Towards modeling and optimization of features selection in Big Data based social Internet of Things

Awais Ahmad, Murad Khan, Anand Paul, Sadia Din, M. Mazhar Rathore, Gwanggil Jeon, Gyu Sang Chio

PII: S0167-739X(17)30453-3
DOI: https://doi.org/10.1016/j.future.2017.09.028
Reference: FUTURE 3683

To appear in: *Future Generation Computer Systems*

Received date : 21 March 2017
Revised date : 27 June 2017
Accepted date : 10 September 2017

Towards Modeling and Optimization of Features Selection in Big Data based Social Internet of Things

Awais Ahmad, Murad Khan, Anand Paul, Sadia Din, M. Mazhar Rathore, Gwanggil Jeon, Gyu Sang Chio

***Abstract*— The growing gap between users and the Big Data analytics requires innovative tools that address the challenges faced by big data volume, variety, and velocity. Therefore, it becomes computationally inefficient to analyze and select features from such massive volume of data. Moreover, advancements in the field of Big Data application and data science poses additional challenges, where a selection of appropriate features and High-Performance Computing (HPC) solution has become a key issue and has attracted attention in recent years. Therefore, keeping in view the needs above, there is a requirement for a system that can efficiently select features and analyze a stream of Big Data within their requirements. Hence, this paper presents a system architecture that selects features by using Artificial Bee Colony (ABC). Moreover, a Kalman filter is used in Hadoop ecosystem that is used for removal of noise. Furthermore, traditional MapReduce with ABC is used that enhance the processing efficiency. Moreover, a complete four-tier architecture is also proposed that efficiently aggregate the data, eliminate unnecessary data, and analyze the data by the proposed Hadoop-based ABC algorithm. To check the efficiency of the proposed algorithms exploited in the proposed system architecture, we have implemented our proposed system using Hadoop and MapReduce with the ABC algorithm. ABC algorithm is used to select features, whereas, MapReduce is supported by a parallel algorithm that efficiently processes a huge volume of data sets. The system is implemented using MapReduce tool at the top of the Hadoop parallel nodes with near real-time. Moreover, the proposed system is compared with Swarm approaches and is evaluated regarding efficiency, accuracy and throughput by using ten different data sets. The results show that the proposed system is more scalable and efficient in selecting features.**

***Index Terms*— SIoT, Big Data, ABC algorithm, feature selection.**

I. Introduction

THE most important paradigm that fills the gap between physical world and cyber world is the Social Internet of Things (SIoT). Recent advancement in the field of IoT has led to the digitization of the physical world where the configuration of novel applications and services are high in demand. For such advancements, a variety of things is grouped together to share information with the help of Internet. These things include radio frequency identification (RFID) tags, sensors, actuators, mobile equipment, computers, medical sensors, etc. and are connected to each other through wires or wirelessly. The evolved things in the SIoT can sense the physical environment, collect data, transfer or disseminate data, process data for appropriate applications, and communicate with other things. Hence, SIoT came up with a power technology that helps in understanding the physical world and to response to outer stimuli. Hence, an ultimate solution that gives us insight to the real-world in real-time.

Apparently, advancements in the field of SIoT pose new challenges when it comes to implementation [1]. Since SIoT is a mixture of heterogeneous things, making it quite different from traditional networks, thus it becomes more complex due to its scalable property [2]. As a result, the evolved things cannot directly apply to SIoT. Moreover, given the complex, heterogeneous nature of SIoT, various things communicating with each other under its umbrella consume a high volume of memory, processing power, and high bandwidth. Thus, IoT tends to generate a huge volume of data, referred to Big Data. The term Big Data is categorized into some specific types of data sets, which comprises formless data, which resides in the data layer of the technical computation applications as well as the web. Usually, Big Data comprises 3 V's, such as Volume (referred to as the size of the data sets), Velocity (referred to as high-speed processing and analyzing) and variety (referred to as different data sources, i.e., heterogeneous networks). To

This research was supported by Kyungpook National University Bokhyeon Research Fund, 2015.

Awais Ahmad is with the Department of Information and Communication Engineering, Yeungnam University, Korea (e-mail: aahmad.marwat@gmail.com).

Murad Khan is with the Department of Computer Science, Sarhad University of Science and Information Technology, Peshawar, Pakistan (e-mail: muradkhan23@gmail.com).

Anand Paul is with the School Of Computer Science and Engineering, Kyungpook National University, Daegu, 702-701, Korea (e-mail: paul.editor@gmail.com).

Sadia Din is with the School Of Computer Science and Engineering, Kyungpook National University, Daegu, 702-701, Korea (e-mail: research.2486@gmail.com).

M. Mazhar Rathore is with the School Of Computer Science and Engineering, Kyungpook National University, Daegu, 702-701, Korea (e-mail: rathoremazhar@gmail.com).

Gwanggil Jeon is with Department of Embedded Systems Engineering, Incheon National University, Incheon, Korea (email: gjeon@inu.ac.kr)

Gyu Sang Chio is with the Department of Information and Communication Engineering, Yeungnam University, Korea (e-mail: castchoi@ynu.ac.kr).

cope with such constraints, the ideal solution is the green IoT. IoT can minimize emissions and pollution by exploiting environmental surveillance that is used to decrease the operational cost as well as power consumption [1, 3-6]. Consequently, the most crucial challenge is how to effectively reduce the costs as well as power consumption of things in SIoT - the primary focus of this paper.

Moreover, in the current scenario of Big Data, various standards and platforms have been introduced by relational database vendors. These are used for data aggregation as well as data analysis. These platforms are either software or simply they just provide analytical services (usually runs on third party servers). These techniques are unable to select appropriate features in Big Data.

Therefore, based on the explanations above of Big Data analytics, feature selection is one of the core issues in this technology. Feature selection includes image classification, cluster analysis, data mining, pattern recognition, image retrieval, etc. [7]. However, features selection is a quite decisive technique to efficiently analyze Big Data, in which one subset from the Big Data is considered to remove irrelevant, noise, and redundant features. Moreover, the mentioned tasked minimizes computational complexity and cost, while enhancing the accuracy of data analysis.

