IMPACT: International Journal of Research in Applied, Natural and Social Sciences (IMPACT: IJRANSS)
ISSN(P): 2347-4580; ISSN(E): 2321-8851
Vol. 5, Issue 6, Jun 2017, 51-60

A NOVEL DATA STREAM CLUSTERING ALGORITHM IN HEALTHCARE IOT

AMEER N. ONAIZAH

College of Computer Science and Mathematics, University of Kufa, Iraq

ABSTRACT:

The Internet of Things (IoT) is going ahead to creating direction to make seals integration of network information and physical object. As an important technique of data analysis, clustering attempts to find the underlying pattern structures embedded in unlabeled information. Unfortunately, most of current clustering techniques that could only deal with static data become infeasible to cluster a significant volume of data in the dynamic industrial applications. We propose a method which determines how many different clusters can be found in a stream based on the data distribution in healthcare application. After selecting the number of clusters, we use an online clustering mechanism to cluster the incoming data from the streams. In the proposed algorithm, two cluster operations, namely cluster creating(initialization Buckets) and cluster merging, are defined to integrate the current pattern into the previous one for the final clustering result and k-medicos is employed to modify the clustering centers according to the new arriving objects. Finally, experiments are conducted to validate the proposed scheme on two cases in terms of clustering accuracy and computational time.

KEYWORDS: Iot, Stream Clustering, Healthcare, Buckets.

INTRODUCTION

The existence of (IoT) has not been returned for long time. But it has been found ear list since 1800sconcurent with communication machines. Machines helped to widen the range of communication to be directly since (first landline) telegraph in the 1830s,then in 1840s developed to (wirless telegraph), while the first voice transmission through radio coming in June 3, 1900, computers developed start in the 1950s. Making necessary development extend in the Internet of Things. [1].

The most important component of the IoT is the internet itself. In 1962, was the started when the internet take apart of DARPA (Defense Advanced Research Project Agency). And in 1969 developed in to the ARPANET which the public comes to use it in supplying the commercial service in the 1980s, till it developed into our modern Internet. In the ear list 1993, Global Position Satellites (GPS) become reality. System with high function include 24 satellites a constant and provided by Department of Defense, Then the private owned quickly possess it Later the orbit come when replace the commercial satellites with it Basic communication provided from much of the IoT by landline and satellites[2].

The IPV6's was important motif and additional to involving the functional of IoT Expanding the space of address come as intelligent remarkably decision. For the (Computer History Museum) Steve Lisbon states "the address space expansion means that we could assign anIPV6 address to every atom on the surface of the earth and still have enough addresses left to another 100+ earths." Put by other ways, we are not led to run out of internet addresses anytime soon. [3].

The “Internet of Things” (IoT) can notion is not more that science stuff except the main part which it reality of use it in every day our life. Round earth electronic and digital interconnected devices estimated more than 13 billion operations together. The equivalent each human has 2 devices nearly on the earth. The (smart home) so called devices is a live example for the IoT such remote controlled appliance programmable thermostats. Growth the using of IoT and largest it to indicate applications of technology in every sector of the virtually and economy from commercial and dustier environment to public safety and health

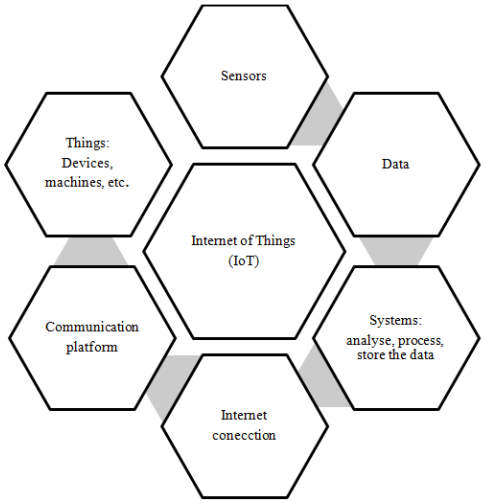

Figure 1: Internet of Things Components

In this Ul white paper, general introduction to supply the readers about standards and technologies which it supporting the IoT to continuous widespread deployment. The beginning show current overviews and future of applications for IoT. Then the paper discuss make the IoT possible by specific technology and details supplying in progress on current standards development activities. [7].

