IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-10 E-ISSN: 2347-2693

A Relative Study of Data Mining and Big Data on IoT 's Data Streams

Sallauddin Mohmmad[1*], Syed Nawaz Pasha[2], Dadi Ramesh [3] ,Shabana[4]

[1*]Computer Science & Engineering, SR Engineering College, Warangal, India
[2] Computer Science & Engineering, SR Engineering College, Warangal, India
[3] Computer Science & Engineering, SR Engineering College, Warangal, India
[4] Computer Science & Engineering, Sumathi Reddy Institute of Technology for Women,Warangal, India

[*]*Corresponding Author*: *sallauddin.md@gmail.com, +91 9885502477*

Available online at: www.ijcseonline.org

Received: 27/Sep/2017, Revised: 08/Oct/2017, Accepted: 19/Oct/2017, Published: 30/Oct/2017

Abstract— the huge amount of data produced and captured by the Internet of Things (IoT) in current situations. IoT data created by several applications such as from smart cities, monitoring system, health care, object detecting systems and smart systems etc. which all produce the continues data streams. But data of IoT recognized with RFID signal their size is near to 20 bytes. So that it is requiring of special procedures for analyzing, managing and mining the IoT stream data which is identified by RFID. Several software tools have been developed for mining the stream data relatively Big Data tools. While most current research looking into data mining and relative big data tools which are on machine learning of distributed and non-distributed systems with respect to IoT stream data. This paper focused on IoT data format, different kind stream data mining tools according to mining steps and few Big Data tools relative with data mining for handling the stream data of IoT.

Keywords—IoT, Data mining ,Big Data , RFID, Stream Data.

I. INTRODUCTION

The Internet of Things (IoT) paradigm is a collection of intelligent and self-configuring nodes (things) interconnected in a dynamic and global network infrastructure which enables these nodes to exchange data. It is one of the emerging technologies for ubiquitous and pervasive computing. IOT has infrastructure that interoperates with the existing internet [1]. The things in the IOT can be devices such as sensors, actuators or any other software or electronics. Internet of things (IOT) can be taken as an important extension of Internet. The IOT ecosystem consists of wireless connected objects that share information with other objects. In traditional era, information was shared among people. The IOT takes this concept of sharing information and integrated it into sharing information between information world and physical world. IOT connects digital world to the physical world creating whole new era of computing and adding a new dimension to the existing infrastructure. The IOT is generally a dynamic network formed by interconnection of various wireless objects or things. The quantity of goods in a medium-sized super market will be around millions or tens of millions. In supermarket the information of the goods is tracked using RFID. Assume that there are 10 million items need to be tracked and each and read 10 times a day, hence each time generate 100 bytes. As per this the amount of data per day will reach up to 1 GB and for one year it will be around 3650 B.Ths. enormous amount of data is being generated by medium sized super market as such there will be millions of medium and large scale data generating sources [1, 2].Further in the field of real-time monitoring such as ecological monitoring, satellite sensing wireless sensor networks needs to record multimedia information generated from multiple nodes. The amount of data generated through these real time systems will be around 1 TB per day [3]. In some emergency monitoring system the data is generated in real time and continuously in the form of streams [4].

One of the important aspects of the IOT is to maintain the timeliness of the data. The data perceived by the system reflect the existing state of the thing [8]. Therefore regardless of the WSN or RFID system, the data acquisition of IOT is carried out at every second to send the current data to the server and update it. Hence the responsiveness and practicality of the system is the key to reliability [8, 9]. This requires that the software data processing system of IOT must have sufficient speed of operation; otherwise, it may lead to erroneous conclusions and even cause great losses. We can realize that IoT data is tagged with RFID such Sensor data inject into an appropriated application for further data processing which may be a mining, modeling, analytics and visualization then only data users to reach their objectives. IoT data also processed in several layers of abstraction [10].]. The efficiency of the mining depends on the KDD step that is not all the attributes are essential for mining. Hence attribute subset selection or feature selection

methods are used to select essential attributes for mining. The task of selecting essential attributes of big data and processing them is challenging to the traditional data mining algorithms [16, 20]. Hence the traditional data mining algorithms has to focus on distributed processing of the huge data from IOT to improve the system performance and service quality of IOT [25]. It has to utilize the parallel processing of map reduce to improve the efficiency and throughput of the system. This methodology follows the descriptive statistics which generate the output inject into visualization devices, dashboards. We can satisfy this when the data size medium range. Huge data analysis will require more advanced tools and mechanisms which they might be require the predictive analysis including machine learning computing .From this paper we are going to describe about the study of data mining tools which they also relative with Big Data analytics on Iota's data Stream . In the section II we are going to discuss on IoT data representation at user interface and how RFID converted into XML and EXI with a layer model .Section III conveys the complete data mining paradigm for IoT stream data with a multi-layer data mining model .We prolonged the discussion with classification, clustering, association and time series oriented traditional and trendy tools for stream processing. Section IV focused on the Big Data tools and challenges related to the data mining .In this section we try to make you understand about small data, huge data and IoT data origin points etc.Section V will talk about theme of the paper which explained with a diagram .From that section we can analyze the what is predictive data mining and descriptive data mining .we completely went into the machine learning distribute computing tools and non-distributing computing tools .In The section conclusion we talk about what are things have to reader can understand from this paper and also initiated our future work which will become enhance to this paper.