Among feature selection algorithms, various scheme are proposed that are classified into two broad categories, i.e., filter approaches [8-13], and wrapper approaches [14-17]. In filter based techniques, before classification process, the filtration process is performed since their independent nature of usage of classification algorithms [17]. Moreover, in this approach, a weight value is computed for each feature, so that features with better values can be selected to represent original Big Data sets. Apparently, wrapper technique produces a set of nominee feature by altering (adding and removing) features to compose a subset of features. Afterward, accuracy is employed to evaluate the outcome of the feature set. The later technique outperforms filter technique in their results. Also, other evolutionary methods, such as Ant Colony Optimization (ACO) [9, 11, and 18], Particle Swarm Optimization (PSO) [19], Bat Algorithm [20], and Artificial Bee Colony (ABC) [21] are also proposed that increases the computation efficiency. Apparently, other traditional mechanisms have also been seen in active storage systems that address the Input/Output (I/O) bottleneck challenges for scientific application. Such potential is increased with the increase in the demand for the data. Though, the prototype of the active storage [23-24], the primary focus is given to read-intensive operations. Since it provides an easy way to identify common operations, i.e., lookup, amongst several data analysis approaches, kernel analysis is predefined in libraries, known as processing kernels [24]. On the other hand, various other techniques for writing-intensive applications are not well addressed that are common in scientific areas as well as feature selection. Since rapid advancement in the size of the output, write performance operations I/O systems becomes more important [25]. Moreover, due to incredible growth in this scientific application, several other challenges are noticed. These challenges include storing huge amount of data and memory allocation to this application, processing and analyzing these data without having intelligent technique. Several other techniques, such as prototype reduction (PR) is one of the useful remedies for class distribution. Prototype reduction splits the original data into different subsets, which can be individually addressed. After that, PR combines each moderately compact set into a global solution. In addition, torrents of event data are essential to be distributed among various databases, which required large mining algorithms needed to be distributed in the networks of the computer. However, the design of existing active storage and feature selection algorithms have some major drawbacks, which are i) it is difficult to extract features in real-time continuous data, ii) to process kernel design pattern is not suitable for write operations, ii) usually, write operations require more computation power than read operations, and iii) to process massive volume of data is sometimes hard to processes by using some traditional intelligence and processing tools.

A. RESEARCH MOTIVATION

In this world, various devices are connected to other devices and things with the help of Internet, 3G/4G, wireless LAN, etc. This enables a rich infrastructure for Internet of Things that aims to connect various things (for instance, cellular phones, wireless body area network, Wi-Fi, access points, etc.) with distinctive addresses and allows these devices to interact with each other in an efficient way, hence generating Big Data. Thus, the primary goal is to select best and optimal features in the SIoT Big Data, so that to reduce the energy consumption involved in the SIoT during communicating Big Data over the Internet.

In this paper, while aiming for a selection features, we focus on SIoT Big Data and present various technologies toward IoT. Specifically, we propose a system architecture that exploits IoT for the different application, such smart cities, smart home, traffic management, healthcare system, which is a novel paradigm in modeling and optimizing Big Data in IoT. The system architecture is used to aggregating Big Data, exploiting feature selection algorithm and forward the data toward Hadoop ecosystem. Also, the IoT-based Big Data architecture is based on feature selection. Thus, feature selection in Big Data using convolution method is considered. Such feature selection aspect pitches us against the problem of optimization. We intend to solve this problem of an optimizing feature selection by considering Ant Colony Optimization technique. Eventually, future directions and open challenges are discussed regarding SIoT in 5G network. To the best of our knowledge, this work is the first that exploits the realization of the feature selection for IoT system given the Big Data. We hope and believe that this work will be useful for the SIoT systems and will provide state-of-the-art guidance for research vis-à-vis SIoT and big Data.

B. RESEARCH CONTRIBUTION

The main contribution of this paper is as follows.

- At first, we present a hierarchical framework for the extracting feature in IoT Big Data. The framework first internments the scalable features of IoT, which helps in the extension of network commodities.

- The exploitation of Kalman filter is used to enhance the efficiency of big data analysis in a real-time environment. Moreover, the proposed ABC algorithm assists the architecture in extracting features in Big Data.
- The proposed data feature selection algorithm based on ABC improve the accuracy of the system. We have tested ten data sets with the proposed feature selection based on ABC to check the performance of the system in more explainable mode. Moreover, the proposed scheme is compared with well know Swarm approaches to test its capabilities on various platforms.

C. *ORGANIZATION*

The rest of the paper is organized as follow. Section II briefly describes the latest background and related work. Section III presents the detailed description of the proposed scheme that includes system architecture of Hadoop-based ABC. Section IV comprises extensive simulations of the proposed scheme. And finally, Section V offers a conclusion.

II. Background And Related Work

Feature selection is a process that is responsible for the election of a feature's subset. These can be labeled as a search into a state space, where full search technique can be applied in which all the spaces are traversed. This approach seems impractical for identifying a large number of features. Some of the related technique in which heuristic approach considers features, which cannot be selected at each iteration for evaluation. On the other hand, random search engenders random subsets in the search space. There are several bio-inspired, and genetic algorithms use such techniques [17, 18, and 19].

As mentioned in the earlier section, several evolutionary methods are used for feature selection, in which Swarm Intelligence technique can be found for feature selection. A scheme based on rough set approach along with ABC algorithm [26]. The proposed algorithm is used for reducing the dimension of different medical data sets. Apparently, the same algorithm is used for feature selection using neural networks. The major drawbacks of the above two mentioned techniques are the computational efficiency and cost. When there is increase the data sets the computational efficiency decreases and cost increases.

To use some conventional methods using Hadoop ecosystem, usually, the framework of Hadoop MapReduce run over Hadoop Distributed File Server (HDFS), which has the advantage of multiple local disks on a computer node providing better data-locality [27]. Though, majority of the HPC clusters [28, 29] used to follow traditional Beowulf architecture [30, 31]. In such systems, the computer nodes are provided with a very light weight operating system, or sometimes with a limited capacity of local storage [32]. Simultaneously, they are all connected to a parallel file system, called as Lustre. Lustre provides an efficient and scalable data storage facility. Figure 1 delineates an example scenario of deployment of Lustre system and the operation of YARN MapReduce on modern HPC clusters.

The major drawback of this architecture has the limited capacity of local disks since they inhibit the working of MapReduce on a large data sets. These inconsistencies lead toward attenuation of MapReduce running on HPC clusters. Moreover, recent studies also verify that MapReduce does not provide significant results when it combines with HPC cluster [10] [31, 32]. These limitations lead us toward a question whether storage system with Lustre gives us local storage capabilities that facilitate MapReduce, which gives us efficient results on HPC clusters?

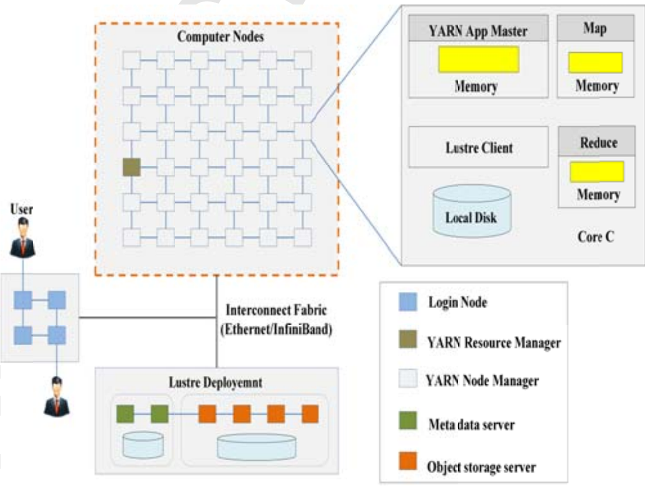

Figure 1. YARN MapReduce running over a typical Lustre

The majority of the Lustre system installed on HPC cluster using Lustre as local storage. These local storage are for traditional MapReduce functions where these functions can be completed in two steps, i.e., read and write operations. These operations give us high-speed data shuffles path since read and write have high throughput on Lustre systems. Though, the time required for a transmission inside Lustre depends on many factors, such as interconnection of clusters, data load, and other variations, etc. These factors, when combined, generate an overhead on the traditional MapReduce functions [27]. Moreover, recent studies have also been proposed to enhance the function of MapReduce design [33, 34, and 35] to speed up the function of MapReduce. However, these systems are facing some other limitations related to the local memory, processing big data sets, dividing the job of the map and reduce function, results in storage in a real-time scenario, etc.