Related Work

Clustering broblem has many approaches, but we will give close study only for particular approach:

Daniel Puschmann, PayamBarnaghi, and Rahim Tafazolli[8], these researchers enables to adaptively process large volumes of dynamic data online based on the current situation. This paper shows how proposed method adapts itself to the changes. Also demonstrate how the number of clusters in a real-world data stream can be determined by analyzing the data distributions.

Qingchen Zhang; Chunsheng Zhu; Laurence Tianruo Yang; Zhikui Chen; Liang Zhao; Peng Li [9], in this proposed, the proposed algorithm, two cluster operations, namely cluster creating and cluster merging, are defined to integrate the current pattern into the previous one for the final clustering result and k-mediods is employed to modify the clustering centers according to the new arriving objects. Finally, experiments are conducted to validate the proposed scheme on three popular UCI datasets and two real datasets collected from industrial Internet of Things in terms of clustering accuracy and computational time.

Mingze Wu, Yitong Wang, and Zhichao Liao[10],New algorithm proposed in this paper to cluster multi - type sensor data current to provides future predicate for installment data of a nominate farm product. This proposed method based on stand of agricultural IoT.

Big Data Era: a Challenges Issues in Iot

With the rapid development of IoT, big data, and cloud computing, the most fundamental challenge is to explore the large volumes of data and extract useful information or knowledge for future actions. Figure 2 explain the IoT way where in this paper deals with healthcare data. In IoT field, the key features can be account as a big data, they are as the following. [9].

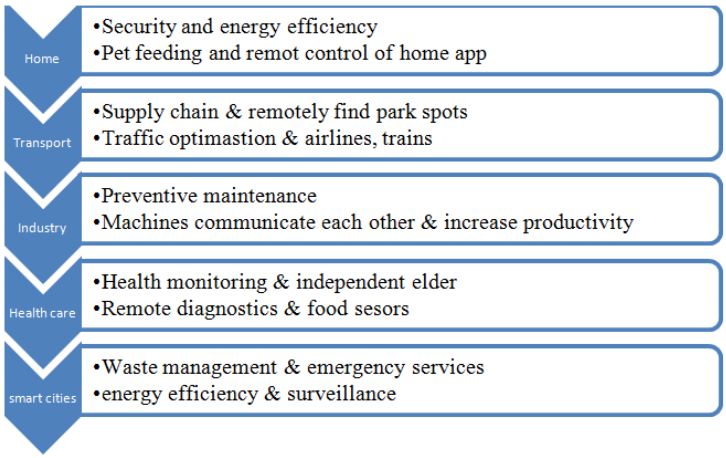

Figure 2: IoT Ways

The key characteristics of the data in IoT era can be considered as big data; they are as follows [9].

- Read and write large amount of data: PB (pet bytes) even TB ((terabytes) and ZB (zeta bytes) refers to the data's amount, so we need to discover mechanisms that be more effective and fast.
- Data kinds to merge and Heterogeneous data provenances: variety of data provenances provides a big area for data. In follow example combating data what we require in sensor data, social media, camera data, and so on. These data have various shape, number, string, binary, byte and so forth. Also the communication with various types of system and various types of devices what we need with necessary to extra data from WebPages.
- Complex knowledge to extract: a large amount of data contains knowledge which it hidden deeply in it, the knowledge unobvious, and to find the associated of different data need to analyze the properties of data.
- When big data and IoT come allot of challenges appear; the quality of data is slow when the data quantity is big and data sources are different according to data various which it possess a lot of various type inherently and representation form, and the data is as -structured and heterogeneous. To analyze challenges in data extracting, data mining system area, data mining algorithm. The following is a précis for challenges. [10].
- The initial challenges are to incoming, educing a big amount of data from supplies location. It contains different data. The big challenges when find the error and also harder in data correction. We want to agree with heterogeneity, noise and set of data.

- A big challenge is modifying traditional algorithm to a big data environment in data mining algorithm area.
- Second challenge in big application and how to mine incomplete and uncertain data, share data among various systems and applications is an effective and security solution in data mining system and one of the most important clangs. Such as medical records, banking transactions, should me matter of interest.
- This paper deals with first challenges and try to put a suitable solution to addends this challenge and studying the methods and analysis of stream data. Figure 3 explain the suggested big data mining system [11].