II. IOT DATA AT USER INTERFACE

Initially, RFID data at application interface represented as the XML related format.XML is used for representing the IoT data it is also using in data encoding methods It brings the more benefits for IoT data [41, 42].XML data format is portable in nature huge support for all kind of tools, APIs and libraries also very well reliable and readable for devices. Most probably IoT devices are self-managing and organizing devices which are having few amounts of RAM and ROM memories .they will use low bandwidth and operating cost. Here XML increase the RAM use and bandwidth utilization. To reduce that conflict alternate XML representation is Efficient XML Interchange also called as EXI[41,43]. EXI provide same as XML compatibility. Many other XML implementations are introduced for IoT data like Wireless Binary XML also called as WBXML and Fast Infoset also called as FIS[44]. Such all kind of data formats is used for data encoding from source to the delivery point. IoT device RFID data formatted in XML delivers in a communication network[41,44]. it will reach to the Intention Manager. Internet manager sends this information to an appropriate destination in the same format using data channel. IoT can also use the Cloud service then we can explain the process in three steps 1)collect the data from different devices and summarize 2)implement the IoT cloud service 3) Set Data Channel to Transmission of data as Intention object.Data encoding (O&M and JSON).IoT data will encoded into object notation format and directly be used GISand can also implemented in SOS as O&M. Already available XML encodings for the O&M and JSON data model[41].

A. Sensor metadata (SensorML)

SensorML is a special model perform the encoding on data then produce and provisions the metadata about sensors. then the measurement methods in a XML format[41].

B. JSON Web Signatures (JWS)

JSON is a very good scripting language to represent data in the IoT systems. It is simple, structured, and semantical named language. for IoT web security purpose JWS is using .it is providing privacy for devices and security fi IoT systems.JSON Web Signature (JWS) is used for security side

ofIoTsystems[41,42,43].

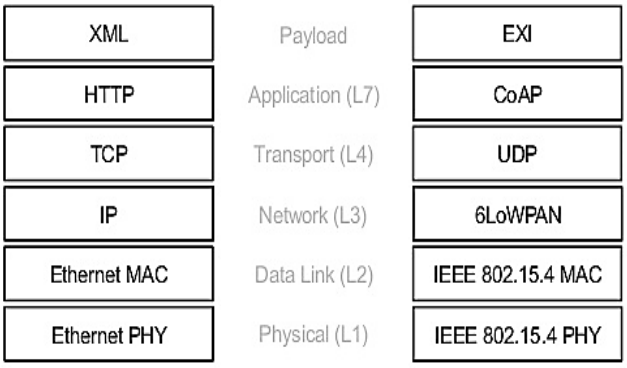

Figure 1. Web base and IoT based layer architecture

III. DATA MINING MODELS FOR THE IOT

A. Multi-layer data mining model for IoT

Multi-layer data mining model for IoT which consists of the following four layers: data collection layer, data management layer, event processing layer and data mining service layer[7].

*1)data collection layer :*The data collection layer is responsible for collecting the data. It collects the data from devices such as RFID readers, sinks, GPS etc[10].A number of problems may arise during data collection such as energy-

efficiency, misreading, repeated reading, fault tolerance, data filtering and communications[11]. Hence these isues must be resolved in the data collection layer.

*2)Data management layer:*It is responsible for managing the data collected during the data collection.It applies centralized or distribute database or data warehouse to manage collected data. The collected data is preprocessed using various preprocessing techniques such as data cleaning,integaration,transformation and reduction[12]. Data format of RFID data stream which has the format as EPC,location,time where EPC marks objects ID.After data cleaning, the data is stored in stay table, info table and map table[5,7].

*3) Event processing layer :*Event processing layer is responsible for analyzing the events in IOT efficiently. It performs event-based query. The observed primitive based events are filtered and complex events are obtained which are desired by the user. It is done by aggregating, organizing and analyzing the data according to events[2].

*4)Data Mining service layer:*The data mining service layer is built based on data management and event processing. It performs various object-based or event-based data mining services such as classification, forecasting, clustering, outlier detection, association analysis or patterns mining, are provided for applications e.g., supply chain management, inventory management and optimization etc.It adopts the service-oriented architecture[4,8].

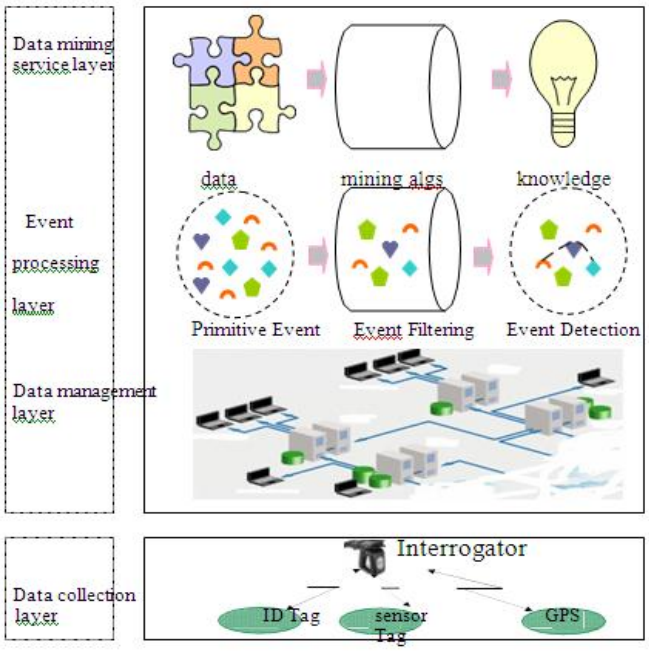

Figure 2. Multi-layer data mining model for IoT .