MapReduce programming paradigm is producing a large amount of datasets that is responsible for the extensive diversity of real world tasks [36]. MapReduce divides input data into small independent chunks which dealt in a parallel manner completely. The MapReduce architecture classifies the maps outputs and sends to the reduce job. The input and output of the task are kept in the file system. MapReduce is a parallel programming model, which perform three main task at the same time, i.e., simplicity, load balancing, and fault tolerance. The Google File System (GFS) normally inspired from the

MapReduce model gives the reliability and efficiency of data storage required for large databases applications [37].

MapReduce model motivated by functional languages. Functional languages have a map and reduce primeval exist in functional languages. Depending on the framework requirement numerous execution can be feasible in MapReduce platform. Few recently executions are existed in literature work e.g. networked machines clustering [36], shared memory multi-core system techniques [38, 39], graphic processors and asymmetric multi-core processors approach [40].

Google launched one of the most famous implementation that exploit a huge number of clusters of computers, which are joint through switch Ethernet. Google MapReduce scheme reduce the cost of clusters of machines for wide range distributed applications. MapReduce approach makes simpler and easier the establishment process. It is based on real-time execution, and it does not define preplan execution scheduling of the node [41].

MapReduce paradigm can perform a parallel execution on distributed nodes. The core purpose of MapReduce model is to clarify huge data implementation as well as low-cost cluster machines. It is also obtained fault toleration and balancing the load for each cluster to make this simpler and easier for operators. Map and Reduce are two basic entities of the MapReduce model. Google is the owner of indigenous MapReduce, so it is not for public use [41]. Even though, the idea of MapReduce primeval simply the distributed computing system. The original MapReduce structure is very important to obtain required and efficient performance [42]. The Google's MapReduce Structure has originally distributed file system which identifies the data location and accessibility [36].

In addition to that, Particle Swarm Optimization technique is another remedy used for feature selection and processing of large data sets. The technique is proposed for feature selection as filter method [17] or wrapper method [43 - 42]. A wrapper method is proposed based on BAT algorithm with OPF classifier. The problem of increase in complexity and efficiency of the system decreases using above techniques. Also, it takes more time to extract the features from large data sets. Sometimes, it also happens that removal of noise algorithms affects the original data sets, and inappropriate data is considered for feature selection.

Therefore, based on the aforementioned related techniques of convolution methods and traditional Hadoop techniques, system architecture is required that select the best and optimal features from the massive volume of data. For this purpose, this paper proposes a technique based on ACO algorithm in Hadoop ecosystem. ACO algorithm efficiently selects the best and optimal feature, whereas Hadoop ecosystem combines with ACO to generate the best and efficient results.

III. Proposed Scheme

This section comprises IV-Tier layered architecture that supports ABC algorithm based on Hadoop ecosystem. Afterward, we provide a detailed working of ABC algorithm in Hadoop environment for IoT.

A. *IV-Tier Layered Architecture*

The multilevel active storage and processing aim to extract features from Big Data. The system architecture is composed of four layers. Each layer is supported by different functionalities enables read and write operations high effectively. In this section, we will first introduce the layered architecture that supports complete system design for feature extraction. Afterward, we will present the design and working of the designed system.

Based on the needs of analyzing big data, we propose an IV-Tier architecture model as shown in Figure 2. The designed model assists different objects to interact with each other using the shared medium. The proposed architectural model integrate various data generated by difference application, under the same domain, i.e., internet of things, which supports the research community to provide the generalized framework and architecture that can help the domestic users in the case of security, healthcare, elderly age people and kids, and transportation system, machine-to-machine network, wireless sensor network, and vehicular network, etc. As Figure 2 shows that the proposed IV-Tier architectural model consists of four layers.

Tier I: Data Generation handles data generation through various objects and then collecting and aggregating that data. Since a different number of objects are involved in generating the data. Therefore, an enormous number of heterogeneous data is produced with various formats, a different point of origin, and periodicity. Moreover, various data have security, privacy, and quality requirements. Also, in sensor's data, the Metadata is always greater than the actual measure. Therefore early registration and filtration technique are applied at this layer, which filters the unnecessary Metadata, as well as redundant data, is also discarded.

Tier-II: This layer provides end-to-end connectivity to various devices. Moreover, data is aggregated at this point generated from different devices and arrange them in the proper format.

Tier-III: Feature Extraction and Processing Layer is the primary layer of the whole system architecture, which handles feature extraction and processing of data. Since we need a real-time stream of the data and offline data analysis. Therefore, we need a third party real-time tool to combine with the processing server to provide the real-time implementation. To provide real-time implementations, Strom, Spark, VoltDb, and Hupa can be used. For instance, to be very specific in the case of data analysis, the implementation part could be achieved by using MapReduce, whereas, for feature extraction, the incorporation of ABC algorithm assists the proposed scheme in the better acquisition of features from large data sets. At this layer, the same structure of MapReduce and HDFS is used. With this system, we can also use HIVE, HBASE, and SQL supposed for managing Database (in-memory or Offline) to store historical information.

Tier-IV: Service layer is the lowermost layer responsible for incorporating the third party interfaces to objects and human. This layer can be used autonomously as a single site, merged with other locations, or deployed in cloud interface. There are different other features as well. For instance, the unique global ID management is the key element in the application layer that handles identifying the object throughout the universe. Vendor control is another feature deals with the definition of the activities duly performed by different objects. The proposed architectural layers involve different objects that need intelligent power to interact with a human. For this reason, a smart algorithm is required at the application level that could efficiently and effectively interact with the human. Various tasks could be performed by these features, such as request generator, session initiating, setting up communicating rules, interact with heterogeneous objects and terminating the session.