Service	Classification	Clustering	Association	Time series Analysis	Other analysis
Data Processing	Distribute file system (HIDFS)	Programming (Map reduce)	Real-time Analysis (Storm)	Batch analysis (Hadoop)	Workflow (Oozie)
Data gather	Real-time Data receiver	Data parser	Data queue	Batch data Extractor	Data merging
Raw data	Structure Data		Semi structured data		Unstructured data
Devices	Sensor	Camera	RFID	Printer	Other IoT Devices

Figure 3: The Suggested Big Data Mining System

PROPOSED METHOD

In this part, we will deal with how one might cluster a stream briefly. The idea that came in mined where is there a sliding window of N points, and we able to question about clustered or centric of the better cluster created from the last M a

M in this dots, the sliding window contains the most newest N point, for any $m \leq N$ prelisted subsection of points in the stream is our goal.

When the point of the stream lives, we make no limited concerning the space, in case the Euclidean space the replay to the question is the cancroids of the elected clusters. In case the space may be non - Euclidean the replay is the clustered of the elected clusters. Where any of the "clustered" definition may be used (the problem is much easier) with assume that chosen all stream element with calculated don't very long the streams. Then streams sample is good enough to trail the clusters. And we can in effect ignore the stream after awhile, however, the assumes of stream model normally be that stream elements raid is varies with time, for example, the clusters centric immigrate slowly as time run out, or clusters may extend split, contract or merge. There for, split of a BDMO algorithm which it refers to (for the authors, B. Babcock, M. Datar, R. Motwani, and L. O'Callaghan). The original version of the algorithm is more and much complication in structures, to supply performance guarantees in the worst case.

The BDMO Algorithm stand on the methodology for counting in stream for once, where the streams points are divided into, and briefed by buckets whose the buckets size refers to the points number, rather than I witch it the number of stream elements, However, the size of buckets follow the constraint which there are two or one for each size, achieve to specific limit, anyways, we don't take in account that assumption the size of admissible bucket sequence start with 1. They are required only, where each size is twice the pervious size, e.g., 3, 6, 12, 24, [12].

The bucket content is consist of the bucket size and bucket timestamp which it most newest point that contributes to the buckets, therefore, and can be notation modulo N by timestamp, a set of notations that the clusters performing into the points of that buckets have been divided. These records contain:

(a) The number of cluster points

(b) The clustered or centric of the cluster.

(c) Any ells parameters requisite to enable us to combine clusters and preserve approximation to the full set of parameters for their combine cluster.

Step One: Initializing Buckets

P will be the size of the smallest bucket, a power of 2, for that, new bucket producing in every P stream element, in the most newest P points. The timestamp of the newest point in the bucket is the timestamp for this bucket. And every point in the cluster may leave by itself and depending on the strategy that chosen, we already perform a clustering of these points. For example if K -means algorithm chosen, then (assuming k < p) we cluster the points for K clusters via some algorithm

The methods that use at first to cluster, whatever we assume that the possibility to calculate the points in each cluster and count the clustered or centric for the cluster. This mentioned information comes to be a part of the records for each cluster, whatever; we also calculated other cluster parameters that it need in combining process.

Step Two: Merging Buckets

What it happens, new bucket creating, then it require to rehearsal the sequence of buckets. First, if some of these buckets has a timestamp that is further than N time units prior to the current time then no one of that buckets is in the window, and we may exclusion it by the schedule. We may force to create three buckets of size P, in case the oldest two of the three must be combine. In case we forced to combine buckets of increasing size when the M combining may create two buckets of size 2p, to merge two consecutive buckets, we need to do several things [10]:

- The size of two combine buckets equal the size of the bucket in twice
- The timestamp for the joint bucket is the timestamp of the more recent of the two sequential buckets.
- We should regard if to combine clusters, instead, we require you calculate the parameters of the combine cluster. We will work on this section of the algorithm, via regard many cases of the certain of merge ring and ways to assessment the needed parameters.

Cases Study

In this searcher, two cases will be discussed to illustrate the proposed algorithm.