The IOT has created a data overwhelming that maintains large amount of different valuable information. The task of how to handle this huge data and extract useful information out of this data has emerged in the recent past[12]. The data that is generated from IOT can be classified as "data of the things" that refers the data that describes location, identity, state so on and the "data produced through things" that refers to data generated by things[2,11]. The former one contains data that optimizes the performance related to the infrastructure, systems and things of IOT whereas the latter includes the data that is produced as a result of interaction between the system and the human or between humans or between systems. Due to the above characteristics the data from the IOT has been defined as type of "big data" because it is high in volume and also in velocity. These characteristics help us realize that big data is not that is something related to business but something really bigger than that. The amount of data that is being generated worldwide is thought to be around zeta byte or yotta byte[13].The data analysis tools that are available today are not powerful enough to analyze and handle such huge amount on data generated form IOT[14].The issue of storing huge zeta byte of data in a single storage system and processing such information is a huge task. Hence the data storage and processing has shifted for traditional storage and processing to big data storage and processing. one way of handling the data big data from sensors is to restrict the sensor from collecting unnecessary data and collect only necessary data to be stored and process[3,12]. Hence the focus is to reduce the complexity of the input data. One way to reduce the complexity is to make use of the principal component analysis which is a dimension reduction stratergy.The PCA reduces the number of dimensions using the feature reduction technique. Other technique to reduce the input data is pattern reduction which reduces the number of patterns instead of dimensions[17]. There are other data reduction techniques such as attribute subset selection, numerosity reduction and so on. It is predicted that the organizations will be generating and storing huge data due to the services, applications and platforms they adopt and it is a big challenge for the organizations to extract useful information from such huge data to be able to make better decisions. Many user friendly applications can be developed by analyzing the data from sensors such as smart city and smart homes[15,16]. The interesting patterns can be found in the big data of IOT by applying the KDD(Knowledge Discovery from Data) which is used to find the hidden patterns by applying sequence of steps. The steps involved are data cleaning, data selection, data transformation and data reduction. These steps when applied to big data of IOT can be helpful in generating useful patterns for the IOT data which in turn helps in decision making step. This is useful to get knowledge out of the data[20]. The efficiency of the mining depends on the KDD steps that is not all the attributes are essential for mining. Hence attribute subset selection or feature selection methods are used to select essential attributes for mining. The task of selecting essential attributes of big data and processing them is challenging to

the traditional data mining algorithms[16,20]. Hence the traditional data mining algorithms has to focus on distributed processing of the huge data from IOT to improve the system performance and service quality of IOT[25]. It has to utilize the parallel processing of map reduce to improve the efficiency and throughput of the system.

B. *Classification:*

Classification is one of the data mining functionality which classifies or builds the model for categorical attribute based on the training data and then uses this model to classify new data. The classification technique is called as supervised learning[20].

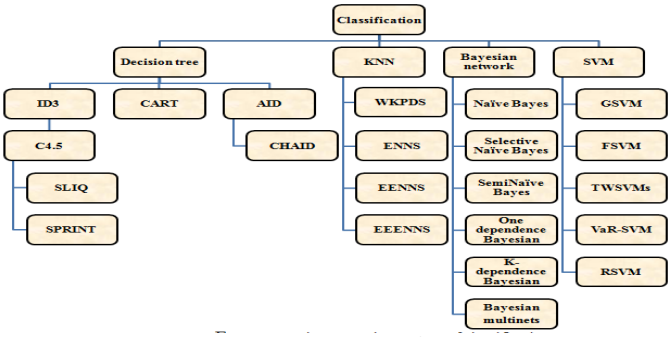

Figure 3. Data mining classification

1) Decision Tree:

- ID3: Iterative Dichotomiser 3 is a simple process of tree learning calculation also called as ID3.it is utilized the twofold parts [17,28].
- C4.5: Next level towards the child of the tree is C4.5 calculation.it is an enhanced adaptation of ID3; C4.5 is processed utilizing the multi-way parts it is proportion to part criteria, it is also simple to enhance for ID3 [18,32,33].
- SLIQ (Supervised Learning In Quest) :In classification SLIQ is can able to classify a large amount of data in the less amount of time[30,31].
- SPRINT (Scalable Parallelizable Induction of Decision Tree calculation): When it constructs the tree for classification it will not see about capacity limitation about data and data collection[35,36].
- CHAID (chi-squared programmed collaboration finder):Mostly this process concentrate on separating the data collection in different formats which they contrast to reaction variables[31].

2) The KNN (K-Nearest Neighbor): calculation based on the Nearest Neighbor result values which are used to identify the nearest neighbor of the visited object. The main thought of the KNN algorithm is to discover the K-closest focuses [28]. Wavelet-Based K-Nearest Neighbor Partial Remove Search (WKPDS) is an upgraded version of KNN [29],also few upgraded versions are ENNS,EENNA and EEENNS all are used find nearest neighbor in the classification.

*3) Bayesian systems:*Bayesian systems tell irregular factors in classification bayesian sense they pretend non-cyclic diagrams. The examination incorporates Bayes [32, 33], particular Bayes [34], Bayes , one-reliance Bayesian classifiers [36], K-reliance Bayesian classifiers [35], Bayesian arrange increased Bayes , unlimited Bayesian classifiers , and Bayesian multi nets .

4)Support Vector Machines: SVM verification process gives parallel classifier, isolating hyperplanes which take input as nonlinear mapping of dataset .it produces the vector results into the multidimensional cubes [32]. Basically, SVM used for content arrangement, showcasing datasets in classification, restorative analysis of information[33]. SVM upgraded with GSVM, FSVM , TWSVMs , VaR-SVM and RSVM .

C. *Clustering*

Clustering basically grouping the similar objects.Hierarchical clustering divide the similar group into subgroups relatively.that sub groups will again from the root grop; those sub grouping process take Various level of grouping procedures.