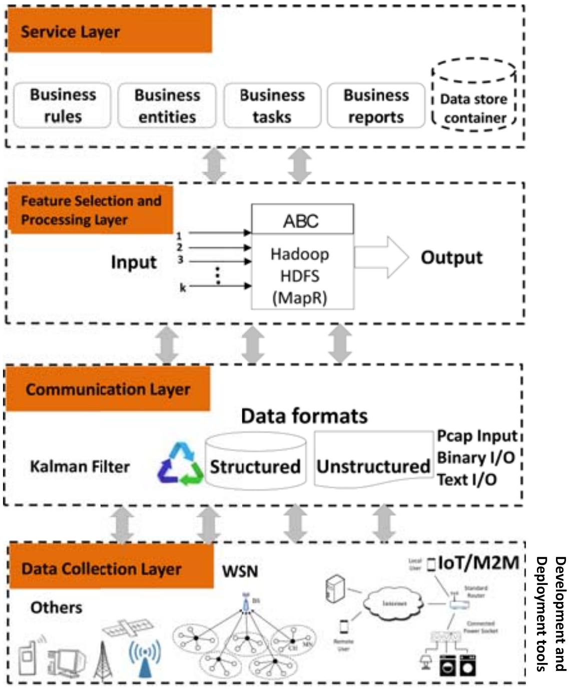

Fig. 2. Four-tier communication model

A. HABC: Hadoop-based ABC algorithm for SIoT

To elaborate the architecture of the proposed system architecture, a service scenario is shown in Figure 3. The proposed encompasses SIoT system development, i.e., smart traffic control department, smart weather forecast department and smart hospital and health department. Components above are liable for the collection of heterogeneous data within the SIoT network. Thus, acting as the bottom level of the proposed framework. These components are further connected with the smart decision and control system via heterogeneous access technologies such as GSM, Wi-Fi, 3G, and 4G. The autonomous decision making uplifts the reliability as well as the practicability of the proposed scheme. Upon receiving the collected data, intelligent decisions are carried out by the smart decision and control system, situated in the middle level of the smart city framework. Moreover, the middle level regulates the events conforming to the made decisions. The event generation is taken place at the top level (application level), upon the reception of autonomous decisions.

A realistic SIoT environment not only includes a prodigious amount of data but also complex and comprehensive computation and multiple application domains. The realization of the SIoT system implementation relies on all forms of data and computation due to their indispensability [46]. The smart environment notion aims to optimize residential resources, to reduce traffic congestion, to provide efficient healthcare services, and to perform the water management. The acquisition of data associated with the daily operational activities become vital regarding achieving the preceding aims. However, the data acquisition has become tedious and challenging due to the massive amount of data created by people and other connected devices. To process further, the phenomenon of interest from the real world are sensed and identified. Consequently, converted into digital data employs various mechanisms. Low cost and energy efficient sensors have become a promising mechanism to acquire heterogeneous data from the urban SIoT. The city becomes smarter, along with the expansion of the number of connected devices [47]. Hence, the realization of the proposed smart city architecture begins with the extensive deployment of heterogeneous sensors within the city suburbs. These sensors are liable for the collection of real-time data from the neighboring environment. The deployed context determines the type of collected data i.e. smart home, vehicular transportation system, healthcare management system, and meteorology system.

The bottom layer of the proposed scheme consists of multiple components. The key concern of the smart home is to enhance the energy utilization of the residential buildings. The home appliances are equipped with a sensor, which determines the real–time energy consumption and convey to the middle layer afterward. The data processing layer defines a threshold value for particular household's energy consumption. A data filtration process is performed by the data aggregation techniques to determine the values exceeding the threshold, thus optimizes further processing. Consequently, the decisions made at the middle level proceed to the smart community development in application level, which notifies energy consumption of a particular household to the respective residents. Meanwhile, it empowers the energy usage customization of residential buildings. The prime objective of the vehicular transportation system is to reduce the city traffic congestion. The data processing level defines the mean time that is taken to travel between two stated points. The sensors implanted on the roadsides collect vehicle entrance and departure between two points. The embedded aggregation techniques determine the roads with congestion by analyzing the current travel time of stated locations, which exceeds the defined mean time. Thence, it autonomously generates alternative paths and notifies the travelers via the application level. The utmost goal of the meteorology department is to ascertain the weather conditions and other environmental parameters. For example, the sensors implanted in certain locations determine the carbon monoxide (CO) concentration of the city. These sensors convey the acquired data to the middle level to filter and process accordingly to facilitate decision making and event generation.

The proposed architecture occupies multiple communication technologies; ZigBee, Bluetooth, Wi-Fi, and data and cellular networks to transmit sensed data to the data management and processing level.

The data management and processing level act as the mediator between the data acquisition and application levels. Since, the crucial processes such as filtering valuable data, analyzing, processing, storing, decision-making and generating events are carried out in this layer. So, this layer is Considered as the brain of proposed framework. To perform the tasks above, multiple modalities are embedded into this layer. Initially, the enormous amount of sensed data is filtered by aggregation mechanisms to obtain valuable real-time and offline data. The MapReduce paradigm is used for the data analysis, while manipulation and storing is performed by Hadoop distributed file system (HDFS), HBASE, and HIVE.

