amount of time. Instead of that we should have a mechanism which can handle the lost data from sensors. We discuss some applications of IoT and the impact of loss of data.

Index Terms: Linear Discriminant Analysis, multi sensor structure, Data fusion.

I. INTRODUCTION

IoT is a platform where it contains the global network with connected devices, which collect data and share for to be used by the applications. A formal definition is Internet of Things (IoT) is a growing network of physical objects and devices, called "Things," as well as individuals.

IoT enables many sensors to interconnect with each other for transmitting the data without human intervention, due to which it influencing the nations by various applications like smart cities, smart meter, smart home, healthcare monitoring systems, intelligent cars, smart manufacturing plant, and real-time traffic monitoring, air quality detection in environment, forecasting applications etc. [1]

Every one of the above applications have a unique thing to achieve, based on the domain application need sensors are deployed and they are called as sensing applications mentioned in figure 1 which have the ability of sensing the devices that are associated with the sensors to monitor by capturing the data.

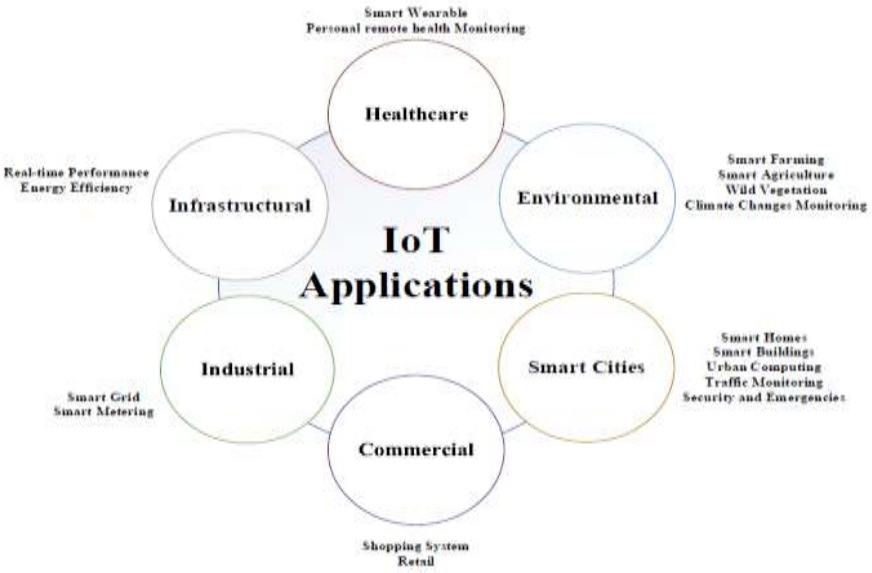


Figure 1: IoT Applications in different fields

There are different factors which makes IoT applications to be efficient are low-cost sensors, error tolerant in data communication with high speed. In real time both the devices and sensors together produce and transmit information which will be not efficient in case of incorrect or insufficient for data processing.

Rest of the paper is organized as follows 2 applications are discussed, where loss of data is not an issue for GPS navigation system on the other hand health care systems in medical field the loss of data

matters and we have shown with a simple case study of regression analysis to predict home ownership based on age and educational background and Pima Indians Diabetes Dataset analysis.

II. GPS NAVIGATION SYSTEM

Detection of vehicles provides a significant benefit like safety management, security and traffic control etc[2]. GPS navigation system is popularly used in today world for to get the location information in the form of latitude and longitude along with other satellite information.

In general way one need to have internet connection on their device in order to finds the user address as its don't have an alternative way to access a geocoder but now one can still obtain their latitude and longitude values without needing a connection.

We consider the real time steps which takes place when the accident is found. With the help of GPS vehicle tracking, we able to track the vehicle of a person. In the real time scenario, the vehicle can be detected using different things like 1. Using GPS sensor 2. Using GSM (mobile network services) 3. Application which uses backend server

When we want to use the location service GPS location sensor will be activated. If at all GPS is turned in active then the mobile device uses GSM network provider for to detect the user's location in figure 2 .in this way user can be tracked by either sending their location information through network socket to another user. By using some mobile application also one can upload their location information to a server where another user can query their location information.

