Wireless Personal Communications (2020) 114:409–430
https://doi.org/10.1007/s11277-020-07369-0

Intelligent Data Fusion for Smart IoT Environment: A Survey

Ihsan Ullah[1] · **Hee Yong Youn**[2]

Published online: 18 April 2020

Abstract
Efficient data collection and communication are key tasks in smart IoT environment consisting of a large number of devices. Here imprecise data are generated due to the interferences between the devices and harsh operation condition, and therefore data fusion is needed to gather and extract useful data from multiple sources. A number of approaches for data fusion have been proposed which are based on probability, artificial intelligence, or evidence theory to efficiently aggregate the data. The techniques allow the system to be cognitive and intelligent in terms of decision-making under the uncertainty of data and limited resource. In this paper a comprehensive survey on the data fusion techniques for smart IoT system is presented. The challenges and opportunities with data fusion are also delineated. It will be useful for the researchers in developing the applications and services based on smart IoT environment, which require intelligent decision making.

Keywords Machine learning · Data fusion · Clustering · Smart IoT environment · Evidence theory · Decision making · Wireless sensor network

1 Introduction

The smart environment of IoT in modern real-world consists of tiny devices equipped with sensors, actuators, and computational elements. These devices are connected through mostly wireless network for collecting data from the environment and inferring the status based on them [1]. The smart environment usually consists of heterogeneous devices providing diverse services as shown in Fig. 1 [2]. The heterogeneous devices may generate imprecise or noisy data deteriorating the inference accuracy, while the events and data produced in the smart environment are related with each other. Here it is necessary to implement a method of sophisticated data integration dealing with various sources. Note that it is challenging to efficiently fuse a large amount of probably noisy data and then infer an accurate result. Moreover, it is required to process the data

✉ Hee Yong Youn
youn7147@skku.edu

Ihsan Ullah
ihsan@skku.edu

[1] Electrical and Computer Engineering, Sungkyunkwan University, Suwon, Korea

[2] College of Software, Sungkyunkwan University, Suwon, Korea

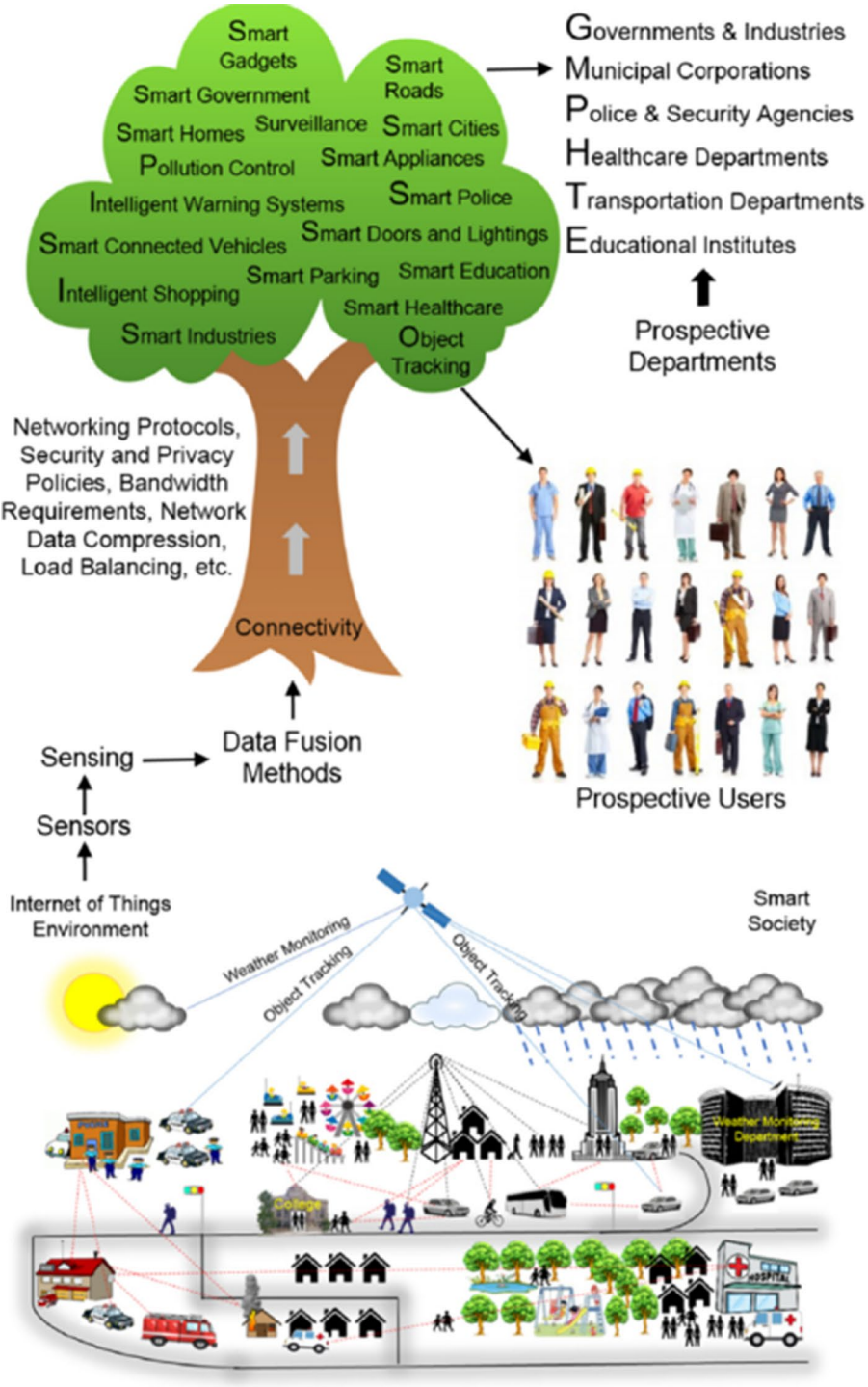


Fig. 1 The smart IoT environment requiring data fusion [2]

based on different contexts and inference condition. The smart environment needs the context-aware operation to achieve high performance with minimal energy consumption and networking overhead.

Wireless sensor network (WSN) is commonly used to monitor and gather the required data from the target area. It consists of a number of sensor nodes which are distributed in high density to reliably cover the target area [3, 4]. The sensor nodes are limited with respect to the communication and computation power, and therefore collaboration between them is required to collect and transmit data to the base station (BS). The dense distribution in the target area, however, causes the data redundancy problem due to spatial and temporal correlation of the nodes. Outlier in the sensory data is another problem which is aggravated by the instability of the communication environment. It reduces the integrity of the data and performance of the entire system. Considering such unstable and erroneous characteristics of WSN, machine learning technique is expected to be effective in exploiting the collected data and improving the performance of the system.

The decision-making system is the core component of smart IoT environment, and its accuracy relies on the integrity of the data obtained with the sensor nodes. The sensor data might be corrupt due to sensory deprivation, restricted coverage, imprecision, and uncertainty, which significantly degrade the quality of decision. Also, the spatial and temporal redundancy of the data decrease the performance of WSN [3], and transmission of redundant data consumes large energy which eventually shortens the lifetime of the entire network. By minimizing redundant data, a significant amount of resources can be saved and the network performance can be enhanced. In addition, the uncertainty and inconsistency in sensor data may result in wrong inference on the environment. Therefore, enhancing the data integrity is the key to increasing the accuracy of decision-making. Data fusion is a discipline concerning with how multi-source data are merged to increase the integrity of data. It allows to effectively deal with noisy data of dynamic environment, and helps the decision-making process based on the available information [4]. Fusion of sensor data is one of the crucial tasks with WSN, and numerous data fusion mechanisms have been proposed to filter and merge the sensor data before sending to the sink and decision-making system [5–7]. Figure 2 shows the structure of decision making system consisting of data filtering, fusion, and processing.