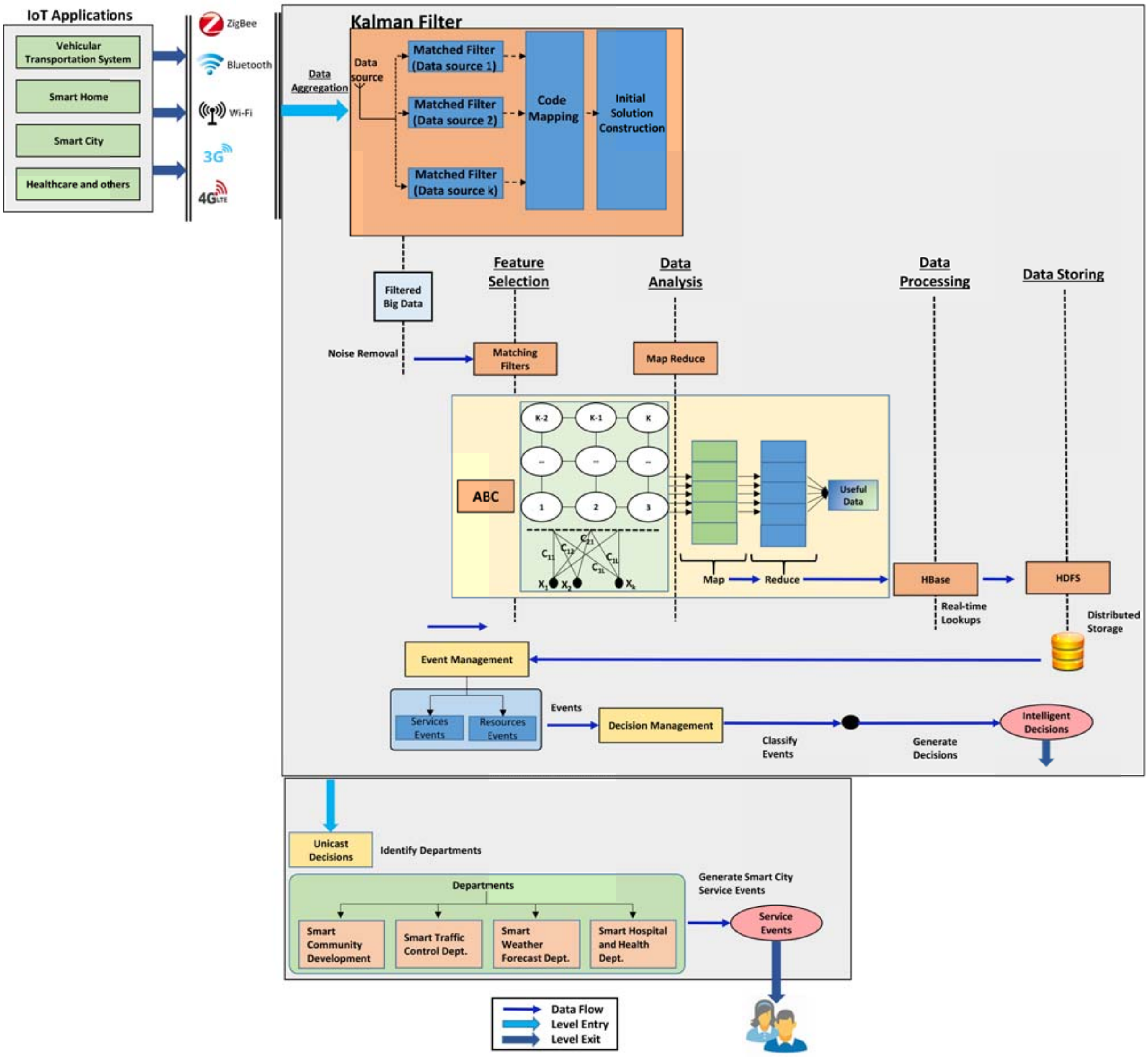

Fig. 3. Proposed HABC system architecture for Big Data in SIoT

The aggregation techniques enhance the data processing efficiency by applying data filtration. Kalman Filter (KF) is used to perform data filtration in the proposed framework [48]. The KF is an optimal estimator, which removes noise from the sensed data [49-50]. Algorithm I shows the working mechanism of KF in different steps for sensor data filtration. It initially assumes the current state f_k is evolved from the previous statef_{k-1}. The current state observation is denoted by h_k. $\hat{f}_{k|k-1}$ represents the estimation of f at time k, while the estimation accuracy is denoted by $G_{k|k-1}$. It deduce valuable data from a large set of indirect and uncertain data. Since the KF works recursively, it processes data on arrival. Thus, it assures the real-time operation of the smart city. Moreover, it facilitates immediate processing with a minimal memory consumption. As KF removes noise from data, the data processing level utilizes its capability to infer the best estimate from a larger set of real-time data. Thereupon, the KF is manipulated to determine valuable data corresponding to the predefined threshold values. For example, the roadside sensors of the streets and roads generate a massive amount of city traffic data. Nevertheless, further processing of uncongested street data is a superfluous task. Thence, the KF determines best fitting sensed data in accordance with the predefined thresholds. Ultimately, it reduces the amount of futile data resulting a swift analysis.

Algorithm I: KF Working

```
1. Initialization
   $T_k$ – State transition model (applied to the previous state $f_{k-1}$ )
   $O_k$ – Observation model
   $Q_k$ – Covariance of the process noise
   $R_k$ – Covariance of the observation noise
   $C_k$ – Control input model (applied to the control vector $v_k$ )
   $w_k$ ~ □ (0, $Q_k$)
2. Computing the new state $f_k$ using the previous state $f_{k-1}$
   $f_k = T_k f_{k-1} + C_k v_k + w_k$
   $h_k = O_k f_k + u_k$        $u_k$ ~ □ (0, $R_k$)
3. Current state estimation from the previous state
   Predicted state
       $\hat{f}_{k|k-1} = T_k \hat{f}_{k-1|k-1} + C_k v_k$
   Predicted covariance
       $G_{k|k-1} = T_k G_{k-1|k-1} T_k^T + Q_k$
4. Combine current prediction with the current observation
   Current observation
       $\tilde{x}_k = h_k - O_k \hat{f}_{k|k-1}$
   Observation covariance
       $S_k = O_k G_{k|k-1} O_k^T + R_k$
   Optimal gain
       $K_k = G_{k|k-1} O_k^T S_k^{-1}$
```

The proposed scheme stores and processes data in Hadoop framework. Thus, MapReduce has been selected as the mechanism for analyzing filtered data. MapReduce works in two steps. First is the mapping process where the set of filtered data is converted into another set of data. Next is the reduce process which combines the data created in mapping process and results in a set of values that are reduced in amount. Data storing and processing plays a major role in the realization of a smart city. As shown in Figure 3, the proposed framework utilizes multiple techniques i.e. HDFS, HBASE, HIVE, etc. to facilitate the above requirements. The storage demand of the proposed smart city is facilitated by HDFS, which is the primary storage of Hadoop. Since the storage of HDFS is distributed; it augments the MapReduce execution on smaller subsets of larger data cluster. Also, HDFS assists the scalability demand of the Big Data processing. To favor the autonomous decision making, the real-time read/write facility over the complete cluster is essential. Hence, HBASE is used to enhance the processing speed on Hadoop as it offers real-time lookups, in-memory caching, and server side programming. Further, it enhances the usability and the fault tolerance. HIVE provides querying and managing facility over the large amount of data that resides on the Hadoop cluster. Since SQL cannot be used to query on HIVE, we have used HiveQL to query the data on Hadoop cluster. Finally, the derived intelligent decisions are transferred to the application level of the framework.

1) HABC Algorithm

The proposed HABC algorithm is used for feature selection in large data sets, which are generated by IoT network. Abc algorithm consists of three phases, i.e., food source, employed bees, and unemployed bees [51]. These are described as follows.

- Food source represents the solution of the given problem.
- Employee bees are used to find out the different food sources. Also, they are used to store the quality of the information and share this information with other bees in the honeycomb.
- Finally, the unemployed bees are classified into two types, i.e., onlooker and scout bees.
  - Onlooker bees receive shared information from employed bees, which are used to find the better quality food sources.
  - Scout bees are those when employed bees get exhausted to find sources; they turned into scout bees. They are the one who tries to find new sources of food.

The pseudocode of the ABC is given in Algorithm II [52].