Case1: where we use a k - means approach in Euclidean space maybe represent the simplest case. And we perform clusters by calculate of their android and points. Each cluster has precisely k cluster. For this we realize p=k or P larger than K in other pick and when we initially create a bucket the P points cluster in to K clusters and we must get the best between K cluster of the first and second bucket. And "best" here refers to the matching that decrease the to total if the spaces between the matched cluster centric.

In our assumption the clusters don't develop allot between sequent buckets, for that we don't consider combine two cluster from the same bucket. Thus, we would foresee to find between each of two neighbors bucketed a representation of any one of the K "true" cluster that exist and found in the stream

Through determined to combine two clusters one cluster from each bucket, the number of the point in the combine clusters is same with the number of points. In two clusters the sum is equal surly the weighted are range in the centric of the combine clusters is same with the centric of the two clusters. Where the measuring is by the number of points in the clusters that is, if points n1 and n2 in two clusters, respectively with and centric c1 and c2 (the latter are d-dimensional vectors for some d), then the combined cluster has n = n1 + n2 point and has centric.

$$c = \frac{n1c1+n2c2}{n1+n2}. \quad \textbf{(1)}$$

Case 2: The procedure in in case 1 be served when the cluster slowly changing. Suppose we might foresee the centric of the cluster combine sufficiently quickly when the centric of the two sequential buckets be matching then we will be face a vague state, there is not evident which of two clusters best match give- cluster from the another buckets. In this situation there is one way to protect which it to make more than one K clustered for each bucket, even if we know that, when we ask, we should to combine with property K cluster. For example we perhaps select P to be much combine than K, and through combination, only combine cluster when the result enough consistent the hierarchical strategy can use to make better combine, so as to maintain>K cluster in each bucket.

To be exact suppose to put edge for the total spaces that between all points of the cluster and it is centric. We can implicate rating of this aggregate in the record for a cluster ' and it is an additional to the calculate points and the android of a cluster. when we prepare the bucket. We can exactly calculate the aggregate. These parameters stay a rats only, as well as we merger cluster. We suppose two cluster combine and want to calculate the sum of the space for the cluster that combine. The symbols use for the centric and calculate in case 1 and for additional 1st s1 s2 be the total for the two clusters. Then we have to establish the radians of the combined cluster to be this is the space sating of the new centric C and any point X.

$$n_1 \mid c_2 - c \mid^2 + n_2 \mid c_2 - c \mid + s_1 + s_2 \quad \textbf{(2)}$$

Thisisratingofspaceforthenewentroidpointanypoint Xtobethespaceofthatpointtoitisoldcentroid(thisspacesums1+s2,thelasttwoyermsintheaboveexpression)addedtoitthespacefromtheoldcentroidtothenew(thesespacesisthetotalofthefirsttwotermsoftheupperexpression).Attentionthatthespacethatisaboveboundedbythestraingleinequality.

$$n_1 \mid c_2 - c \mid^2 + n_2 \mid c_2 - c \mid^2 + t_1 + t_2 \quad \textbf{(3)}$$

The alternative is the changing to the sum of square of the distance instead the sum for the distance and from the two points into the centric. If these sum for the two cluster are t1 and t2, respectively, then we able to product rating to the similar sum in the new cluster as :

In order to explain the proposed algorithm, the demons trod examples below describe briefly the steps of this algorithm.

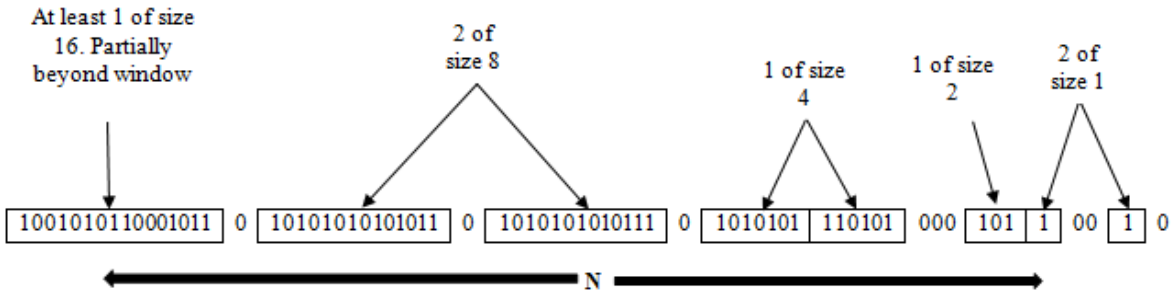

Figure 4: The Selection of Sliding Window in Proposed Algorithm

Three Characteristic of the Maintained Buckets

- Each **one** or **two** buckets with the similer**power-of-2**number of **1s**as shown in figure 4
- Buckets do not interfere in timestamps
- Buckets are saved via size