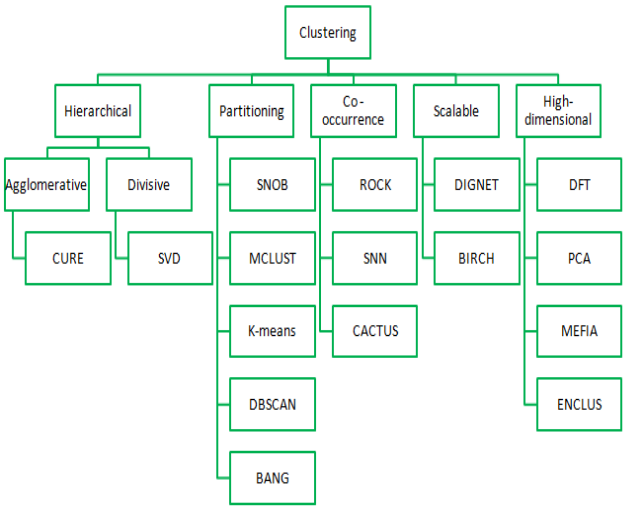

Figure 4. Data Mining clustering

1)Hierarchical clustering : Hierarchical clustering divide the similar group into subgroups relatively .that sub groups will again from the root grop; those sub grouping process take Various level of grouping procedures .basically two models are preferable they are arrangements, agglomerative are following the base up model . and divisive is following the top-down model[27,28]. The agglomerative grouping start with single-point value of data then progressively and recursively combines at least two of the related groups their implementation version is CURE (Clustering Using Delegates) . The divisive grouping begins with a solitary bunch which consisting of all data values and recursively split that bunch into sub clusters . divisive implementation version is SVD (Singular Value Disintegration)[30] .

2)Partitioning calculations: Partitioning calculations find the elements either by iteratively comparing between subsets or by differentiate ranges of most populated groups with information. The related research implementations are SNOB ,DBSCAN, MCLUST [31],BANG and k-medoids.

3) Co-Occurrence : Co-Occurrence create the clusters based on occurrences relatively.thair implementations are ROCK,SNN and CACTUS[28,32].

*4) Scalable bunching :*It can able create the cluster even the data in large size .it can able scale according with data elements.thair research adaptability issues keepsthe registering time can able to recognize the memory prerequisites.Scalable bunching implementations are DIGNET and BIRCH[36] .

*5) High dimensionality clustering :*This methodology implemented for dataset with different qualities and high dimension which occupy the large memory and different values .their implementations are DFT , PCA ,ENCLUS and MEFIA [16,31].

Table 1. IoT data sources for Data Mining

Mining Algorithm	**Goal**	**Data Source**
Clustering	1)Network performance enhancement. 2)Inhabitant action prediction. 3)Provisioning of the needed services . 4)Managing the plant zones . 5)Relationships in a social network.	wireless sensor , X10 lamp and home appliances, Raw location tracking data , GPSand sensor for agriculture ,RFID, smart phone, PDA, and so on
Classification	1)Device recognition. 2)Traffic event detection. 3)Parking lot management 4)Inhabitant action prediction 5)physiology signal analysis	RFID ,GPS, smart phone, and vehicle sensor,Passive infrared sensor,RFID, sensor, video camera, microphone, wearable kinematic sensor, wireless ECG sensor and so on .

D. Association

Association rule mining is the process of discovering frequent patterns, associations and correlations from data sets found in various kinds of databases such as relational, transactional or other data repositories. The association rule can be sequenced rule which generate frequent patterns such as the apriori algorithm and pattern growth algorithm.

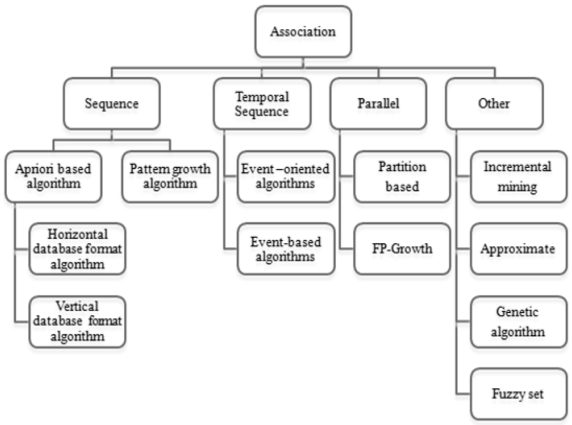

Figure 5. Data Mining Association .

*1)The Apriori Algorithm:*The apriori algorithm finds frequent itemsets in transactional databases using level wise iterative search.The algorithm first finds frequent-1 itemsets then frequent-2 and so on.It generates the frequent itemsets based on support confidence framework. The draw back of apriori is that it generates more number of candidates.Frequent patterns can be generated using apriori form two data formats as vertical data format and horizontal data format[31].

*2) Pattern Growth:*FP Growth:This algorithm overcomes the disadvantage of generating candidates and is used to generate frequent itemsets without candidate generation.It uses a

prefix-tree structure for storing the information of frequent patterns.

*3)Temporal sequence:*The temporal sequence databases will store data that relates to time instances.It stores the fact data.

*4)Genetic algorithms:*These algorithms follow greedy approach in discovering the association rules.It follows the process of natural selection where it is going to generate new population of strings from old population in a iterative manner.

*5)Fuzzy set:*The fuzzy set consists of the association rules which have sharp cut offs. The numerical attributes are converted to fuzzy attributes. For example the numeric value of income is converted to fuzzy attribute like high,medium,low.

E. Time Series

A time series indicates a collection of values gained by sequential computation over time. Time-series data mining calculated from natural ability to visualize the shape of data.

*1)Time Series Representation:*The model-based approach is based on the assumption that the data changes with respect to time parameter[28]. The goal is to find the parameters of such model representation. Two-time series models are considered to be similar if they are produced by the same set of parameters deriving the underlying model. There are several temporal parametric models which include statistical model by feature extraction, ARMA models, Markov chains.Markov chains are simpler as they fit well for shorter time series but their expressive power is limited[31].