Algorithm II. ABC Algorithm

```
1. Initialization Phase
2. Repeat
     i. Employee bees
        $a_{ij} = a_j^{min}$ + rand ($a_j^{max}$ - $a_j^{min}$)        (1)
        i = 1, 2, 3, . . .N, j = 1, 2, 3, . . . K, where N and D are food sources and optimization parameter
    ii. Unemployed bees
        $v_{ij} = x_{ij} = \Phi_{ij}$ ($x_{ij}$- $x_{kj}$)        (2)
        Food source $v_i$, parameter j, where j and k are random variable.
        $\text{fitness}_i = \begin{bmatrix} \frac{1}{1+f_i} & if\ f_i \geq 0 \\ 1 + abs(f_i) & if\ f_i < 0 \end{bmatrix}$        (3)
        $p_i = \frac{\text{fitness}_i}{\sum_{n=1}^{F} \text{fitness}_i}$        (4)
          a. Onlooker bees
          b. Scout Bee
3. Memorize the best probable results
4. Until (cycle = maximum cycle per number or a maximum CPU time)
```

One the opposite of the optimization algorithms, where solutions to various problems are expressed by vectors along with the real values, the contestant solution for feature selection constraints are represented by bit vectors. In a given scenario of HABC, each food source is linked with a bit vector (size N, where N is an overall number of features). The location in the vector agrees with the overall number of features, which need to be evaluated. In this case, if the value of the agreed feature is equal to 1, then this condition shows that feature is a part of the evaluated subset. However, if the value agreed feature is equal to 0, then this condition shows that feature is not a part of evaluated subset. Moreover, food source store its quality (i.e., fitness). This is given by the precision of the classifier using the feature subset designated by a bit vector.

The steps of the proposed HABC for feature selection are given below.

1. In Hadoop processing system, when the noise is removed from data sets using Kalman filter, it is required to find the best and lowest number of the features. For this reason, the proposed system exploits forward search strategy [7]. In this technique, there are total N number of food sources that contains N number of features. This follows the same technique of assigning a bit vector of size N (each subset shall contain the only 1feature).
2. Afterward, feature subset of each food source is assigned to classifiers, where it uses accuracy as a fitness value (accuracy is stored in a fitness of food source).
3. Since the classifiers are now assigned, now it is required to determine neighbors for chosen food sources. It is achieved by exploiting parameter of modification rate (MR), in which employee bee visit each food sources and explore their neighbors. To extract features, neighbors are created from a bit vector of the initial food source. In HABC, the neighborhood is performed by using equation 2 using MR [7]. A random and uniform number R_i (range: 0 and 1) is generated for each position of the bit vector. In case, if the value is less than perturbation parameter MR, then the feature is injected into subset. Apparently, the vector is not modified by using equation (5). Thus, submitting a subset's feature to the classifier, and use the accuracy fitness.
$$a_i = \begin{bmatrix} 1 & if\ R_i < MR \\ a_i & otherwise \end{bmatrix} \tag{5}$$
4. If the quality of the newly found food source is better than that of the exploration food sources, then the neighboring food source is considered as latest one. This information is shared with the other bees. In HABC, the size of the data is exponentially increases, so this process is continued until the selection of the best feature in Hadoop.
5. The onlooker bees are used to collect information about the fitness/correctness of the food sources, which are visited by employee bees. Their job is to select the food source based on two parameters, i.e., better probability exploration or better fitness/correctness. Thus, it is required to store the best fitness values.
6. Finally, a scout bee is created for an unrestrained food source, where a new food source is created, in which N number of features are randomly created and submitted to the classifier. Thus, the newly found source is assigned to scout bees, and the employee bees performed their job again.

IV. Experimental Results

The proposed architecture is tested against the well-known swarm optimization approaches such as Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), and Genetic Algorithms (GA) in the context of accuracy, the number of features, etc. Moreover, the classification of data and the parameters used by the ABC mechanism is explained in the upcoming sections. Furthermore, the classification performance is analyzed based on the standard accuracy metric. The features obtained by each approach are tested on the same dataset for repeating the experiments for ten times. Finally, the results obtained are taken with the average of all the ten experiments.

A. Datasets

The proposed optimization and feature selection are tested on ten datasets from the UCI machine learning Repository [53]. Each dataset is tested for each learning algorithm using multi-cluster Hadoop system. The description of each dataset used in the analysis is shown in Table 1. With the number of attributes (features), instances and classes. Each dataset is analyzed mainly based on the number of features. Because we have two prime interests i.e. 1) to compute the features and apply the decision mechanism on various thresholds and 2) To test the data and extract the features of interest. Moreover, the dataset is obtained from the UCT machine learning repository which is a high authentic and rich repository of various sources [53].They are maintaining the repository with great care, and up to now, they have 351 datasets from various sources and fields. However, there is still need of datasets from IoT-based environments such as smart homes, cities, etc. in various formats such as ASCII, text, etc. Hadoop analyzed the data in text and number format and therefore, it is essential to present the data in that format, so the researchers can easily test the data using Hadoop.

TABLE 1. Data sets used in the analysis

Data Set	Number of Features	Number of Instances
Gas sensors for home activity monitoring	11	919438
Air Quality	15	9358
GPS Trajectories	15	163
Indoor User Movement Prediction from RSS	4	13197
3D Road Network	4	434874
Water Treatment Plant	38	527
Hepatitis	19	155
Housing	14	506
Cloud	10	1024
Twitter Data set for Arabic Sentiment Analysis	2	2000

TABLE 2. The accuracy of proposed system on UCI datasets

Data Set	Total Number of Features	Selected Number of Features	Total Features (accuracy %)	Average Accuracy (%)

Gas sensors for home activity monitoring	11	6	71.28	92.49
Air Quality	15	7	73.28	84.56
GPS Trajectories	15	5	65.24	75.25
Indoor User Movement Prediction from RSS	4	2	75.56	86.32
3D Road Network	4	1	74.56	82.45
Water Treatment Plant	38	13	81.23	94.31
Hepatitis	19	12	66.43	91.36
Housing	14	4	54.85	82.14
Cloud	10	6	71.84	77.86
Twitter Data set for Arabic Sentiment Analysis	2	1	84.56	96.63

B. Simulation Environment

All the experiments are performed on a multi-cluster Hadoop installed on Ubuntu 14.04 LTS with a Core i5 3.4 GHz processor and 8GB RAM. The proposed feature selection algorithm is implemented in Java programming language. Moreover, the data classification is performed in Java along with WEKA [54] and LibSVM [55] libraries.

C. Classification Performance

The performance evaluation and accuracy of the proposed feature selection based on ABC mechanism is obtained using a 10- fold cross validation with partitioning mechanism of K different partitions. One of the set is used as a primary partition, and the rest K-1 is used as training set. However, this process is repeated ten times and each time one of the set becomes the primary and rest of them becomes the training set. The results are accumulated after each test, and the final results are obtained by taking the mean of all the ten partitions. Moreover, the features established in some test is normalized using Z-score mechanism [55], which is subtracting the average value for each feature set and divide it by the standard deviation of the set. The rest of the parameters in each algorithm is shown in Table 2. All these parameters are set to some standard values and therefore, it does not affect the performance in each case.