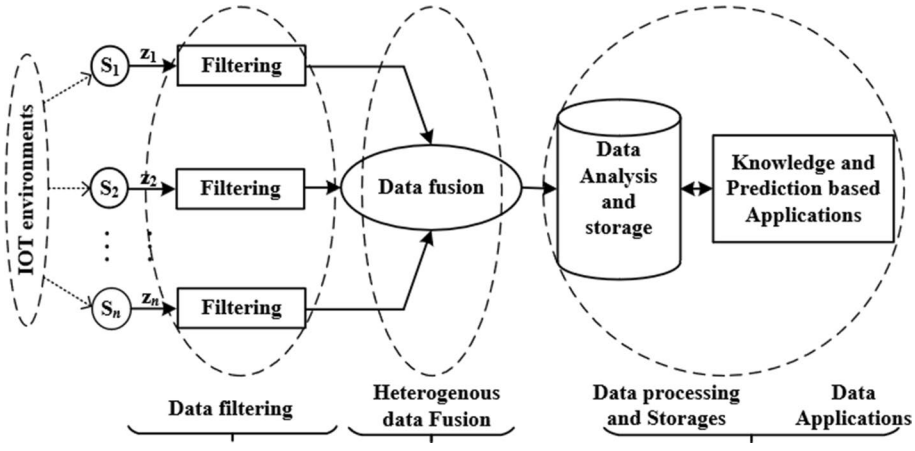


Fig. 2 The structure of decision-making system consisting of data filtering, fusion, and processing

Intelligent data fusion is important to improve the accuracy of decision making process for the following reasons:

- The IoT system usually operates in dynamic real-time environment, and thus it is necessary to establish a smart network which can efficiently adjust its operation according to the operational condition.
- WSN is often used for gathering data from unreachable, dangerous, or critical location [8] such as fire or water leakage detection. The system designers need to utilize a robust technique that is able to make correct and reliable decision based on available knowledge, and also gain new knowledge from the experience.
- WSN is usually deployed in complicated environment, and thus it is quite hard to build an accurate mathematical model on the target operation, e.g. event or outlier detection. Data fusion based on various techniques including machine learning is imperative to efficiently handle such complicated problem and situation.
- In the machine-to-machine (M2M) communication of IoT environment, smart decision-making and control are required [9]. With artificial intelligence techniques [10], different levels of knowledge can be used to make a decision and the tasks are dynamically performed based on the contextual information.
- It is not easy to extract important correlation between the data and accurately fuse them if the amount of data is large. Here machine learning techniques are expected to be effective.

Data fusion is applied to combine the data of multiple sources in effective and accurate way. In WSN environment it used to integrate the multi-sensor data and transmit them to the BS [11]. Due to spatial and temporal correlation of adjacent sensor nodes, a significant amount of redundant data are generated which need to be reduced. The outliers in the data are caused due to unexpected events or malicious attacks on the network, while the noises and errors reduce the integrity of the data [12]. Without cleansing or filtering redundant or erroneous data, the fused data might not be useful. Data fusion is classified mainly into two approaches based on the employed network structure, centralized approach and cluster-based approach as shown in Fig. 3.

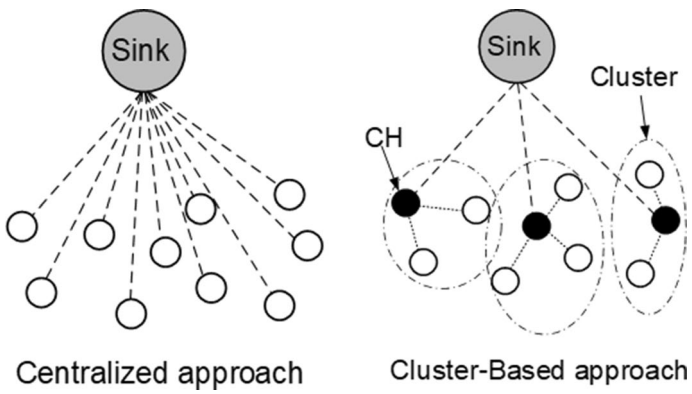


Fig. 3 The two approaches employed for data fusion

The centralized approach is used to filter and fuse the data at sink level so that the end-to-end data transmission delay can be reduced since the data of the highest priority must be transmitted with a minimum transmission delay. But the centralized filtering of the sensed data may limit the inference accuracy, and increases the network load by sending noisy and redundant data to the sink. The cluster-based approach has been developed to reduce the temporal and spatial data redundancy and outliers in the collected data at the cluster head (CH). The objectives of the data fusion techniques are to collect data using minimal resources. Figure 4 describes the data fusion operation with clustered WSN using machine learning technique [13, 14]. Here the researchers attempted to efficiently filter and merge the sensor data at the CH before sending them to the sink.

This survey paper aims to investigate various data fusion schemes which are employed with WSN, and compare their features. The main contributions of the paper are as follows:

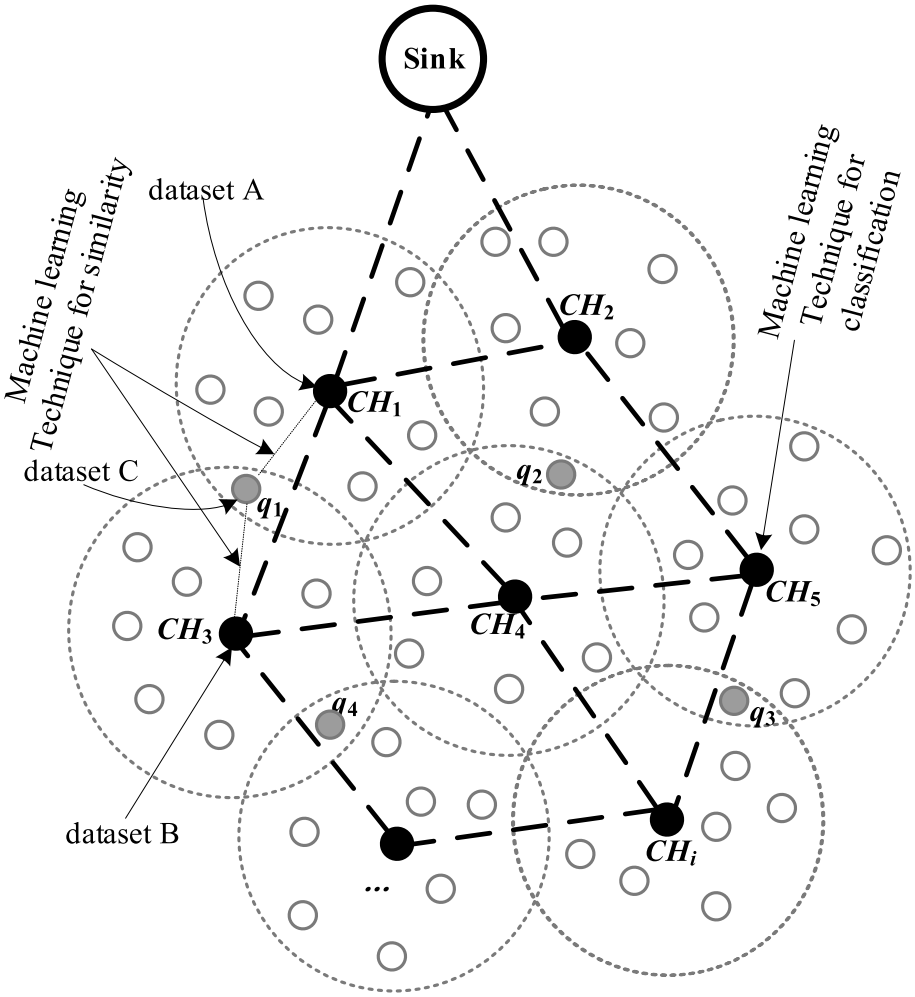


Fig. 4 The fusion of data with clustered WSN using machine learning technique

- The multi-sensor data fusion schemes based on various technique or theories for WSN and IoT environment are introduced which can help the researchers in developing a smart cognitive system.
- The challenges and opportunities for data fusion are explained considering the characteristics of sensor data including uncertainty, noise, inconsistency, redundancy, and outliers, etc.
- The mathematical models applicable to multi-sensor data fusion are discussed for the applications to WSN and IoT.

The rest of the paper is organized as follows: in Sect. 2, the data fusion techniques are discussed. The opportunities and challenges are explained in Sect. 3, and the conclusion is made in Sect. 4.

2 Approaches for Data Fusion

In smart environment the data from a single source may not be sufficient for making an accurate decision. Hence, multi-sensor data fusion and inference are required handling heterogeneous data. The data fusion techniques are classified into three categories with respect to the employed method [15]; the probability-based, AI-based, and evidence-based technique. They are summarized in Table 1.

- *Probability-based* Recursive operators and Bayesian analysis
- *Artificial Intelligence (AI)-based* Neural networks (NN) and Fuzzy Logic
- *Evidence theory-based* Dempster–Shafer theory

2.1 Probability-Based Method

In this subsection various probabilistic techniques proposed for data fusion with IoT are reviewed. Bayesian inference is one the most popular probabilistic methods developed for data fusion [41–46]. It needs relatively small number of sample data required to train the system, and allows dealing with the heterogeneity of information based on the probabilistic occurrence of the events in the environment. In [47] a data fusion scheme based on hard and soft sensor is proposed. It presented the cloud-enabled Bayes network for consolidating heterogeneous, real-time data streams from the target region to accomplish actionable intelligence from the computer-based decision supportive network. The data fusion information group (DFIG) model is shown in Fig. 5.