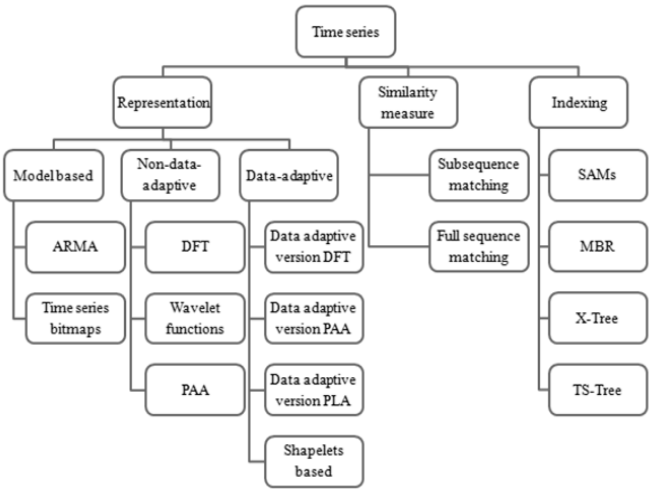

Figure 6. Data Mining Time Series

In non-data adaptive representation the transformation parameters remain the same for every time series data regardless of its nature[29]. The first non-data adaptive representation was drawn from spectral decomposition. The Discrete Wavelet Transform (DWT) was used in the seminal work of agrawal which projects the time series data as function of sine and cosine curves. This DWT uses the scaled and shifted versions of mother wavelet function which a gives a multi-resolution decomposition where low frequencies are measured over larger intervals and hence provide better accuracy[29,35]. Some of the wavelet functions include Haar,Daubechies and Coif lets.Other approaches more specific to time series have been proposed. The Piecewise Aggregate Approximation (PAA) introduced by Keogh represents a series through the mean values of consecutive fixed-length segments. An extension of PAA includes a multi-resolution piecewise Aggregate Approximation.

Data Adaptive approach, the parameters of the transformation are modified based on the data availability by adding a data sensitive selection step.All nondata adaptive methods can be transformed into data adaptive. For spectral decompositions, it usually consists of selecting a subset of the coefficients. Some of the data adaptive techniques include SVD, Sorted coefficients etc[30].

2) Similarity Measure:

The similarity between time series is measured based on the notion of shape. The similarity can be derived based on the characteristics of the series such as amplitude, scaling, temporal warping, noise and outliers. One of the methods for similarity is based on the Euclidean distance which calculated similarity based on the distances between the objects. There are various measures of similarity such as Minkowski and Manhatten.A similarity measure should be consistent with our intuition and provide the following properties.

- It should be able to recognize similar objects, even though they are not mathematically identical.
- The measure should be consistent.
- It should give importance to the most salient features on both local and global scales.
- It should be able to identify and distinguish arbitrary objects, without restriction on time series.
- A similarity measure should be universal in the sense that it allows to identify or distinguish arbitrary objects, that is, no restrictions on time series are assumed.
- It should be resilient to distortions and be invariant to set of transformations.

- It should abstract from distortions and be invariant to a set of transformations.

3) Indexing

An indexing allows fast accessing of the data in large databases. A number of indexing techniques have been proposed. Most of the techniques involve a dimensionality reduction in order to index time series using spatial access method. In this method the time series X is represented as a point in n-dimensional space. The space is partitioned into regions of hierarchies for efficient retrieval.The B-Tree is one of the hierarchical indexing structure which works on one-dimensional data.Mutidimensional indexing structures include R-tree which uses data organized in Minimum Bounding Rectangles(MBR).The problem with Minimum Bounding Rectangles is that when we summarize the data in MBR, the sequential nature of time series cannot be recorded. Moreover the R-trees suffers from overlap.The X-tree (extended node tree), uses a different split strategy to reduce overlap.The A-tree (approximation tree) uses VA-file-style (vector approximation file) quantization of the data space to store both MBR and VBR (Virtual Bounding Rectangle)lower and upper bounds[30,31].

IV. THE IOT AND BIG DATA CHALLENGES

Speed improvement of IoT, the graph of data generation in the worldwide increased lot .this dramatically change in the technology of IoT created new research mainly in data generating ,transmitting ,storing ,mining and analysis the data. IoT data created in big volume , that data in variety in formats .For such kind of data need of managing with big data tools[7,18] .The data of IoT is in stream format in real time ,so that we need to emerging mining tools which should be relative to big data tools[9]. When we go into the big data many challenges regarding about data storing analytics will arise depends upon qualities of the knowledge in IoT.

Large volumes of data generated, captured from IoT devices .A single sensor unit data is near to 20 bytes.Here data produce in a continues manner from several sensors for one single machine.Whenever in an ITS system consisting of millions of sensors work in real time[24].Per day they generate TB of data, some applications can able to produce data in PB(PetaBytes) and ZB (zeta byte).

IoT generated data, not inhomogeneous ,data is not in same formats .different kind of data in different formats will generate from various devices[17].Such data might be structured ,semi-structured and unstructured.For example information from sensors , the data stream from cameras, data from web-based system and data from statistical measure devices etc.So that with respect to the data format we need to store and mine the data.To engage the data of IoT we need to maintain complex structures[19,24] . If IoT network is create with less than 10 sensors then we no need to prefer for distributed environment .If it is in a big network like smart cities then signals ,data and connections are continuously pinging .here we need to use of distributed computing .which goes to new research from cloud computing .

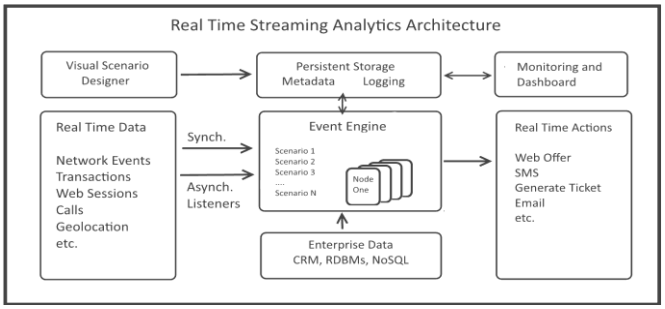

Figure 7. Real Time Streaming Analytic Architecture

Big Data storage operation is done by the HDFS of Hadoop.Generally, HDFS consisting of the network of data nodes. They arranged in a racks manner which can stride any type of data value but they consider that data segment minimum memory is 64MB .but IoT created individual data element size nearby 20 bytes. Then we need to understand the which kind of tool and algorithm is the need for storage arrangement when data segment size is very small.