D. Discussion on Results

UCI dataset provides a variety of features and classes. These various features and classes widely influence the performance and accuracy of a feature selection algorithm. The performance analysis in the context of accuracy with a selected number of feature and a full set of the feature is shown in Table 2. As shown in Table 3, the selected feature are quite less than the other original feature list. However, the accuracy of the system is superior to the original data feature for all data. In comparison with other data set, the proposed feature selection perform superior in the context of accuracy. In some datasets such as and Air Quality, 3D network Traffic, and cloud the accuracy is worst while in some such as GPS trajectories and Twitter Dataset for Arabic Sentiment Analysis, the accuracy is almost equal to the other methods. However, the accuracy of the proposed scheme is considerably good in all the cases.

TABLE 3. *Comparative analysis of proposed scheme with other Swarm Approaches.*

Data Set	PSO (%)	ACO (%)	GA (%)	ABC (%)
Gas sensors for home activity monitoring	85.90	78.56	78.26	92.49
Air Quality	91.70	93.25	89.28	84.56
GPS Trajectories	74.23	68.47	65.24	75.25
Indoor User Movement Prediction from RSS	75.98	77.58	75.56	86.32
3D Road Network	88.85	91.24	87.56	82.45
Water Treatment Plant	88.63	86.89	81.23	94.31
Hepatitis	87.65	88.59	82.46	91.36
Housing	67.56	74.45	75.85	82.14
Cloud	83.52	86.29	71.84	77.86
Twitter Dataset for Arabic Sentiment Analysis	89.58	87.24	88.34	90.63

To be more realistic, the proposed system based on ABC optimization algorithm is compared with a single node Hadoop and Java query based system as shown in Figure 4. In the single node Hadoop each time the data is processed without any optimization system. On the other hand, Java query based system is tested with all the other swarm approaches. Similarly, the filtration system is used to remove the noise from the data before passing it to the Hadoop ecosystem. We gradually increase the data size with time and observe the effect of it on the proposed system, a Hadoop-based system without other swarm approaches, and Java query system with other swarm approaches. The implementation of such systems helps in processing Big Data in real-time. This further enhances the decision system with real-time events generations. The proposed system efficiently processes the data in real time and generates results which help the citizen in making a decision in real time. For example, on processing the environmental data in real-time help the citizens to avoid going to those places which are highly polluted. Initially, the single Hadoop node and Java query based system process the data with similar speed as shown in Figure 5. But, with the increase in data set size, the speed of processing is highly decreased. However, the proposed scheme efficiency is significantly high compared to the single node Hadoop and Java query based system.

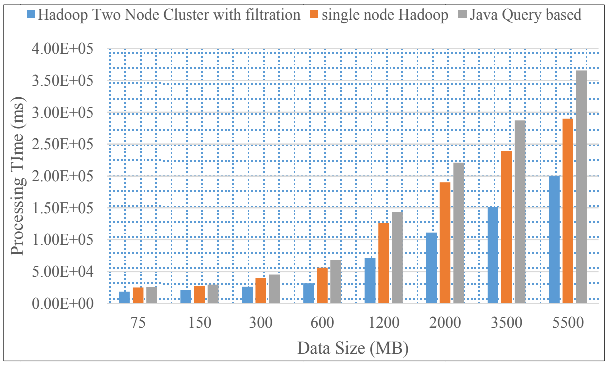

Fig. 4. Processing Time Analysis

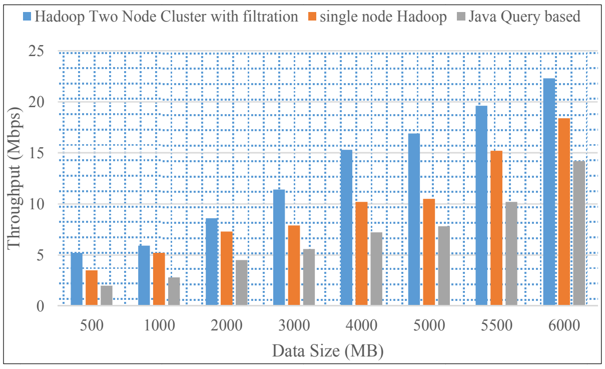

Fig. 5. Efficiency of the system in context of throughput

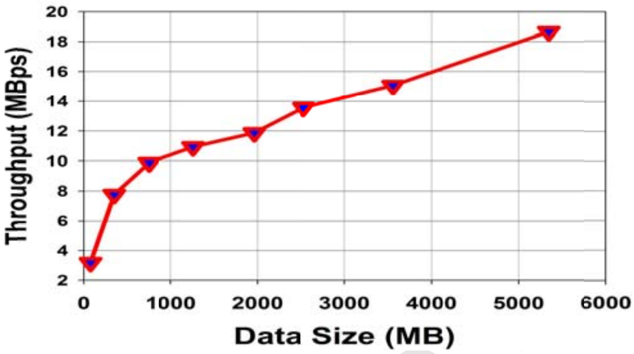

Fig. 6. Throughput of the proposed system

Moreover, we also evaluate the throughput of the proposed system by increasing the size of the data sets. As shown in Figure 6, the throughput is directly proportional to the size of the data sets. When there is an increase in the size of the data sets, the throughput also increases, hence increase the system sufficiency.

To test and validate with various other datasets, we measured the processing time as good throughput on healthcare datasets as shown in Figure 7 and 8. In the figure, the proposed scheme takes few seconds to process GBs of data. To be more specific, it takes seventy seconds to process 2GB data on a single node of the parallel processor. Moreover, if we increase the size of the datasets that due to the nature of the proposed system, the throughput is maximized. Therefore, it is concluded from these results that the proposed system with parallel processors of the system gives us very efficient results than ordinary simple processing tools.

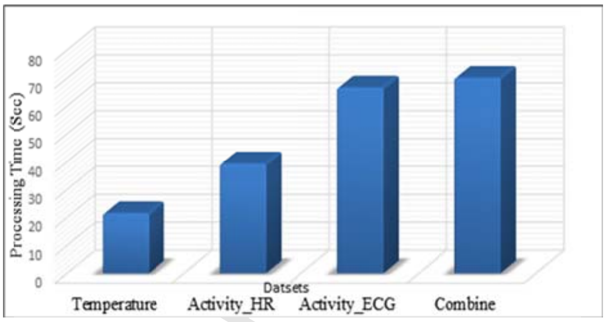

Fig. 7. Processing time of healthcare datasets

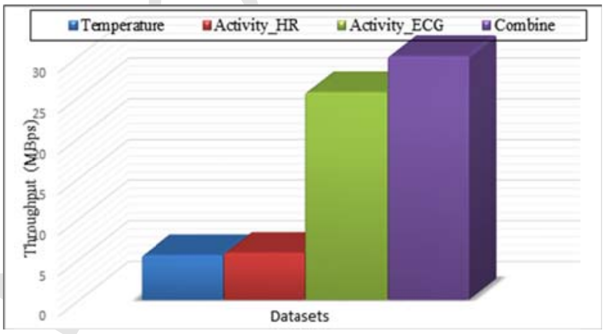

Fig. 8. Throughput of our proposed scheme

V. Conclusion

Undeniably, scientific discoveries and latest innovations can benefit significantly from huge volume aggregated data and simulated data. Moreover, data scientists can gain insights and apprehend the singularities behind the data more efficiently. However, it is based on the assumptions that the designed feature selection algorithms can deliver effective and efficient results. In prior research, various work has been done on the feature selection to make IoT more effective, which give a significant amount of results. However, still, the prior systems are lacking capabilities. Therefore, this paper describes the system architecture for feature selection in Big Data IoT. The proposed scheme is based on the four layers architectural model then aggregated the data, remove erroneous or redundant data, and select features efficiently, which is useful for the enhancement of computational capabilities using Hadoop server giving high-performance computing. The whole system is implemented using enhanced MapReduce with the features of ABC algorithm to select features and process large data sets MapReduce to process other data with Hadoop ecosystem to achieve the efficiency and real-time processing. The results proved that the use of ABC in Hadoop ecosystem dramatically increase the efficiency of the whole system in selecting features.