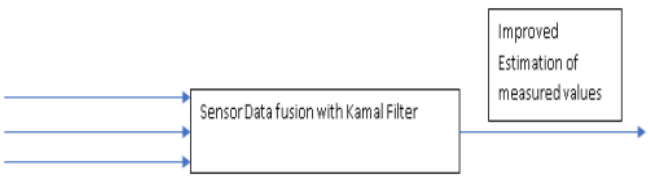
Figure 2: GPS, GSM and server usage for detecting the location

In the above process one can track the object because if one of the methods fail one can use the alternative method to track the object.

Let's take for example to describe how multi-sensor structures provide a better solution than single sensor structures.[7]

In navigation to find the location of a target variable is most effectively done by using a data fusion.

Data-fusion-based or hybrid systems is to increase accuracy of location estimation by combining information gathered from different sensors. we need some tools to combine sensor information and Kalman filter is one of the basic tools used in data fusion.
There are mainly two filter structures used in data fusion which are centralized and decentralized[4]. When we are able to read more than one sensor information, Kalman Filter is a useful algorithm to combine all data gathered from different sources. Filtering can be done in, both, centralized and decentralized structures. We described centralized mechanism.



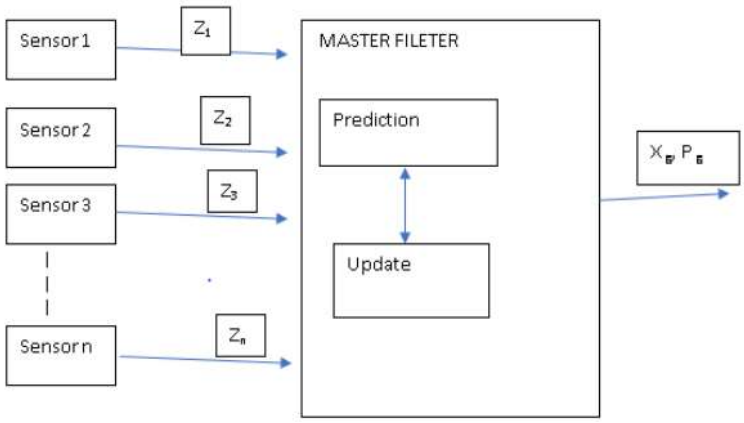


Figure 3: Centralized Kalman filter structure

$Z_i(K)$ is measurement information from different sensors, P_g representing the uncertainty of the estimates.

Now extending these ideas where if a person met into accident, the immediate thing one can do is calling ambulance, in ambulance even though basic treatment is done still the condition of the patient can only be brought to normal level by treatment of the doctor. So, in these processes the condition of the patient can be updated to the hospital team with a prior information about the accident. So that the arrangements can be made for protecting the health of the person.

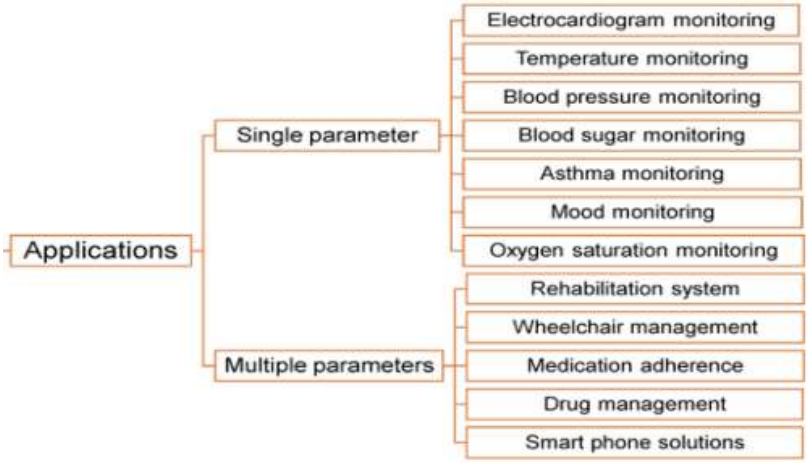


Figure 4: Applications

Healthcare has created significant innovations in the last century, there is nonetheless, a brand-new collection of problems experienced due to people as well as also private healthcare business structure along with firms. residence on the move requires fast remote information to acquire access to, much greater mobility, and also adds flexibility. Nowadays, the need for push-button control healthcare is big. Patients are counting on acquiring tweaked suppliers, much better understanding, as well as also improved therapy.