Data mining process continues with a linear process integrated with few number of tools.In generally IoT needs of mining paradigm as well analytical paradigm .according to an application we need to assess the right tool for IoT data extraction.

V. DATA MINING AND BIG DATA RELATIVITY FOR IOT

Basically the IoT data is stream-oriented data. The devices of IoT continuously produce a different kind of data. Such data may be structured,semi-structures and unstructured format. IoT systems network has to be connected to Cloud in a big network like ITS, smart cities and smart grids etc .without cloud interaction we can't communicate the devices in real time streaming data process pool[21].Either side of analysis ,In a small network like vehicles and smart Home (let us consider if they are not connected to other network and we want to connect the devices within their architecture) devices no need of interact with distributed computing technologies.Even they will produce the streaming data and we have to do all stream data management operations.Traditionally, most of the information processing in the Internet needs fall under online analytical processing

(OLAP).OLAP performs the all mining tasks using a simple procedure called ELT (extract, transform and load)[15,27]. Several data sources compute their data aggregation, cube materialization ,filtering etc association with few number of tools in OLAP. this methodology follows the descriptive statistics which generate the output inject into visualization devices, dashboards. We can satisfy this when the data size medium range. Huge data analysis will require more advanced tools and mechanisms which they might be require the predictive analysis including machine learning computing .

In web-scale data model data intelligence have adapt to parallel and distributed computing. MapReduce is atandard programming paradigm in distributed analytical environment .This programming mostly using in Hadoop open source tool Hadoop and its ecosystem such as Mahout have to be very successful platform in the process huge data.not only IoT many other applications are generating in the form of a stream. Batch data is small snapshot of streaming data which can access in a small interval of time[22,23] .So that more researches concentrated on abstraction of streaming model. In this streaming model information arrives at high speed and data processing algorithms should be process it in one pass interval under designed constraints of space and time[26] .

On the one hand, MapReduce is not suited to express streaming algorithms. but in this progress MapReduce does not suited for streaming algorithms[20,21] .it is limited by memory ,bandwidth of machine . By combining the Examples of these engines include Storm, S4 and Samza.

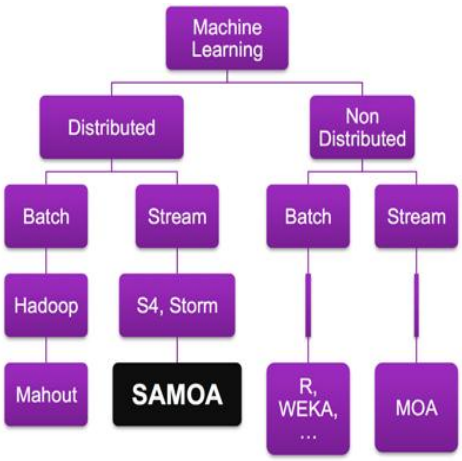

Figure 8. Machine learning Architecture

A) Distributive Computing

Mahout: Mahout is scalable machine learning algorithms focusing on the areas of Clustering and classification. Mahout is having Java libraries for common maths operations like linear algebra and statistics. Using Mahout clustering, classification and batch depended filtering is processed on top of Hadoop using the MapReduce framework.

*1)Hadoop:*Hadoop is used for processing the huge data of big data in distributed storage.Hadoop is an open source software .the processing of data done by MapReduce programming framework. Hadoop consists two operations one is storing the huge data in manageable format another one is analytics the huge volume of data[21,22,37]. the storage part of Hadoop is called as Distributed File System (HDFS). Hadoop data analytical processing part done by the MapReduce programming. Hadoop is open source software and integrated with many tools.

*2)SAMOA (*Scalable *Advanced Massive Online Analysis):* SAMOA which is the integration of several algorithms to mining the stream data in Big Data.Unlike MOA, it can perform the mining in the distributed environment.SA MOA is an integration of classification, clustering and regression processes with mining of frequent itemset and frequent graph.Mostly the running stream data create the stream data environment such kind of data implemented with XML based notations and XML notations implemented in SAMOA[39]. SAMOA framework can process the mining in distributed computing[38].

3)S4: S4 is r real-time, distributed,event-driven and decentralized, processing the streaming data continuously[37]. S4 has a scalable and symmetric architecture which make all the data nodes or devices in a cluster are identical.S4 architecture different from classic master-nodes architecture. S4 uses the ZooKeeper as the communication layer to interact the nodes within the nodes of a cluster. But, this process is completely transparent to users, creation of S4 is implemented with a combination of MapReduce and Actors model. The process of S4 is performed by Processing Elements (PEs).IoT real-time streaming data is streamed between PEs[30][40].

4)Storm: Storm is the open source data streaming framework which integrated with other queuing system and also bandwidth systems.The storm also integration of other tools like ZooKeeper ,Nimbus, processing nodes supervisor.Storm represents the data flow model which shows the continues data flow through network entities[23,38] . That continues data flow is called as stream data which is having the sequence of tuples. Storm recognize the data in the sequence of tuples.Storm operate on big volume of data with respect to memory speed. So that Storm requires the interfaces for data sources and storage nodes For better output Storm

combined with HBase to improve the real-time architecture for processing streaming data.

B) Distributive Computing

1)MOA: In the non distributed system the Stream data mining performing by MOA in real time which is an open source software. MOA is an integration of classification, clustering and regression processes with mining of frequent itemset and frequent graph.Mostly the running stream data create the stream data environment such kind of data implemented with XML based notations and XML notations implemented in MOA[38]. MOA framework process the mining in the single machine.

2)R-Language: R language used for statistical computing .it is an open source language.R-language can implement in enterprises, cloud. It can capable to do analysis and visualization up to 16TB in one iteration[38,39] .R-language mining is Text Mining , Social Network and Graph mining also perform the data manipulations.