When the comma g of new bit in drop the final (oldest) bucket if its end-time is previously to ***N*** time units before the current time

Two cases Current bit is**0**or **1**

- Don't need for the any changes If the current bit is **0**
- In case the current bit is **1**
- Generate a new bucket of size 1, for just this bit End timestamp = current time
- If supposed now three buckets of size 1, merge the oldest two for the a bucket of size two
- If supposed now three buckets of size 2, merge the oldest two of the a bucket of size four
- And for thus, as shown in figure 5.

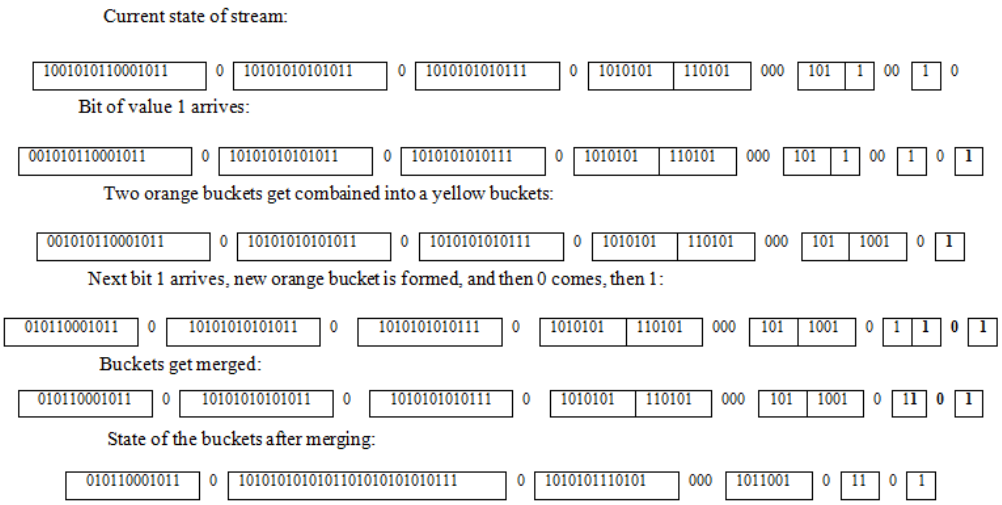

Figure 5: The Strategies of Proposed Algorithm

RESULT AND DISCUSSIONS

In this paper different health related data with heart rate used for the clustering of heart attack and reveal these myths. An example dataset from UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/Statlog+(Heart)] from Centre for Machine Learning and Intelligent Systems has been used here. 270 patients dataset has been restructured such a way that keeping in mind the results which are generally obtained from healthcare cloud platform. The data set consists of 270 observations from 250,450 sensor data and 8 variables (out of which 4 numeric variables, 3 binary values and one is id) which are as follows: Patient id, Sex, Age, Resting Blood Pressure, Serum Cholesterol, Fasting blood sugar if more than 120 mg/dl, maximum heart rate within 24 hours, Heart diseases In real life, more important parameters may be needed to measure the patients who are in danger period for heart attack. This paper can classify this data field on the base of collection procedure. Here it is in three ways. Two live data blood pressure and heart rate will be collected by using wearable sensor devices. These data will be stored in a separate temporary storage and from there highest heart rate and average blood pressure in a certain period will be recalculated from this temporary storage and finally stored in main database. Main database will consist with patients' static information like age, sex and other periodic information such as blood sugar and serum cholesterol. In this dataset binary values have been used for fasting blood sugar measurement. If the blood sugar level more than 120 mg/dl, this result is treated as one and if this is 120 mg/dl or less than that, this is defined as normal or zero.