The DFIG model supports various control functions based on the spatial/temporal/spectral differences of the sensors. The levels of DFIG model are as follows [47]:

- Level 0 Data Assessment (DA):
- Level 1 Object Assessment (OA):
- Level 2 Situation Assessment (SA):
- Level 3 Impact Assessment (IA):
- Level 4 Process Refinement (PR):
- Level 5 User Refinement (UR):

Table 1 The approaches and nature of the techniques developed for data fusion

Method	Technique	Study
Probability-based	• Bayesian approaches • Recursive operators	• Bayesian data investigation [16] • Bayesian method to multisensor data fusion [17] • Dynamic Bayesian: inference and learning [18] • Dynamic and active data fusion with DBN [19] • Multi-sensor fusion through adaptive Bayesian network [20] • Bayesian activity recognition in residence [21] • Adaptive Bayesian system for sensor data fusion [22] • Bayesian method to covariance estimation and data fusion [23] • A recursive fusion filter for angular data [24] • Multi-scale recursive estimation, data fusion [25]
Artificial Intelligence-based	• Fuzzy Logic • Artificial Neural Networks (ANN)	• Data aggregation based on the event-driven and NN [26] • A fuzzy-logic-based data fusion WSN [27] • Fuzzy-based data fusion for fault detection in WSN [28] • Data aggregation scheme using NN in WSN [29] • Neural networks and statistical learning [30] • An RBFNN-based data aggregation algorithm for WSN [31] • Fuzzy systems and data mining [32] • A hierarchical fused fuzzy deep NN for data classification [33] • Multi-sensor data fusion in cluster-based WSN using fuzzy logic [32] • Fuzzy logic-based data fusion for autonomous multi-sensor [34]
Evidence theory-based	• Dempster–Shafer theory	• Bayesian and Dempster–Shafer (DS) fusion [35] • Application of DS theory for data fusion [36] • A distributed data fusion system based on DS theory [37] • Combination of evidence in DS theory [38] • Extended DS theory in context reasoning for smart environments [39, 40]

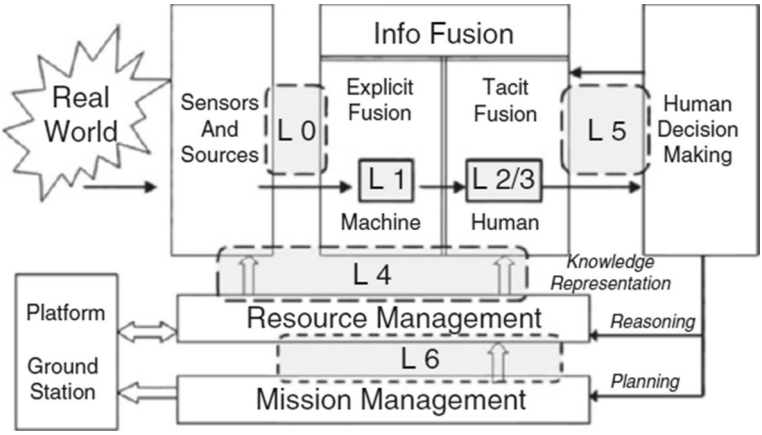


Fig. 5 The data fusion information group (DFIG) model [47]

The Dynamic Bayesian networks (DBNs)-based [20] adaptive data fusion scheme was proposed for various applications such as identification and detection of object. It considers the previous belief under the current observation of the phenomena to get subsequent estimation. Figure 6 is an example of the model.

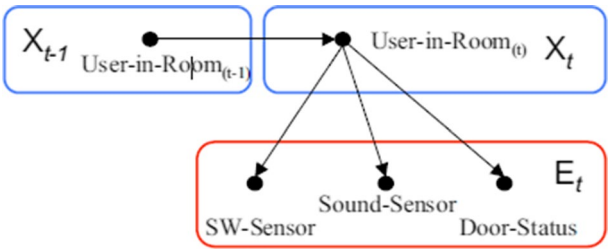
In [48] information aggregation and image data mining are achieved using the Bayesian technique. It applied Bayesian inference to acquire an estimate of a given physical parameter based on the perceptions obtained with various sensors. The paradigm of Bayesian information and knowledge fusion is shown in Fig. 7. Here the issue of fusion of two data sets requiring the combination of knowledge is observed, in a form of the determination of the priori models, M_1 and M_2, as in Eq. (1). A Bayesian methodology for data fusion can be formulated to maximize the posteriori probability [48]:

$$p\left(\Theta|D_1, D_2, |M_1, M_2\right) = \frac{p\left(D_1|\Theta, M_1\right)p\left(D_2|\Theta, M_2\right) \cdot P\left\{p\left(\Theta|M_1\right)p\left(\Theta|M_2\right)\right\}}{p\left(D_1, D_2|M_1, M_2\right)} \quad (1)$$

where P denotes the prior data in the hypothesis of two distinct models.

Nowadays the improvement in monitoring for animal health care is rapidly growing. In [49] the animal health monitoring scheme was proposed by using Bayesian algorithm for enhancing the productivity and monitoring the health of the animals. A data mining approach based on Bayesian Networks (BN) was introduced in [50], which integrates the quantitative and qualitative knowledge into a comprehensive probabilistic information

Fig. 6 An example of DBN [20]



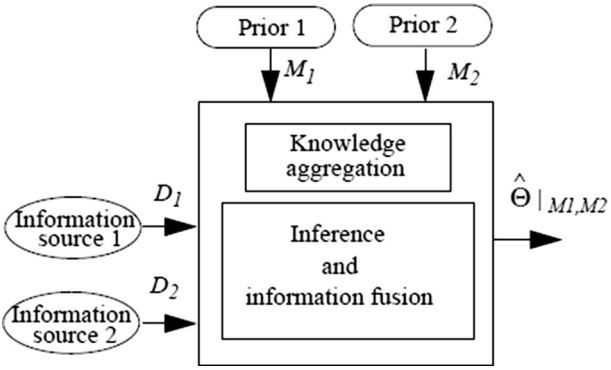


Fig. 7 The paradigm of Bayesian information and knowledge fusion [48]

prototypes and inference in WSN. Similarly, a Bayesian-based model [51] was proposed to fuse the measured temperature data from smart building. It extracts knowledge with a few sensor measurements, and then predicts the spatial temperature distribution for posterior estimation. In [17] three filtering approaches, Pre-Filtering, Post-Filtering, and Pre-Post-Filtering were proposed to fuse the sensor data. It proposes an approach for filtering and combining the sensor data using modified Bayesian fusion algorithm with Kalman filter to effectively handle the uncertainty and inconsistency problem.

2.2 Artificial Intelligence-Based Method

The artificial intelligence-based data aggregation and fusion techniques can effectively classify and abstract the information, and extract important features and knowledge from the data [52, 53]. The sink nodes can handle the fusion and classification of the data extracted from multiple sources using the back-propagation network (BPN) technology. Here the location and time limitation are considered to reduce the data gathering latency (Fig. 8).

With the fuzzy-based data fusion algorithm [27], an unfixed fusion weight is assigned to the CH. The weight is computed using fuzzy-logic dealing with various parameters such

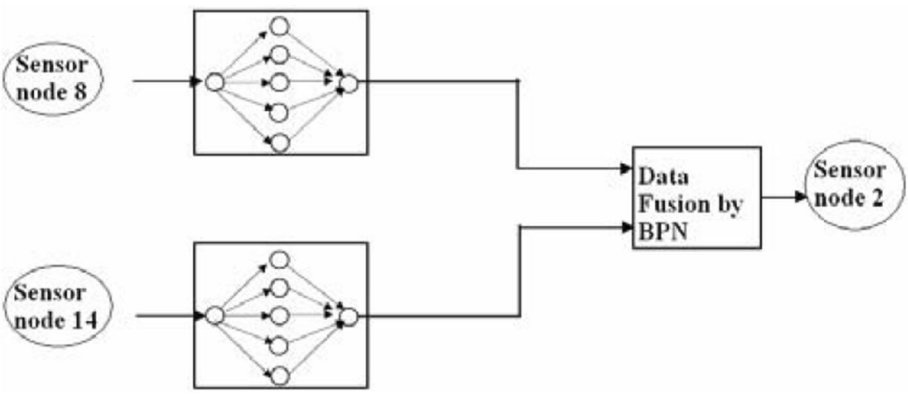


Fig. 8 The data fusion model for decision making based on BPN [53]