3)WEKA: WEKA is a collection of machine learning algorithms and it is an open source application. WEKA is pure java based platforms their database connection created by JDBC with any RDBMS package .primitive tasks including data pre-processing Classification, regression , clustering , association rules and visualization. It allows GUI for the interface , and data visualization and analysis.

VI. CONCLUSION

The Internet of Things created new research and challenges in the area of data mining and analysis. We know that IoT data is in stream data paradigm which evaluates in real time. stream data processing need of special advanced tools for IoT data management because create data in huge size. Such data analysis might be in the predictive analysis including machine learning computing. Predictive nature of data analysis can adapt to distributed and non-distributed computing procedure. According to data size, technology which implemented and machine which adapted we need to perform that Data mining and data analysis for KDD. Several tools are available for data mining relative with Big Data analytical tools.In this paper we discussed many tools to manage the stream data which created by IoT.For this initially, we started with a model which tell about IoT data generating points, the format of data in the application interface, mining procedure for stream data and data representation.We also focused on several tools related to distributed computing for mining S4 ,Mahout ,Storm ,Hadoop and SAMOA as well for non distributing computing related mining tools like R-language ,WEKA ,MOA .From this paper, we went to say about stream data mining in the real time with respect to predictive manner.In future, we to continue the same research and survey on stream data for analytics.IoT technology is progressively increasing meanwhile it is creating more challenges about data management.In the Current generation, there is no common solution for mining big data streams, nothing but there is no reliable tool for accommodating the executing data mining as well as at the same time machine learning algorithms on a distributed stream processing engine. Our research will also concentrate on the data mining and machine learning algorithms on distributed stream processing.

REFERENCES

[1] C.-W. Tsai, C.-F. Lai, and A. V. Vasilakos, "*Future internet of things: open issues and challenges*," Wireless Networks, vol. 20, no. 8, pp. 2201–2217, 2014.

[2] A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and C. A. C. Coello, "*A survey of multiobjective evolutionary algorithms for data mining: part* I," IEEE Transactions on Evolutionary Computation, vol. 18, no. 1, pp. 4–19, 2014.

[3] J. Zhang, B. Iannucci, M. Hennessy, K. Gopal, S. Xiao,S. Kumar, D. Pfeffer, B. Aljedia, Y. Ren, M. Griss et al.,"*Sensor Data as a Service–A Federated Platform for MobileData-centric Service Development and Sharing*," in ServicesComputing (SCC), 2013 IEEE International Conference on.IEEE, 2013, pp. 446–453.

[4] Y. Zhang, M. Chen, S. Mao, L. Hu, and V. Leung, *"CAP: crowd activity prediction based on big data analysis,"* IEEE Network, vol. 28, no. 4, pp. 52–57, 2014.

[5] L. Ramaswamy, V. Lawson, and S. Gogineni, "Towards a Quality-centric Big Data Architecture for Federated Sensor Services," in Big Data (BigData Congress), 2013 IEEE International Congress on, June 2013, pp. 86–93.

[6] S. Haller, S. Karnouskos, and C. Schroth, "*The Internet of Things in an enterprise context,"* Future Internet Systems (FIS), LCNS, vol. 5468. Springer, 2008, pp. 14-8.

[7] Shahid Tufail, M. Abdul Qadeer, "*Cloud Computing in Bioinformatics: Solution to Big Data Challenge*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.9, pp.232-236, 2017.

[8] M. Chen, S. Mao, Y. Zhang, and V. Leung, *Big Data: Related Technologies, Challenges and Future Prospects*, SpringerBriefs in Computer Science, Springer, 2014.

[9] J. Wan, D. Zhang, Y. Sun, K. Lin, C. Zou, and H. Cai, "*VCMIA: a novel architecture for integrating vehicular cyber-physical systems and mobile cloud computing*," Mobile Networks and Applications, vol. 19, no. 2, pp. 153–160, 2014

[10] X. H. Rong, F.Chen, P. Deng, and S. L.Ma, *"Alarge-scale device collaboration mechanism,"* Journal of Computer Research and Development, vol. 48, no. 9, pp. 1589–1596, 2011.

[11] F. Chen, X.-H. Rong, P. Deng, and S.-L.Ma, "*A survey of device collaboration technology and system software*," Acta Electronica Sinica, vol. 39, no. 2, pp. 440–447, 2011.

[12] L. Zhou, M. Chen, B. Zheng, and J. Cui, "*Green multimedia communications over Internet of Things,"* in Proceedings of the IEEE International Conference on Communications (ICC'12), pp. 1948–1952, Ottawa, Canada, June 2012.

[13] P. Deng, J. W. Zhang, X. H. Rong, and F. Chen, "*A model of large-scale Device Collaboration system based on PI-Calculus for green communication,"* Telecommunication Systems, vol. 52, no. 2, pp. 1313–1326, 2013.

[14] P. Deng, J. W. Zhang, X. H. Rong, and F. Chen, "*Modeling the large-scale device control system based on PI-Calculus*," Advanced Science Letters, vol. 4, no. 6-7, pp. 2374–2379, 2011.

[15] J. Zhang, P. Deng, J. Wan, B. Yan, X. Rong, and F. Chen, "*A novel multimedia device ability matching technique for ubiquitous computing environments,"* EURASIP Journal on Wireless Communications and Networking, vol. 2013, no. 1, article 181, 12 pages, 2013.

[16] G. Kesavaraj and S. Sukumaran, "*A study on classification techniques in data mining,"* in Proceedings of the 4th International Conference on Computing, Communications and Networking Technologies (ICCCNT '13), pp. 1–7, July 2013.

[17] S.Babu, G.Fathima , "*Decision Trees for Mining Data Streams Based on the Gaussian Approximation*", International Journal of Computer Sciences and Engineering, Vol.4, Issue.3, pp.35-38, 2016.