In this part, the experiment that done stand on the actual data of agric true Internet of Things. Since December 2011 to July 2013 collected the sensor data stream from producing of Wuxi and Binzhon in China. It contain of 270 observation of 23 kinds of farm product and 205, 437 sensor data. The proposition of clustering algorithm apply on the data collection

Figure 6offer the outcome of the proposed algorithm, the window size set to 24, window sliding space set to 20, and combine threshold set to 0.25, the real data for any causes. Some lost data observation clustered for the same data. And for that to view the outcome more clearly. In figure 6 this cluster is not show.

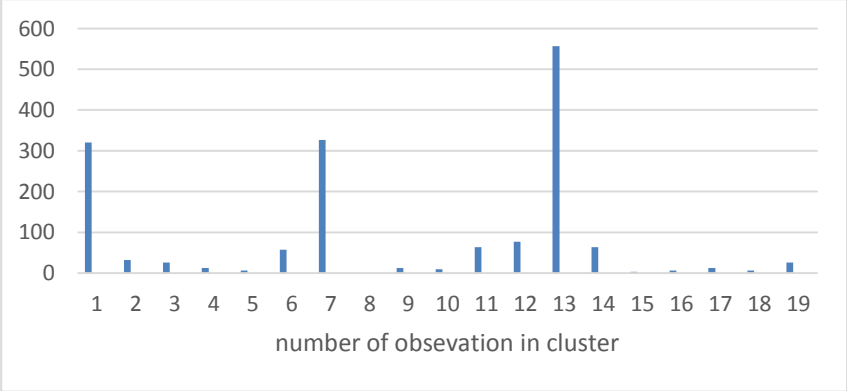

Figure 6: Result of Proposed Algorithm (W=64, S=20, I=0.25)

Through the clustering result. In cluster 13 can see 567 batches which all these 567 batches the most of consist the similar kind of farm products and totally ingeneral data there 23 kind of batches. That view the performance of our clustering alogarithm from the side. J is the similer merger threshold and we wanted if we desir to specific the similer kind of produce which is stand on data through the cycle of production, in this application pasically.

The cluster in this size met the request of determining the farm products quality

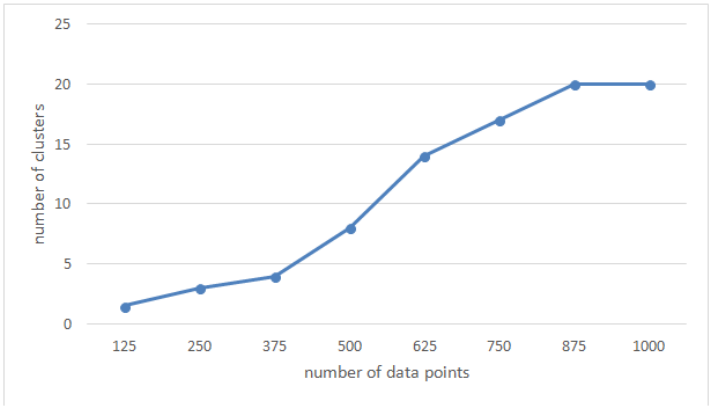

Figure 7: Quantity of Clusters Changes with the Amount of Data (W=64, S=20, I=0.25)

Figure 7views the shift of quantity of the cluster through the amount of data points. Mostly lost data observation distributed some at the beginning of the data stream, forthus,at the begin number of the cluster stays two only, at then, show increasing in data points number, and the number before cluster increasing too.

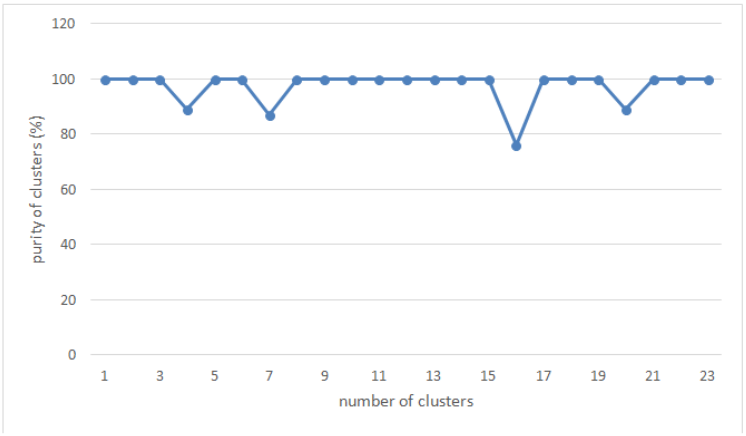

Figure 8: Purity of the Clusters (W=64, S=20, I=0.20)

Figure 8our algorithm views the cluster generation purity. It is manually counted. We take in considering that, this purity for our application is high enough.