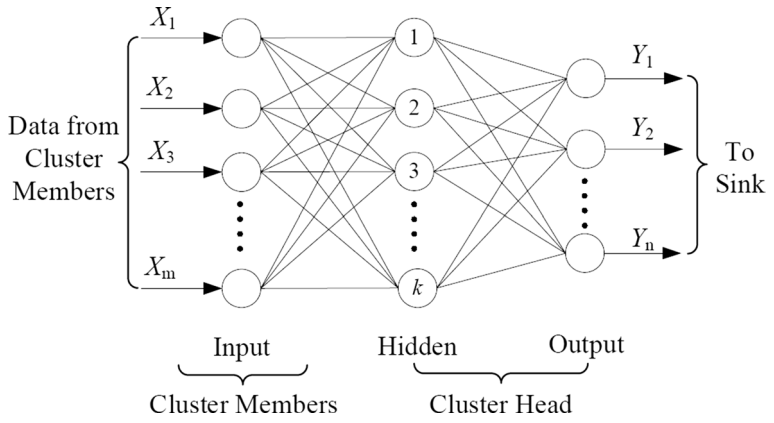


Fig. 9 The schematic model of BPNDA [29]

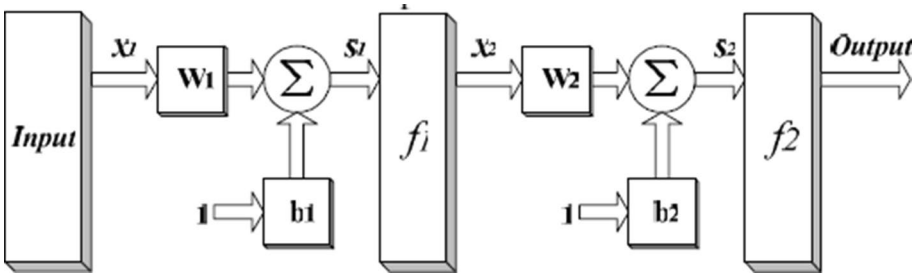


Fig. 10 The structure of BPNDA [29]

as delay, amount of data, and reliability. The structure of Back-Propagation Networks Data Aggregation (BPNDA) [29] scheme are shown in Figs. 9 and 10, respectively. Here a data aggregation scheme for WSN was proposed to reduce the communication traffic, save the energy, and improve the accuracy of information-gathering. The collected data from sensor were processed at CH using Back-Propagation neural network before transmitting them to the sink.

In [28] a fuzzy logic is used to separate the occurrences of failure in the data based on the existing false positive instances. It explores the use of various context information to statistically estimate the network condition with negligible overhead. An energy efficient context monitoring framework is presented in [54] which adjusts the monitoring policy based on the learning of associations between the attributes. The schemes in [13, 55, 56] employ self-organized map (SOM) as a clustering approach which is a three-layer neural network of input, middle, and output layer. $X = \left(x_1, x_2, \ldots x_d\right)^T$ represents the input layer and it is fully connected to middle layer to give result to output neural layer, $Y = \left(y_1, y_2, \ldots, y_m\right)$ as shown in Fig. 11. The training process of SOMDA iteratively updates the synaptic weights of the winner and its neighbors' neurons. At each training step, a sample vector, $x_{i,d}$, is randomly selected from the input dataset. As training progresses, the algorithm calculates the Euclidean distance between every weight and input vector x_d. The node with a weight vector of closest distance to the input vector is tagged as the best-matching unit (BMU), j^*.

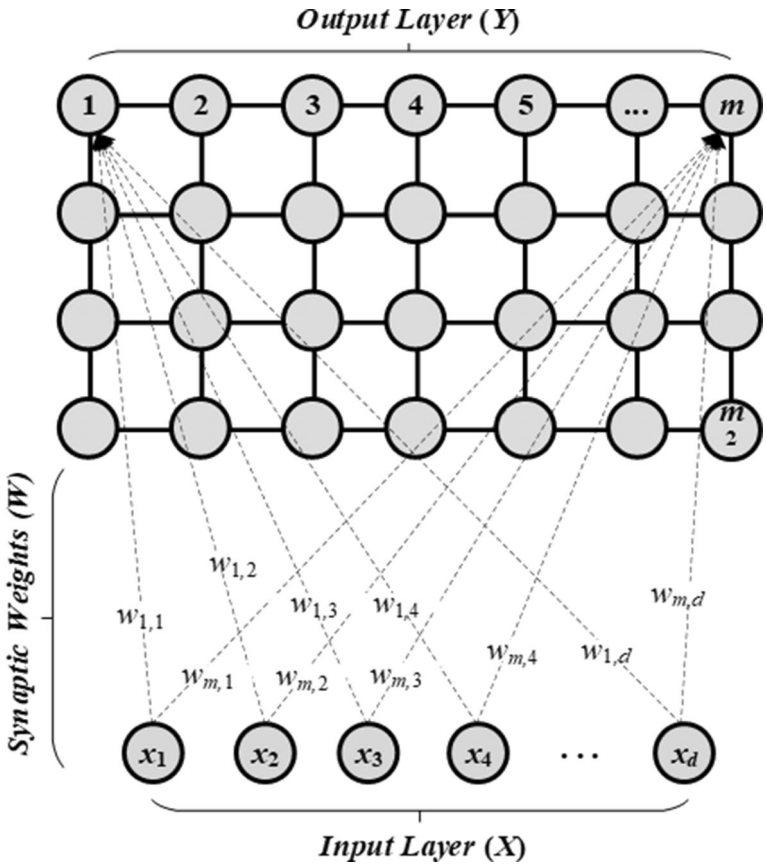


Fig. 11 The diagram of the SOM network [13]

$$j^* = \min_j \left(\sqrt{\sum_{i=0}^{d} \left(x_i - w_{im} \right)^2} \right) \tag{2}$$

The synaptic weight vector, $W_k = \left(w_{k,1}, w_{k,2}, \ldots w_{k,m} \right)$, is the directed links between the input layer X and out layer$Y$, where $k \in \left\{ 1, 2, \ldots, m^2 \right\}$ expresses the index of kth node of the output layer as shown in Fig. 11. The synaptic weight at time $(t+1)$, $w_j(t+1)$, is obtained as follows.

$$w_j(t+1) = w_j(t) + \alpha(t) \cdot h_{ci}(t) \left[x_i - w_j(t) \right] \tag{3}$$

where α and t represent the learning rate factor and the iteration of the training process, respectively. The Gaussian neighborhood function, $h_{ci}(t)$, indicates how strongly the neighbor neurons are connected around the winner during the learning process, and all the neurons close to each other are arranged in the two-dimensional grid as shown in Fig. 12. It is specified as:

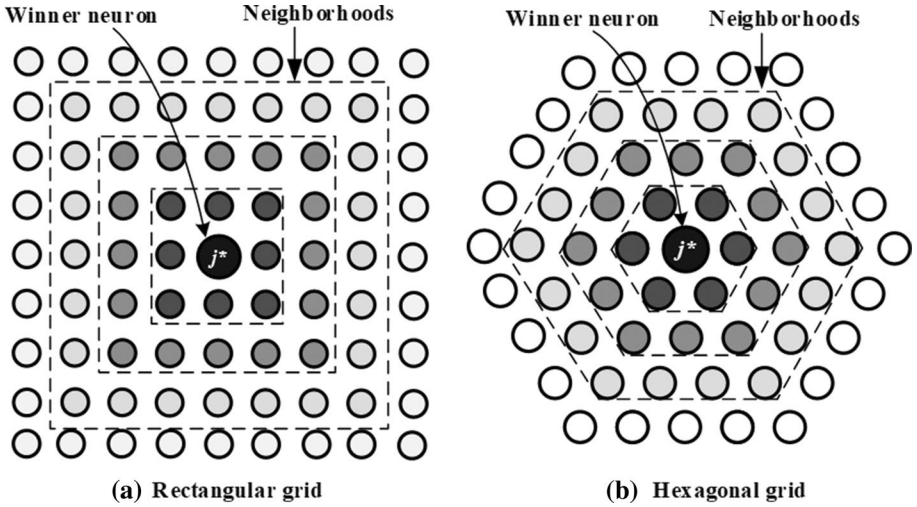


Fig. 12 The grid representation of the SOM-NN [13]

$$h_{ci}(t) = \exp\left(-\frac{r_c, r_i^2}{2\sigma^2(t)}\right) \tag{4}$$

where r_c and r_i represent the location of the winner neuron_c and neuron_i, in the grid and $\|r_c, r_i\|^2$ is the distance between them.