[18] Q. Jing, A. V. Vasilakos, J. Wan, J. Lu, and D. Qiu, "*Security of the internet of things: perspectives and challenges*," Wireless Networks, vol. 20, no. 8, pp. 2481–2501, 2014.

[19]] B. Chandra and P. P. Varghese, "*Fuzzy SLIQ decision tree algorithm*," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 38, no. 5, pp. 1294–1301, 2008.

[20] J. Shafer, R.Agrawal, and M.Mehta, "*SPRINT: a scalable parallel classifier for data mining*," in Proceedings of 22nd International Conference on Very Large Data Bases, pp. 544–555, 1996.

[21] Chang, V., 2015. *Towards a big data system disaster recovery in a Private cloud*. Ad Hoc Networks, 000, pp.1–18.

[22] [23]Sakr, S. & Gaber, M.M., 2014. *Large Scale and big data: Processing and Management Auerbach*, ed., Schilling, D.R., 2014. Exaflop Computing Will Save the World ... If We Can Afford It - Industry Tap.

[23] González-Martínez, J. a. et al., 2015. *cloud computing and education: A state-of-the-art survey*. Computers & Education, 80, pp.132–151.

[24] D. Matolak: *Channel Modeling for Vehicle-to-Vehicle Communications*. IEEE Comm.Mag., pp. 76-83, May 2008

[25] Palaghat Yaswanth Sai, Pabolu Harika, "*Illustration of IOT with Big Data Analytics*", International Journal of Computer Sciences and Engineering, Vol.5, Issue.9, pp.221-223, 2017.

[26] R.S. Barga, J. Ekanayake, W. Lu, *Project Daytona: Data Analytics as a Cloud Service*, in: A. Kementsietsidis, M. A. V. Salles (Eds.), Proceedings of the International Conference of Data Engineering (ICDE 2012), IEEE Computer Society, 2012, pp. 1317–1320.

[27] Zhang, X., Wu, J., Yang, X., Ou, H., And Lv, T. 2009. *A novel pattern extraction method for time series classification*. Optimiz. Engin. 10, 2, 253–271.

[28] Zhong, S.,Khoshgoftaar, T., And Seliya, N. 2007. *Clustering-Based network intrusion detection*. Int. J. Reliab.Qual. Safety Engin. 14, 2, 169–187.

[29] Yankov, D., Keogh, E., And Rebbapragada, U. 2008. *Disk aware discord discovery: Finding unusual time series in terabyte sized datasets*. Knowl. Info. Syst. 17, 2, 241–262.

[30] Chhieng, V. And Wong, R. 2010. *Adaptive distance measurement for time series databases*. In Lecture Notes in Computer Science, vol. 4443. Springer, 598–610.

[31] Ouyang, R., Ren, L., Cheng, W., And Zhou, C. 2010. *Similarity search and pattern discovery* in hydrological time series data mining. Hydrol. Process. 24, 9, 1198–1210.

[32] Fuchs, E., Gruber, T., Pree, H., And Sick, B. 2010. *Temporal data mining using shape space representations of time series*. Neurocomput. 74, 1-3, 379–393.

[33] Assent, I., Wichterich, M., Krieger, R., Kremer, H., And Seidl, T. 2009. *Anticipatory DTW for efficient similarity search in time series databases*. Proc. VLDB Endowm. 2, 1, 826–837.

[34] Barone, P., Carfora, M., And March, R. 2009. *Segmentation, classification and denoising of a time series field* by A Variational Method. J. Math. Imag. Vis. 34, 2, 152–164.

[35] Gullo, F., Ponti, G., Tagarelli, A., And Greco, S. 2009. *A time series representation model for accurate and fast similarity detection*. Pattern Recogn. 42, 11, 2998–3014.

[36] Reeves, G., Liu, J., Nath, S., And Zhao, F. 2009. *Managing massive time series streams with multi-scale compressed trickles.* Proc. VLDB Endow. 2, 1, 97–108.

[37] Thu Vu, G. De Francisci Morales, J. Gama, and A. Bifet, "*Distributed Adaptive Model Rules for Mining Big Data Streams*," in BigData '14: Second IEEE International Conference on Big Data, 2014.

[38] J. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "*A Survey on Concept Drift Adaptation*," ACM Computing Surveys, vol. 46, no. 4, 2014.

[39] A. Bifet, G. Holmes, R. Kirkby, and B. Pfahringer, "*MOA: Massive Online Analysis*," The Journal of Machine Learning Research, vol. 11, pp. 1601–1604, 2010.

[40] Charu C. Aggarwal, Jiawei Han, Jianyong Wang, and Philip S. Yu. *A framework for clustering evolving data streams*. In VLDB '03: 29th International Conference on Very Large Data Bases, pages 81–92, 2003.

[41] D. Schall, M. Aiello, and S. Dustdar, "*Web services on embedded devices*," International Journal of Web Information Systems, vol. 2, no. I, pp. 45-50, 2006.

[42] L. Johnsrud, D. Hadzic, T. Hafsoe, F. T. Johnsen, and K. Lund, *"Efficient Web Services in Mobile Networks*;' in Proceedings of 2008 Sixth European Conference on Web Services. IEEE, Nov. 200S, pp. 197-204.

[43] N. Priyantha, A. Kansal, M. Goraczko, and F. Zhao, *"Tiny web services: design and implementation of interoperable and evolvable sensor networks,"* in Proceedings of the 6th ACM conference on Embedded network sensor systems. ACM, 200S, pp. 253-266.

[44] S. Nastic, S. Sehic, M. Vögler, H.-L. Truong, and S. Dustdar, "*PatRICIA -- A Novel Programming Model for IoT Applications on Cloud Platforms*," 2013 IEEE 6th International Conference on Service-Oriented Computing and Applications, 2013, pp. 53–60.