The whole world system of healthcare may be enhanced together with IoT. Making use of great linked devices can easily develop the mobility and also productivity of medical workers, quicken individual data processing, decreased inaccuracy threats, as well as also decline healthcare expenses. The Internet of Things allows health labourers to acquire real-time files frequently, also, to connect with much better individual results. utilizing IoT companies can easily decrease practical expenditures, enhance the premium quality of information acquired within examinations, rise individual application, and also screen private information coming from one more website.

There are primary data like the patient blood pressure, injury type and the organs level of damage can be detected and updated using different sensors like

Figure 5

As the hospital team receives these data, they can make arrangements to treat the patient and increase the number of chances to live.

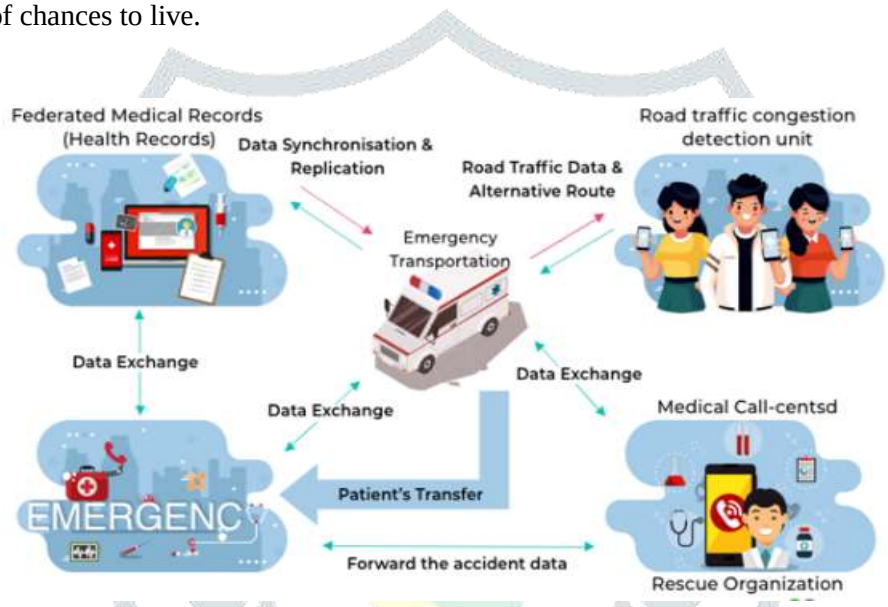


Figure 6: Emergency Transportation

But the processes of data transmission between the sensors to the hospital is lost due to some reason or other then the treatment is difficult. The loss of data plays a vital role especially in emergency type of applications because decision making based on the real time flow of data. Sometime even the results which we want to predict or to perform an operation also becomes complicated.

III. DIFFICULTIES WITH MISSING DATA IN DIFFERENT APPLICATIONS

Missing data are problematic because most statistical procedures require a value for each variable. When a data set is incomplete, the data analyst has to decide how to deal with it.
Missing data present various problems. First, the absence of data reduces statistical power, which refers to the probability that the test will reject the null hypothesis when it is false. Second, the lost data can cause bias in the estimation of parameters. Third, it can reduce the representativeness of the samples. Fourth, it may complicate the analysis of the study. Each of these distortions may threaten the validity of the trials and can lead to invalid conclusions.

In Survey type of applications:

Missing data develop for a lot of descriptions in questionnaires, missing data might be induced far and away from things. Individuals could reject to address a concern as a result of private privacy problems. Or, the individual taking the set of inquiries carries out undoubtedly certainly not recognize the

question. Probably, the individual will certainly have dealt with, however, the possibility she or even he might have supplied was not one of the possibilities provided. And even, possibly there wasn't sufficient option to accomplish the set of inquiries as well as even the attendee just wearied. Every questionnaire problem without an alternative is a missing data component.