The reinforcement learning technique allowing a sensor node (an agent) to interact with its environment using Q-learning [57] is shown in Fig. 13. The maximization of the efficiency of data collection from sensor nodes depends on the movement policy of the mobile element (ME) with which the best position of an ME is decided. Figure 14 depicts how the policy is applied in accessing the reward. In an uncertain environment, the data gathering process by ME is dynamically modeled through the Markov decision processes to enhance the movement of ME [58, 59]. The authors integrated the reinforcement learning algorithm with the data fusion process to develop an adaptive system. It is employs a kernel-based learning method which enhances the efficiency of data integration and fusion.

The Mahalanobis distance-based radial basis function-based Extreme Learning Machine (MELM) [14, 60] is a two-stage data aggregation scheme with the projection stage and clustering stage as shown in Fig. 15. In the projection stage the weights of the link, w_n, are adjusted with the center of the neuron, μ_k, at the intermediate layer. The primary objective of the training process with the neurons in the intermediate layer is to

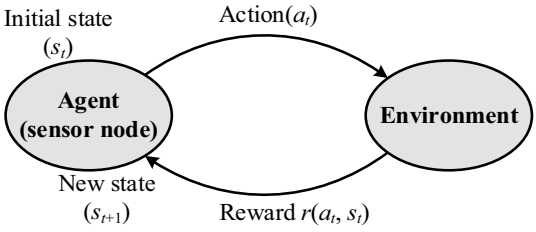


Fig. 13 The structure of the Q-learning method

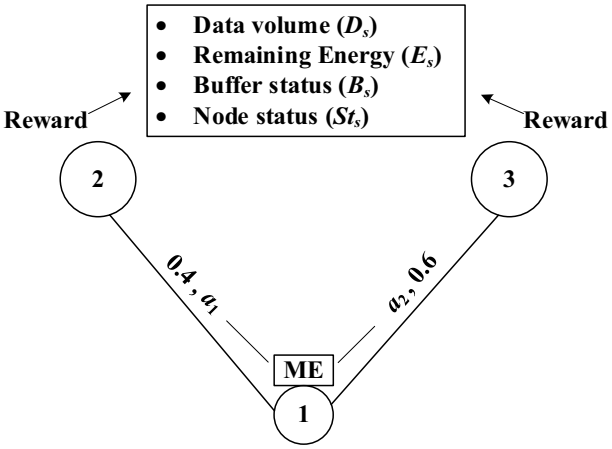


Fig. 14 The movement policy of ME

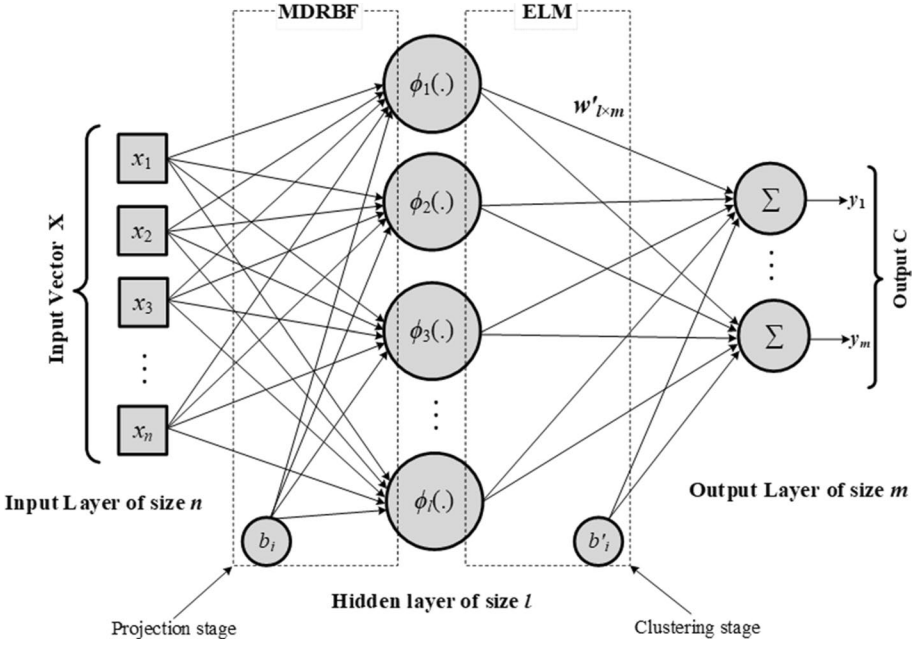


Fig. 15 The structure of the MDRBF-based ELM neural network [14]

place the center of their Gaussian functions as described below. In the clustering stage, the value of neurons is adjusted with output weight, $w_i^{'}$, with the training and tuning process to achieve the target output. The output weight of the clustering stage is analytically determined via mathematical manipulation. As a result, the proposed scheme can improve the accuracy of clustering with small computation overhead. Figure 15 is the structure of the MDRBF-based ELM neural network.

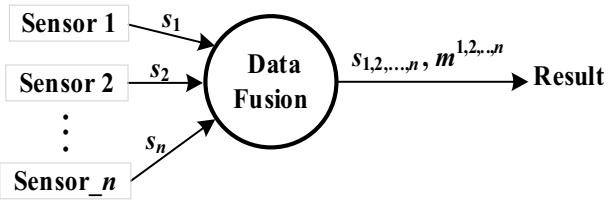


Fig. 16 The DS process with n sensors

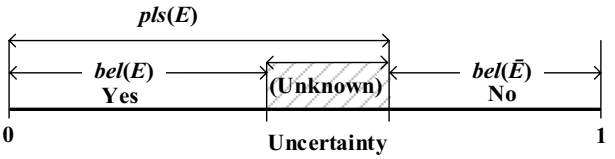


Fig. 17 The relationship between belief, disbelief, unknown, and plausibility function

2.3 Evidence Theory-Based Method

Evidence theory is a powerful and concrete method of fusion which extract precise information from multiple sensor nodes [37]. It transforms multi-source subjective and conflicting information into a decision-making result, and utilizes the combination of mass function from different sources. Dempster–Shafer (DS) is an evidence-based theory, and it is regarded as one of effective approaches for data fusion. The combination rule of DS theory can effectively merge the measures of evidence from different sources as shown in Fig. 16. The relationship between belief, disbelief, unknown and plausibility function in the DS theory are shown in Fig. 17.

A generic evidence fusion scheme [61] was proposed using the DS theory to deal with the uncertainty in the sensor readings and capture the features of the environment. A two-step technique is used to build a belief function from the sensor data, and the rule of combination of three sensor data are expressed as:

$$m^{1,2,3}(E) = \frac{\sum_{s_1 \cap s_2 \cap s_3 = E} m^1(s_1) \cdot m^2(s_2).m^3(s_3)}{\sum_{s_1 \cap s_2 \cap s_3 \neq \Theta} m^1(s_1) \cdot m^2(s_2).m^3(s_3)} \tag{5}$$

Here $m^{1,2,3}(E)$ evidence is obtained using three sensor nodes, $m^1(s_1)$, $m^2(s_2)$, and $m^3(s_3)$. The generalized combinatorial rule of DS theory for n sensor nodes is defined as follows:

$$m^{(1,2\ldots n)}(E) = m^1(s_1) + m^2(s_2) + \cdots + m^n(s_n)$$

$$m^{(1,2\ldots n)}(E) = \frac{1}{1-K} \sum_{\cap_i s_i = E} \left(\prod_{1 \le i \le n} m^i(s_i) \right) C \neq \Theta \tag{6}$$

$$K = \sum_{\cap s_i = \Theta} \left(\prod_{1 \le i \le n} m^i(s_i) \right) \tag{7}$$

A multi-sensor data fusion system [62] was proposed based on the DS theory to allow the detection of residence in a room based on various sources like temperature,

humidity, and light. It assigns a mass to the sensor data by using the mass function to combine all the masses by the combination rules, and then make a decision. Here the occupancy sensing problem is expressed as a classification problem, and each class is considered by a separate set of characteristics. Before computing the mass of the data obtained from a sensor, it is sent to the data fusion center (DC) (shown in Fig. 18) to compute the probability density function. The DC is located within the building premise to increase the accuracy and reduce the cost.

In [63] a DS theory-based fusion scheme was proposed for event detection in twitter. In this scheme two types of data are involved in the fusion, the features extracted from the text using the bag-of-words technique and the visual features extracted by applying the scale-invariant feature transform. The DS theory of evidence is applied so as to combine the data from the two sources, and the method is depicted in Fig. 19. A feature belonging for either text, t, or image, $\bar{t}$, and θ, refers to uncertainty inherit in the theory of evidence. All this constitute the frame of discernment, Θ:

$$\Theta = \left\{t, \bar{t}, \theta\right\}$$

Various techniques proposed for data fusion are compared in Table 2 regarding the employed machine learning approach, complexity, and purpose.