CONCLUSIONS

In this badge, we assume a new algorithm of data stream clustering stand on sliding window (initialization Bucket). And micro - cluster merging. In this badge an adaptive clustering method. Presented that is designed for dynamic IoT data stream. In health care application. The method copes to data drift on the un dealing data stream. And the able of the assume method to specify the number of categories that exist inherently in the stream of data and stand on distribution of data and a quality of the cluster measure. Show the performance of the proposed algorithm the adaptive methods without prior knowledge and is enable to discover inherent categories from the data stream. Finally the proposed algorithm show better performance in view of speed and exposure data in better manner than traditional

algorithms. For the future work we plan to apply the proposed solution to different types of multi-modal data in the IoT domain.

REFERENCES

1. Keith D. Foote, "big data internet of things IoT", August 16, 2016
2. Thorsten Kramp , Rob van Kranenburg, Sebastian Lange, "Introduction of IoT", 05 September 2013
3. Alessandro Bassi, "Enabling Things to Talk: Designing IoT solutions with the IoT Architectural Reference Model", 2013.
4. Z. Sheng, C. Mahapatra, C. Zhu, and V. C. M. Leung, "Recent Advancesin Industrial Wireless Sensor Networks Toward Efficient Management in IoT," *IEEE Access*, vol. 3, pp. 622-637, 2015.
5. C. Perera, A. Zaslavsky, P. Christen, and D. Georgeakopoulos, "Sensing As a Service Model for Smart Cities Supported by Internet of Things," *European Transactions on Emerging Telecommunications Technologies*, vol. 25, no. 1, pp.81-93, 2014.
6. C. P. Kruger and G. P. Hancke, "Implementing the Internet of Things vision in Industrial Wireless Sensor Networks," in *Proc. of the 12th IEEE International Conference on Industrial Informatics*, 2014, pp. 627-632.
7. C. Zhu, L. Shu, T. Hara, L. Wang, S. Nishio, and L. T. Yang, "A Survey on Communication and Data Management Issues in Mobile Sensor Networks," *Wireless Communications and Mobile Computing*, vol. 14, no. 1, pp. 19-36, Jan. 2014.
8. Daniel Puschmann, PayamBarnaghi, and Rahim Tafazolli, "Adaptive Clustering for Dynamic IoT Data Streams", 2016
9. Qingchen Zhang; Chunsheng Zhu; Laurence Tianruo Yang; Zhikui Chen; Liang Zhao; Peng Li, "An Incremental CFS Algorithm for Clustering Large Data in Industrial Internet of Things", 2017.
10. Mingze Wu, Yitong Wang, and Zhichao Liao, "A new Clustering Algorithm for sensor Data Streams in an Agricultural IoT", 2013.
11. B. Babcock, M. Datar, R. Motwani, and L. O'Callaghan, "Maintaining variance and k-medians over data stream windows," Proc. ACM Symp. on Principles of Database Systems, pp. 234–243, 2003.
12. P.S. Bradley, U.M. Fayyad, and C. Reina, "Scaling clustering algorithms to large databases," Proc. Knowledge Discovery and Data Mining, pp. 9– 15, 1998.
13. V. Ganti, R. Ramakrishnan, J. Gehrke, A.L. Powell, and J.C. French:, "Clustering large datasets in arbitrary metric spaces," Proc. Intl. Conf. on Data Engineering, pp. 502–511, 1999.
14. H. Garcia-Molina, J.D. Ullman, and J. Widom, Database Systems: The Complete Book Second Edition, Prentice-Hall, Upper Saddle River, NJ, 2009.