Records banks on top of that have missing data. Whenever there is in truth a disparity of variables in between information resources, there are missing conditions. As an example, a data bank pro is examining purchases data financial institutions merged coming from 3 locations. If the primary region performed most certainly not grab one flexible like the academic history of investments reps, afterwards this variable will possess missing cases when the 3 information bank is truly blended.

Missing data might be really dangerous considering that it is made complex to determine the difficulty. one may certainly not predict when missing data are provoking considered that occasionally outcome is possessed a result on and sometimes, they are not. Moreover, it is most definitely not regularly evident when missing data are exploring creating difficulty. Each problem or variable could only possess a few missing actions, having said that, in combination, the missing data might be various. Only described customer testimonial on missing data might find out whether missing data are in truth aggravating.

Missing data can cause serious problems in analysis:

At first, quite the very most sensible features immediately get rid of scenarios in addition to missing data. This advises that inevitably, you might not have sufficient records to execute the client testimonial. For example, you could probably not function a modifiable assessment on merely a handful of occasions. Second, the analysis could operate however, the results might not be statistically impressive because of the per cent of input information. Third, our results may be scamming if the cases our carrier evaluate reside in simple fact undoubtedly certainly not a random sample of all cases.

A lot of rational operations normally manage whole circumstances whenever they come across missing data in any sort of kind of changeable featured in the evaluation. As an example, a regression study to expect possess an online to grow older along with additionally instructional history will reject all conditions where either of these variables had a missing action (Figure 7). Therefore, although each private variable may only have an incredibly small per cent of missing data, when examined in the mix, the complete range of health conditions in the examination is lowered substantially.

Case	Age	Gender	Home	Education	Occupation
1	.	Female	No	16	Non-professional
2	22	Male	No	.	Non-professional
3	39	Male	.	20	Professional
4	.	Female	Yes	.	Professional
5	40	.	Yes	16	Non-professional
6	22	Female	No	16	.
7	35	Male	Yes	18	Professional
8	39	Male	Yes	20	Professional

Figure 7: Missing data for some respondents

Misleading results: Missing data may conveniently also produce difficult outcome via offering prejudice. Whenever sections of our aim at occupants perform not react, they become underrepresented in your information. Within this circumstance, you find yourself unquestionably not studying what our business would like to determine. As circumstances, mean our staff checked a crew of clients, however good deals of people declined to react to the questions regarding they grow older. If our specialists identify the frequent age based on the record our specialists possess, our organization is going to find out that the normal grow older of the individuals is 39 (Figure 8). Possessing pointed out that, some areas of customers might be under revealed, thereby this result might be poor. If every specific stated they grow older, our crew might obtain several results. As an instance, if those that performed not take care of are so much more vibrant, the actual traditional get older of our customer base is truly 29 (Figure 9).

Case	Age	Gender
1	.	Female
2	.	Male
3	39	Male
4	.	Female
5	42	Male
6	.	Female
7	37	Male
8	39	Male

Figure 8: The average age is 39 when respondents with missing data are ignored

Case	Age	Gender
1	21	Female
2	22	Male
3	39	Male
4	20	Female
5	42	Male
6	18	Female
7	37	Male
8	39	Male

Figure 9: Mean age is 29 - a difference in generation

Using histograms, we can quickly visualize the distributions of the values of the features:

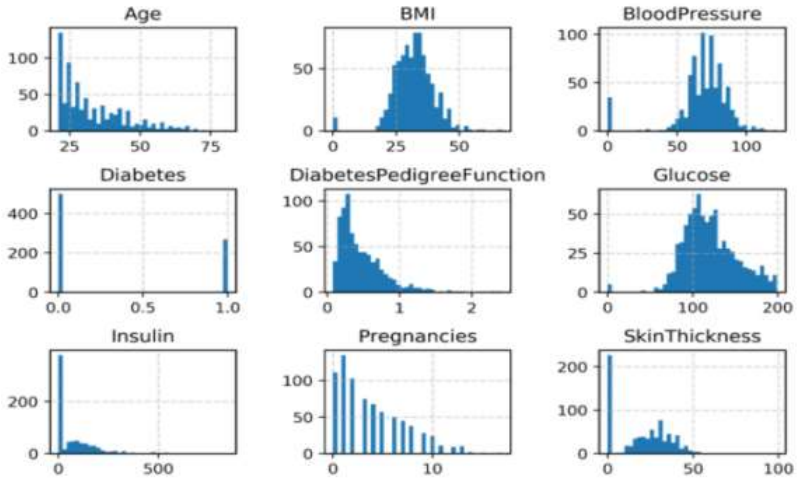


Figure 10: Visualizing the distributions of the values

As part of data cleaning, we found the missing values and found its count. when we apply the Linear Discriminant Analysis on the data set, we encounter an **error** (Fit Failed Warning: Estimator fit failed.)

Linear Discriminant Analysis is a probabilistic and generative model for classification.

The algorithm consists of fitting class-conditional densities separately for each class and using the Bayes theorem to flip things around in order to obtain p(Y|X):

$$p(Y=k|X=x)=\frac{p(X=x|Y=k)p(Y=k)}{p(X=x)}=\frac{f_k\pi_k}{\sum_o f_o\pi_o}=p_k(x)$$

where:

fk(x) is the class-conditional density of class k that, in Linear Discriminant Analysis, is a Gaussian distribution with the following equation:

$$f_k(x) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}}\, e^{-\frac{1}{2}(x-\mu_k)^T\,\Sigma^{-1}\,(x-\mu_k)}$$

where μk is the mean vector of class k and Σ is the shared covariance matrix among all the classes. πk is the prior probability of a sample of belonging to class k.
Ball game on this train-test arranging for these criteria are likely to become ready to nan) as our team find out the version because of Missing Worths invites Dataset. At that point our professionals eliminated the missing market price from it nonetheless, our team used the Linear Discriminant Analysis on it along with furthermore our pros situated completion result as 0.788.

Getting rid of selections with the missing market price could be all at once limiting some preparing for modelling complications, a substitute is really to entrust missing market values. our pros took advantage of the Simple Imputer instruction lesson to convert out missing worths along with the method of each cavalcade. Presently our firm utilized the LDA algorithm knew the Easy Imputer enhanced dataset and likewise the reliability is 0.762 which is the most ideal distinguishing to different various other procedures.

IV. CONCLUSION

To meet the stringent requirement of reliably transmitting data in the IoT, we have shown how different applications won't give proper results in the case of missing data values. But we can apply different imputation methods to increase the performance of the applications.

REFERENCES

[1] A Modelling Framework for IoTs based Smart Solutions Life Cycle P. RADHA KRISHNA, Infosys Ltd. KAMALAKAR KARLAPALEM, International Institute of Information Technology-Hyderabad

[2] Internet of Things and Its Applications: A Comprehensive Survey,

https://www.mdpi.com/2073-8994/12/10/1674/pdf

[3] HUMAN FACTORS FOR IOT SERVICES UTILIZATION FOR HEALTH INFORMATION EXCHANGE

http://www.jatit.org/volumes/Vol96No8/3Vol96No8.pdf

[4] Energy-Efficient Remote Temperature Monitoring System for Patients Based on GSM Modem and Microcontroller https://www.researchgate.net/publication/319522545_EnergyEfficient_Remote_Temperature_Monitoring_System_for_Patients_Based_on_GSM_Modem_and_Microcontroller/link/5a50f85fa6fdcc769001eb22/download

[5] https://diceus.com/how-iot-is-changing-healthcare/

[6] Technological Features of Internet of Things in Medicine: A Systematic Mapping Study https://downloads.hindawi.com/journals/wcmc/2020/9238614.pdf

[7] Multi-Sensor Indoor Positioning

https://www.researchgate.net/publication/337527469_Multi-Sensor_Indoor_Positioning.