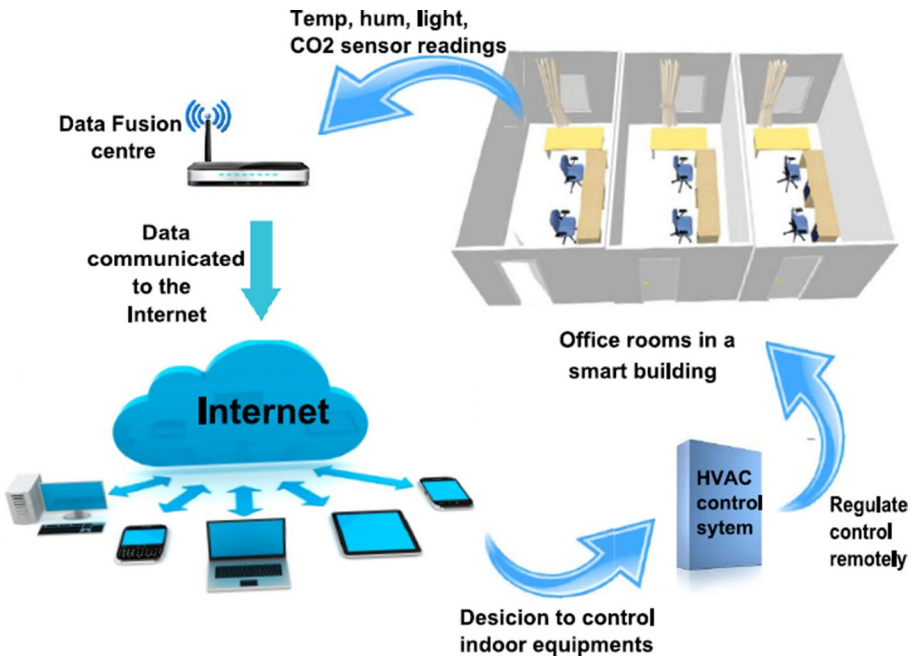


Fig. 18 The IoT structure for residence sensing [62]

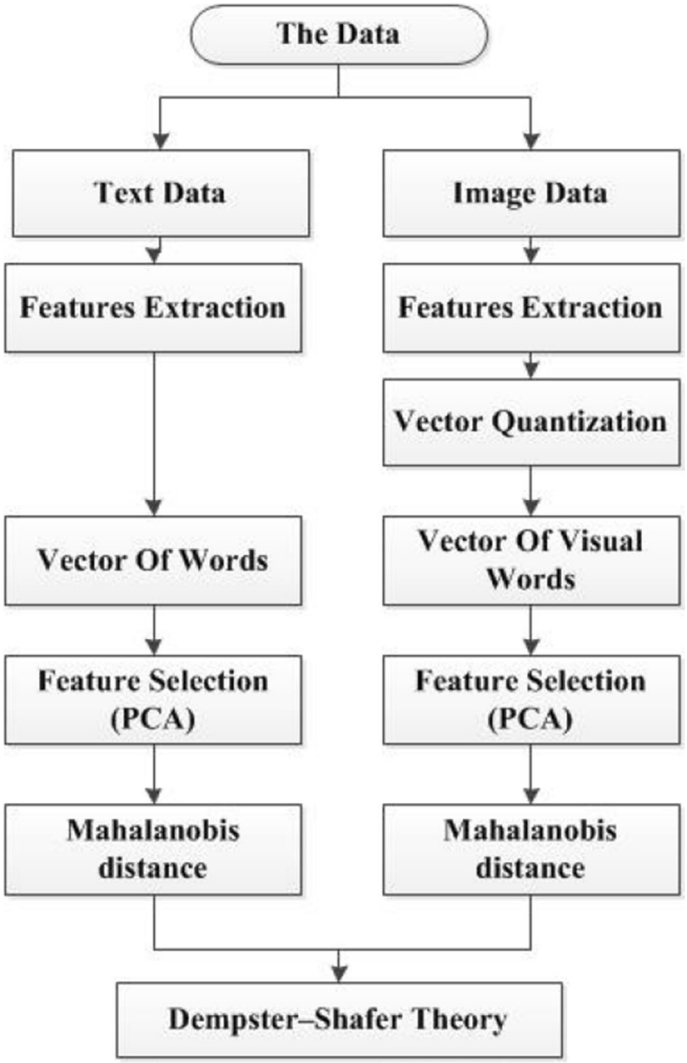


Fig. 19 The block diagram of fusion for twitter data [63]

3 Opportunities and Challenges

A huge amount of data are continuously generated in IoT environment, and it is very challenging to efficiently handle them since the data generated by the sensors are not precise and contain many outliers. Extracting reliable and accurate information is critical because the low-quality data may negatively affect the result of the overall data fusion operation [67, 68]. The opportunities and challenges with data fusion using various techniques are as follows.

Table 2 The comparison of different techniques proposed for data fusion

Study	Machine learning approach	Complexity	Purpose
• Bayesian approach to covariance estimation and data fusion [23] • Design of an Adaptive Bayesian system for sensor data fusion [22]	Bayesian analysis	Moderate	Data fusion and analysis
• Multiscale recursive estimation, data fusion, and regularization [25] • Density-based averaging–a new operator for data fusion [64]	Recursive operator	Low	Data fusion and estimation
• A Fuzzy-logic-based method for data fusion in WSN [32] • An autonomous multisensor data fusion based on fuzzy-logic [34] • Neural networks and statistical learning [30]	Fuzzy logic method	Moderate	Data fusion and clustering
• Redundancy reduction in SOM merging for scalable data clustering [65] • An RBF-NN-Based Data Aggregation Algorithm for WSN [31] • A fuzzy deep neural network-based data classification [33] • A data fusion method of WSNs based on swarm algorithm optimized BP-NN [66]	Neural Network Method	High	Data gathering and classification
• Application of Dempster–Shafer for Data Fusion [36] • Distributed data fusion in the Dempster–Shafer framework [37] • D–S-based data fusion investigation system for WSN [61] • Sensor Data Fusion for Residence Sensing environment using D–S theory [62]	Evidence theory (Dempster–Shafer)	High	Data fusion and decision making

3.1 Opportunities

- Filtering of data: Sensor data are noisy and imprecise, and thus filtering of data is needed to make data more intelligent, decisive, sensible, and precise. Various filters including Kalman filter and Moving-average filter (MAF) could be employed for pre-processing of data [69, 70]. An adaptive approach is also needed to improve the filtering operation of sensor data for real-time IoT environment.
- Data analysis: Analysis of the fused data needs to be accurate and fast to provide timely service. The probabilistic technique such as Bayesian decision network might be effective for analyzing heterogeneous data. The Bayesian approach for the estimation of the covariance of data [23] and Bayesian inference-based data fusion [41–46] are expected to be effective for the integration of sensor data.
- Power consumption: Data fusion and classification need to be efficient to increase the lifetime the WSN and IoT devices by removing outliers and redundant data. Clustering of the nodes based on data similarity and density would improve the power efficiency. Various machine learning technique would improve the power efficiency via effective clustering [13, 14].
- Security and information: The data fusion operation needs to be done in consideration of the security issue which hides and encrypts the information. A new approach integrating the fusion and encryption of data would be important.
- Knowledge and decision-making: Data fusion needs to help extract knowledge from multi-source data to make accurate decision. Evidence theory is a powerful and concrete method of fusion which extract precise information from multiple sensor nodes and take decision based on the fused data [37]. Data mining based on Bayesian network [50] is expected to be effective for integrating the quantitative and qualitative knowledge into a comprehensive probabilistic information.
- Self-organized system: Different contexts may require different sensory capabilities, and it is not desirable to determine a priori the subset of sensors to use. In a real-world scenario, the context conditions may change over time, implying the need for a system capable of dynamically selecting the subset of sensory devices. The SOM-based approaches will be effective for implementing context-aware self-organized system.
- Clustering and classification of data: Since sensors generate uncertain imperfect data containing outliers, a new efficient approach for data fusion is needed to maximize the performance of fusion and hosting network. Here node clustering based on the data density and similarity will play an important role.