References

[1] Ahmad, Awais, Anand Paul, and M. Mazhar Rathore. "An efficient divide-and-conquer approach for big data analytics in machine-to-machine communication." *Neurocomputing* 174 (2016): 439-453.

selection." *Signal processing* 93, no. 6 (2013): 1566-1576.

[53] Machine Learning Repository," University of California, 1987. [Online]. Available: https://archive.ics.uci.edu/ml/index.html. [Accessed 12 4 2016].

[54] "ibsvm," [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/. [Accessed 17 3 2016].

[55] S. Aksoy and R. Haralick, "Feature normalization and likelihood-based similarity measures for image retrieval," *Pattern Recognition Letters,* vol. 5, no. 22, pp. 563-582, 2001.

Awais Ahmad is currently working in the Department of Information and Communication Engineering, Yeungnam University, Korea. He received his BS (CS) from the University of Peshawar and MS (Telecommunication and Networking) from Bahria University, Islamabad Pakistan in 2008 and 2010 respectively. During his Master's research work he worked on energy efficient congestion control schemes in Mobile Wireless Sensor Networks (WSN). During his Ph.D., he was working on Big Data in IoT, SIoT, M2M, and WSN. He has research experience on Big Data Analytics, Machine-to-Machine Communication, and Wireless Sensor Network. Dr. Awais has published more than 50 papers in various IEEE, Elsevier, and Springer journals, and also in reputed conferences, i.e., IEEE Globecom, IEEE LCN, and IEEE ICC in the field of Big Data, IoT/SIoT, and WSN. He is a visiting researcher at CCMP Labs, Kyungpook National University Korea. Also, he was a visiting researcher in INTEL Labs, National Taiwan University, Taipei Taiwan. He serve as a reviewer in many journals/conferences, and TPC member in various conferences. He also received three prestigious awards: (i) Research Award from President of Bahria University Islamabad, Pakistan in 2011, (ii) Best paper nomination award in WCECS 2011 at UCLA, USA, and (iii) Best paper award in the 1st Symposium on CS&E, Moju Resort, Korea in 2013. He is a member of the Institute of Electrical and Electronics Engineers (IEEE).

Murad Khan received his Ph.D. degree from Kyungpook National University, Korea. He is currently working an Assistant Professor in the Department of Computer Science, Sarhad University of Science and Inofrmation Technology, Peshawar Pakistan. During his Ph.D., he was working on handover mobility model in heterogeneous network. Later, he started working on IoT and Big Data. He has published more than 40 articles in various journals and conferences. He was also a Best Research at School of Computer Science and Engineering, Kyungpook National University.

Anand Paul is currently working in The School of Computer Science and Engineering, Kyungpook National University, South Korea as Associate Professor; He got his the Ph.D. degree in the electrical engineering at National Cheng Kung University, Taiwan, R.O.C. in 2010. His research interests include Algorithm and Architecture Reconfigurable Embedded Computing. He is a delegate representing South Korea for M2M focus group and for MPEG. 2004-2010 he has been awarded Outstanding International Student Scholarship, and in 2009 he won the best paper award in national computer symposium, Taipei, Taiwan. He serves as a reviewer for IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transaction on System, Man and Cybernatics, IEEE Sensors, ACM Transactions on Embedded Computing Systems, IET Image Processing, IET Signal Processing and IET Circuits and Systems He gave invited talk in International Symposium on Embedded Technology workshop in 2012, He will be the track chair for Smart human computer interaction in ACM SAC 2014. He is also an MPEG Delegate representing South Korea.

Sadia Din received his Bachelors in Computer Engineering from Comsats Institute of Information Technology Abbottabad, Pakistan. Currently, she is pursuing her Masters Leading Ph.D. degree at Kyungpook National University, Daegu, South Korea. Her research interests include IoT, Big Data analytics, Wireless Sensor Network, and 5G/4G.

Muhammad Mazhar Ullah Rathore received the Master's degree in computer and communication security from the National University of Sciences and Technology, Islamabad, Pakistan, in 2012, and is currently pursuing the Ph.D. degree at Kyungpook National University, Daegu, South Korea. His research interests include IoT, BigData analytics, network traffic analysis and monitoring, intrusion detection, and computer and network security.

Gwanggil Jeon received the B.S., M.S., and Ph.D. (summa cum laude) degrees from the Department of Electronics and Computer Engineering, Hanyang University, Seoul, Korea, in 2003, 2005, and 2008, respectively. He was with the Department of Electronics and Computer Engineering, Hanyang University, from 2008 to 2009. He was with the School of Information Technology and Engineering, University of Ottawa, Ottawa, ON, Canada, as a Post-Doctoral Fellow, from 2009 to 2011. He was with the Graduate School of Science and Technology, Niigata University, Niigata, Japan, as an Assistant

Professor, from 2011 to 2012. He is currently an Associate Professor with the Department of Embedded Systems Engineering, Incheon National University, Incheon, Korea. His current research interests include image processing, particularly image compression, motion estimation, demosaicking, and image enhancement, and computational intelligence, such as fuzzy and rough sets theories. Dr. Jeon was a recipient of the IEEE Chester Sall Award in 2007 and the ETRI Journal Paper Award in 2008.

Gyu Sang Choi received the Ph.D. degree in computer science and engineering from Pennsylvania State University. He was a research staff member at the Samsung Advanced Institute of Technology (SAIT) in Samsung Electronics from 2006 to 2009. Since 2009, he has been with Yeungnam University, where he is currently an assistant professor. His research interests include embedded systems, storage systems, parallel and distributed computing, supercomputing, cluster-based Web servers, and data centers. He is now working on embedded systems and storage systems, while his prior research has been mainly focused on improving the performance of clusters. He is a member of the IEEE and ACM

- Presents an architecture for Social Internet of Things
- Modeling and optimization of big data
- System Architecture based on ABC
- MapReduce with ABC to enhance system efficiency