3.2 Challenges

- Multivariate data analysis: Due to the complexity of the data, analysis and visualization of multivariate data are imperative. IoT environments are heterogeneous due to disparate sources of data and devices. There is quite limited study on the covariance and multivariate analysis of the sensor data. The effectiveness of distributed multivariate outlier detection also needs to be enhanced in term of data communication and energy efficiency.
- Optimization with machine learning model: The researchers have proposed to employ ELM to dramatically reduce the computation time of training. However,

instability may occur due to random selection of the weights and biases of the model. A systematic approach needs to be developed to decide optimal values for target problem.

4 Conclusion

Tremendous amount of data are continuously generated in smart IoT environment, which are usually transmitted through wireless network including WSN. Such data are required to be efficiently collected and analyzed to make decisions on the service. This induces various challenges, and timely, accurate data fusion and analysis of sensor data is one of key issues. The performance of data fusion in the IoT and WSN environment can be significantly improved if the errors and uncertainty in the sensor data are reduced by proper fusion considering the context.

Numerous researches and developments have been made on data fusion utilizing various approaches to face the challenges in big data analysis in WSN and IoT. In this article we have presented a literature survey on data fusion proposed for reliable and accurate operation. Here the schemes combine the data obtained from various sources and extract meaningful information to help the decision process. The opportunities and challenges with data fusion in the IoT and WSN environment are also summarized. There still exist numerous challenges and issues needing attention of the researchers in the future.

Acknowledgements This work was partly supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No. 2016-0-00133, Research on Edge computing via collective intelligence of hyperconnection IoT nodes) Korea, under the National Program for Excellence in SW supervised by the IITP(Institute for Information & communications Technology Promotion) (2015-0-00914), Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2019R1I1A1A01058780, Efficient Management of SDN-based Wireless Sensor Network Using Machine Learning Technique) the second Brain Korea 21 PLUS project.

References

1. Rahmati, A., Shepard, C., Tossell, C., Zhong, L., & Kortum, P. (2015). Practical context awareness: Measuring and utilizing the context dependency of mobile usage. *IEEE Transactions on Mobile Computing., 14*(9), 1932–1946.
2. Alam, F., Mehmood, R., Katib, I., Albogami, N. N., & Albeshri, A. (2017). Data fusion and IoT for smart ubiquitous environments: A survey. *IEEE Access, 5*, 9533–9554.
3. Pinto, A. R., Montez, C., Araújo, G., Vasques, F., & Portugal, P. (2014). An approach to implement data fusion techniques in wireless sensor networks using genetic machine learning algorithms. *Information Fusion, 15*, 90–101.
4. El Faouzi, N.-E., & Klein, L. A. (2016). Data fusion for ITS: techniques and research needs. *Transportation Research Procedia, 15*, 495–512.
5. Collotta, M., Messineo, A., Nicolosi, G., & Pau, G. (2014). A dynamic fuzzy controller to meet thermal comfort by using neural network forecasted parameters as the input. *Energies, 7*(8), 4727–4756.
6. Collotta, M., Pau, G., & Bobovich, A. V. (2017). A fuzzy data fusion solution to enhance the QoS and the energy consumption in wireless sensor networks. *Wireless Communications and Mobile Computing, 2017*, 1–10.
7. Koshmak, G., Loutfi, A., & Linden, M. (2016). Challenges and issues in multisensor fusion approach for fall detection. *Journal of Sensors*. https://doi.org/10.1155/2016/6931789.
8. Paradis, L., & Han, Q. (2007). A survey of fault management in wireless sensor networks. *Journal of Network and systems management., 15*(2), 171–190.

9. Wan, J., Chen, M., Xia, F., Di, L., & Zhou, K. (2013). From machine-to-machine communications towards cyber-physical systems. *Computer Science and Information Systems., 10*(3), 1105–1128.
10. Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning, 2*(1), 1–127.
11. Gilbert, E. P. K., Kaliaperumal, B., Rajsingh, E. B., & Lydia, M. (2018). Trust based data prediction, aggregation and reconstruction using compressed sensing for clustered wireless sensor networks. *Computers & Electrical Engineering, 72*, 894–909.
12. Abukhalaf, H., Wang, J., & Zhang, S. (2015). Outlier detection techniques for localization in wireless sensor networks: A survey. *International Journal of Future Generation Communication and Networking., 8*(6), 99–114.
13. Ullah, I., & Youn, H. Y. (2019). A novel data aggregation scheme based on self-organized map for WSN. *The Journal of Supercomputing, 75*, 3975–3996.
14. Ullah, I., & Youn, H. Y. (2020). Efficient data aggregation with node clustering and extreme learning machine for WSN. *The Journal of Supercomputing*. https://doi.org/10.1007/s11227-020-03236-8.
15. Hall, D. L., & McMullen, S. A. (2004). *Mathematical techniques in multisensor data fusion*. Norwood: Artech House.
16. Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Boca Raton: Chapman and Hall/CRC.
17. Abdulhafiz, W.A., & Khamis, A. (2013). Bayesian approach to multisensor data fusion with Pre-and Post-Filtering. In *IEEE* (pp. 373–378).
18. Murphy, K.P., & Russell, S. (2002). Dynamic bayesian networks: Representation, inference and learning.
19. Zhang, Y., & Ji, Q. (2006). Active and dynamic information fusion for multisensor systems with dynamic Bayesian networks. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 36*(2), 467–472.
20. De Paola, A., Gaglio, S., Re, G. L., & Ortolani, M. (2011). *Multi-sensor fusion through adaptive bayesian networks* (pp. 360–371). New York: Springer.
21. van Kasteren, T., & Krose, B. (2007). Bayesian activity recognition in residence for elders. *Proceedings of the International Intelligent Environments Conference*. https://doi.org/10.1049/cp:20070370.
22. De Paola, A., & Gagliano, L. (2014). Design of an adaptive Bayesian system for sensor data fusion. In S. Gaglio & G. Lo Re (Eds.), *Advances onto the Internet of Things* (pp. 61–76). New York: Springer.
23. Weng, Z., & Djurić, P. M. (2012). A Bayesian approach to covariance estimation and data fusion. In *2012 proceedings of the 20th European signal processing conference (EUSIPCO)* (pp. 2352–2356).
24. Azmani, M., Reboul, S., Choquel, J.-B., & Benjelloun, M. A. (2009). Recursive fusion filter for angular data. In *IEEE* (pp. 882–887).
25. Chou, K. C., Willsky, A. S., & Benveniste, A. (1994). Multiscale recursive estimation, data fusion, and regularization. *IEEE Transactions on Automatic Control, 39*(3), 464–478.
26. Hou, X., Zhang, D., & Zhong, M. (2014). Data aggregation of wireless sensor network based on event-driven and neural network. *Chinese Journal of Sensors and Actuators., 27*(1), 142–148.
27. Wang, Q., Liao, H., Wang, K., & Sang, Y. (2011). *A variable weight based fuzzy data fusion algorithm for WSN* (pp. 490–502). New York: Springer.
28. Shell, J., Coupland, S., & Goodyer, E. (2010). Fuzzy data fusion for fault detection in wireless sensor networks. In *IEEE* (pp. 1–6).
29. Sun, L.-Y., Cai, W., & Huang, X.-X. (2010). Data aggregation scheme using neural networks in wireless sensor networks. In *IEEE* (pp. V1–725).
30. Du, K.-L., & Swamy, M. N. (2013). *Neural networks and statistical learning*. New York: Springer.
31. Wang, J., Wang, K., Caoa, Y., Younn, G., & Kimb, J.-U. (2017). A RBF neural network based data aggregation algorithm for wireless sensor networks. *Fuzzy Systems and Data Mining III: Proceedings of FSDM, 2017*(299), 428.
32. Manjunatha, P., Verma, A., & Srividya, A. (2008). Multi-sensor data fusion in cluster based wireless sensor networks using fuzzy logic method. In IEEE; 2008. p. 1–6.
33. Deng, Y., Ren, Z., Kong, Y., Bao, F., & Dai, Q. (2017). A hierarchical fused fuzzy deep neural network for data classification. *IEEE Transactions on Fuzzy Systems, 25*(4), 1006–1012.
34. Stover, J. A., Hall, D. L., & Gibson, R. E. (1996). A fuzzy-logic architecture for autonomous multisensor data fusion. *IEEE Transactions on Industrial Electronics., 43*(3), 403–410.
35. Challa, S., & Koks, D. (2004). Bayesian and Dempster–Shafer fusion. *Sadhana, 29*(2), 145–174.
36. Yi, P., & Zhang, S. (2017). Application of Dempster–Shafer data fusion technique in support of decision making with big data. *Transportation Research Record: Journal of the Transportation Research Board., 2645*, 32–37.

37. Kanjanatarakul, O, & Denœux, T. (2017). Distributed data fusion in the Dempster–Shafer framework. In *IEEE* (pp. 1–6).
38. Sentz, K., & Ferson, S. (2002). Combination of evidence in Dempster–Shafer theory. In *Citeseer* (vol. 4015).
39. Zhang, D., Cao, J., Zhou, J., & Guo, M. (2009). Extended Dempster–Shafer theory in context reasoning for ubiquitous computing environments. In *IEEE* (pp. 205–212).
40. Júlia, J. (2018). Thesis code for DS theory. Contribute to you-leee/Dempster–Shafer development by creating an account on GitHub. Retrieved 11 December, 2018, from https://github.com/you-leee/Dempster-Shafer.
41. Jaramillo, V. H., Ottewill, J. R., Dudek, R., Lepiarczyk, D., & Pawlik, P. (2017). Condition monitoring of distributed systems using two-stage Bayesian inference data fusion. *Mechanical Systems and Signal Processing, 87*, 91–110.
42. Mil, S., & Piantanakulchai, M. (2018). Modified Bayesian data fusion model for travel time estimation considering spurious data and traffic conditions. *Applied Soft Computing, 72*, 65–78.
43. Taylor, C. N., & Bishop, A. N. (2019). Homogeneous functionals and Bayesian data fusion with unknown correlation. *Information Fusion, 45*, 179–189.
44. Sharma, G., Singh, K., Gupta, G., Shroff, G., Agarwal, P., & Pandey A., et al. (2017). System and method for visual Bayesian data fusion.
45. Echeverri, A. F., Medeiros, H., Walsh, R., Reznichenko, Y., & Povinelli R. (2017). Hierarchical Bayesian data fusion for robotic platform navigation. arXiv:170406718.
46. Xue, J., Leung, Y., & Fung, T. (2017). A Bayesian data fusion approach to spatio-temporal fusion of remotely sensed images. *Remote Sensing, 9*(12), 1310.
47. Blasch, E., Chen, Y., Chen, G., Shen, D., & Kohler, R. (2014). Information fusion in a cloud-enabled environment. In K. J. Han, B. Y. Choi, & S. Song (Eds.), *High performance cloud auditing and applications* (pp. 91–115). New York: Springer.
48. Datcu, M., & Seidel, K. (1999). Bayesian methods: Applications in information aggregation and image data mining. *International Archives of Photogrammetry and Remote Sensing., 32*(7), 4–3.
49. Shinde, T. A., & Prasad, J. R. (2017). IoT based animal health monitoring with naive Bayes classification. *IJETT*. https://doi.org/10.23883/ijrter.2017.3035.qudpb.
50. Chen, Y. M., Hsueh, C.-S., & Wang, C.-K. (2016). Data mining of Bayesian networks to select fusion nodes from wireless sensor networks. *International Journal of Computer Science Issues (IJCSI)., 13*(4), 11.
51. Chen, X., Li, X. (2016). Virtual temperature measurement for smart buildings via Bayesian model fusion. In *IEEE* (pp. 950–953).
52. Gao, J.-P., Xu, C.-B., Zhang, L., Zheng, J.-L., Shu, H., & Yuan, X. (2017). *A method of information fusion based on fuzzy neural network and its application* (p. 01015). EDP Sciences: Les Ulis.
53. Sung, W.-T. (2009). Employed BPN to multi-sensors data fusion for environment monitoring services. *Autonomic and Trusted Computing.*. https://doi.org/10.1007/978-3-642-02704-8_12.
54. Kang, S., Lee, J., Jang, H., Lee, Y., Park, S., & Song, J. (2010). A scalable and energy-efficient context monitoring framework for mobile personal sensor networks. *IEEE Transactions on Mobile Computing., 9*(5), 686–702.
55. Lee, S., & Chung, T. (2004). *Data aggregation for wireless sensor networks using self-organizing map* (pp. 508–517). New York: Springer.
56. Aghajari, E., & Chandrashekhar, G. D. (2017). Self-organizing map based extended fuzzy C-means (SEEFC) algorithm for image segmentation. *Applied Soft Computing., 54*, 347–363.
57. Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning, 8*(3–4), 279–292.
58. Marwaha, S, Tham, C. K., & Srinivasan, D. (2002). Mobile agents based routing protocol for mobile ad hoc networks. In *IEEE* (pp. 163–167).
59. Lu, Y., Zhang, T., He, E., & Comşa, I.-S. (2018). Self-learning-based data aggregation scheduling policy in wireless sensor networks. *Journal of Sensors*. https://doi.org/10.1155/2018/9647593.
60. Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing., 70*(1–3), 489–501.
61. Senouci, M. R., Mellouk, A., Aitsaadi, N., & Oukhellou, L. (2016). Fusion-based surveillance WSN deployment using Dempster–Shafer theory. *Journal of Network and Computer Applications, 64*, 154–166.
62. Nesa, N., & Banerjee, I. (2017). IoT-based sensor data fusion for occupancy sensing using Dempster–Shafer evidence theory for smart buildings. *IEEE Internet of Things Journal, 4*(5), 1563–1570.
63. Alqhtani, S. M., Luo, S., & Regan, B. (2015). Multimedia data fusion for event detection in twitter by using dempster-shafer evidence theory. *International Journal of Computer, Electrical, Automation, Control and Information Engineering, World Academy of Science, Engineering and Technology., 9*(12), 2234–2238.
64. Angelov, P., & Yager, R. (2013). Density-based averaging–A new operator for data fusion. *Information Sciences, 222*, 163–174.

65. Ganegedara, H., & Alahakoon, D. (2012). Redundancy reduction in self-organising map merging for scalable data clustering. In *IEEE* (pp. 1–8).
66. Wu, W., Xu, B., & Cao, M. (2016). A data fusion method of WSNs based on glowworm swarm algorithm optimized BP neural networks. *Revista Ibérica de Sistemas e Tecnologias de Informação., 17A*, 73.
67. Mahdavinejad, M. S., Rezvan, M., Barekatain, M., Adibi, P., Barnaghi, P., & Sheth, A. P. (2018). Machine learning for Internet of Things data analysis: A survey. *Digital Communications and Networks, 4*(3), 161–175.
68. Lee, I., & Lee, K. (2015). The Internet of Things (IoT): Applications, investments, and challenges for enterprises. *Business Horizons, 58*(4), 431–440.
69. Moving Average. (2018). In: Wikipedia. Retrieved 10 December, 2018, from https://en.wikipedia.org/w/index.php?title=Moving_average&oldid=869594777.
70. Shivashankarappa, N., Adiga, S., Avinash, R., & Janardhan, H. (2016). Kalman filter based multiple sensor data fusion in systems with time delayed state. In *IEEE* (pp. 375–382).

Ihsan Ullah received the B.S. and M.S. degree in computer science from University of Peshawar, Pakistan, in 2001 and 2004, respectively, and the Ph.D. degree in computer engineering from Sungkyunkwan University, Suwon, Korea, in 2019. He is currently research fellow in the UTRI (Ubiquitous Computing Technology Research Institute), department of electrical and computer engineering at Sungkyunkwan University, Korea. His research interests include Data aggregation, Data Fusion, IoT (Internet of Things), Intelligent system Artificial Intelligence, Machine learning, Distributed computing and Wireless Sensor Network.

Hee Yong Youn received the B.S. and M.S. degree in electrical engineering from Seoul National University, Seoul, Korea, in 1977 and 1979, respectively, and the Ph.D. degree in computer engineering from the University of Massachusetts at Amherst, in 1988. He had been Associate Professor of Department of Computer Science and Engineering, The University of Texas at Arlington until 1999. He is presently Professor of College of Software, Sungkyunkwan University, Suwon, Korea, and Director of Ubiquitous Computing Technology Research Institute. He has been also Consulting Professor of Software R&D Center, Device Solutions, Samsung Electronics, Korea. His research interests include distributed and ubiquitous computing, IoT, and intelligent system. He has published more than 400 papers in int'l journals and conference proceedings, and received Outstanding Paper Award from the 1988 IEEE International Conference on Distributed Computing Systems, 1992 Supercomputing, and 2012 IEEE Int'l Conf. on Computer, Information and Telecommunication Systems, 2014 The 6 the International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, respectively. Dr. Youn has also been General Chair of IEEE PRDC 2001, Int'l Conf. on Ubiquitous Computing Systems (UCS) in 2006 and 2009, UbiComp 2008, CyberC 2010, Program Chair of PDCS 2003 and UCS 2007. Dr. Youn is a senior member of the IEEE Computer Society.