INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Relationship-Cluster Head Selection and Data Compression Enabled Cluster-Based Aggregation Model for Social Internet of Things

J Meghana[1], J Hanumanthappa[1], S P Shiva Prakash[2]*, Kiril Krinkin[3]

1 Department of Studies in Computer Science, University of Mysore, Mysuru, Karnataka, India
2 Department of Information Science and Engineering, JSS Science and Technology University, Mysuru, Karnataka, India
3 Researcher, Co-evolutionary Artificial Intelligence, Cyprus

OPEN ACCESS

Received: 24-05-2023
Accepted: 22-09-2023
Published: 07-11-2023

Citation: Meghana J, Hanumanthappa J, Prakash SPS, Krinkin K (2023) Relationship-Cluster Head Selection and Data Compression Enabled Cluster-Based Aggregation Model for Social Internet of Things. Indian Journal of Science and Technology 16(41): 3605-3616. https://doi.org/10.17485/IJST/v16i41.1256

* **Corresponding author**.

shivasp@jssstuniv.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background: The Social Internet of Things (SIoT) combines IoT with social networking, enabling objects to interact based on social relationships and facilitating user-device interactions. However, SIoT networks generate massive data that must be transmitted and processed efficiently. We propose a cluster-based aggregation model for SIoT to address this, integrating relationship-cluster head selection with K-Means and data compression using Huffman coding. **Objectives:** The proposed model utilizes K-Means to select cluster heads based on object relationships. Objects are clustered using K-Means, and a Decision Tree considers object profiling and relationships to choose the cluster head. Selected cluster heads employ Huffman coding to compress data before transmission to the sink node. **Methods:** The data are aggregated at the cluster head and it is compressed before sending it to the destination device. Model is evaluated through simulation-based experiments using the SIoT-CCN simulator. Findings demonstrate that our proposed model outperforms existing approaches in terms of energy consumption, network lifetime, and data aggregation accuracy. **Findings:** Evaluation metrics include Silhouette Score, Number Distance between Train and Test K-Means, BIC Score, and Gradient of BIC Scores. Results for our proposed model include a Silhouette score of 0.459 for cluster numbers 2, 3, and 4, -1148.951 distance between train and test K-Means, -5227.080 BIC Score, and -96.445 Gradient of BIC score. For the Relation-Based clustering approach without data compression, the Silhouette score is 0.6266, -203.345 distance between Train and test K-Means, -7266.080 BIC Score, and -61.415 Gradient of BIC Scores. The Data Compression-enabled cluster-based aggregation model without relationship-based clustering achieves a Silhouette score of 0.58, -467.890 distance between train and test K-Means, -8981.786 BIC Scores, and -94.244 Gradient of BIC Score. **Novelty:** Integrating relationship-cluster head selection with K-Means and data compression with Huffman coding to develop a cluster-based

aggregation model for SIoT, demonstrating its potential to enhance SIoT performance.

Keywords: Clusterbased model; Cluster head selection; Data aggregation; Data compression; Social Internet of Things

1 Introduction

The Internet of Things (IoT) [1–3] is described as physical devices or "things" that are connected over a network which consists of devices with software, sensors, and various technologies to get connected and exchange information with other devices and systems over the network. Social IoT is a network of social objects that are considered intelligent based on the concept of a social relationship built between objects. The main objective of SIoT is to allow people to apply restrictions to preserve their privacy, allow objects to have their own social network, and access the outcome of the object's independent inter-object interaction by separating two levels of objects and people. The social network relation [4,5] SIoT consists of many smart objects widely connected across the network and location. These smart objects are of different sizes and have computation capabilities, communication capabilities, and sensing capabilities. However, these smart objects have limited memory, processing resources, bandwidth, power, and a limited lifetime. To increase the object lifetime by reducing the network bandwidth, save energy, and reduce the burden on the network, data aggregation plays a very crucial role. Data Aggregation involves the process of collection, aggregation, and sending the data to Base Stations which are produced by smart objects. In the cluster-based approach, distance is considered to group the devices into several clusters. [6]

Considering certain criteria in every cluster, a head of the cluster is chosen [7–10]. The main task of this head of the cluster is to receive aggregate and transmit the data. Each device within a cluster that has data to be transmitted sends data to the cluster head of the same cluster. Later head of the cluster receives the data and sends the data to the Base Station. Since data is transferred from the head of the cluster to the Base Station, it saves a large amount of energy and reduces the network's bandwidth. Since a member of the cluster sends data to the head of the cluster and the cluster head sends data to the base station it reduces the chances of collision and energy consumption [11,12]. The emergence of the Social Internet of Things (SIoT) has led to the development of many innovative applications that enable social interactions between people and objects. However, the large amounts of data generated by these objects can be challenging to handle. Also, for effective and efficient data aggregation in devices that have formed a social relationship with each other over the network and location, cluster-based data aggregation is very suitable as it is robust in nature, guarantees the accuracy of information shared, reduces the redundancy of the transmitted data, minimizes network's load as the head of the cluster will be the only device that will be sending the data to the base station. Therefore, in this paper, we propose a relationship-cluster head selection and data compression-enabled cluster-based aggregation model for SIoT. The novelty of this proposed approach lies in the integration of relationship-based clustering and Huffman coding for data compression in a cluster-based aggregation model for SIoT. It also takes into account the selection of cluster heads and data consistency, which are important considerations in SIoT. In contrast to existing literature, the proposed approach is a comprehensive solution that combines several techniques for efficient data transmission and storage in SIoT. Furthermore, it provides a robust and effective solution to reduce the data transmission amount between IoT devices and the cloud, while taking into account the selection of cluster heads and data consistency. The highlighting contributions of this work are:

- **Relationship Cluster Identification:** Identified the relationship clusters based on the dependencies among the devices or objects in the SIoT network.
- **Cluster Head Selection:** Used a relationship cluster head selection mechanism to identify the most appropriate cluster head for each relationship cluster. The selection mechanism considers factors such as the device's data processing capability, energy efficiency, and communication range.
- **Data Compression:** Data compression techniques clustering-based compression are applied to reduce the amount of data transmitted and stored by SIoT devices.
- **Cluster-Based Aggregation:** The compressed data from each device within the relationship cluster by the selected cluster head are aggregated, using a cluster-based aggregation mechanism. This mechanism considers factors such as the accuracy of data aggregation and communication reliability.

1.1 Related Works

The authors Ab Rouf Khan, and Mohammad Ahsan Chisthi[13] have given information about the IoT analysis and study of schemes for the aggregation of data with respect to the complexity of time and working principle. The author has designed an algorithm by the name LCA. This algorithm is being evaluated and works well when the cluster has fewer nodes. In comparison with CDA algorithm, the proposed algorithm works well when the cluster's node is increased. Yuiqing Zhang, and Yong Li[14] the author has proposed an algorithm for the identification of fewer unknown devices connected to the building Internet of Things system. The results show that spotting of the devices is obtained online for the dataset selected. Further, the validity of the proposed system called k- means clustering-based electrical equipment was verified. Khan, Ab Rouf[15] the author has compared the three data aggregation algorithms such as tree-based, based on cluster, and centralized aggregation algorithms. The obtained output shows that by considering a total number of 60 nodes the centralized data aggregation algorithm has worked well for the network which has fewer nodes. The algorithms cluster-based and tree-based have given the best output when there is an increase in the number of nodes. In Abhijith H V, H S Ramesh Babu[16] work, the authors have proposed an intelligent mechanism for the aggregation of data for remote sensing networks for IoT devices. The author is successful in avoiding data redundancy by aggregation technique called spatial. The method that has been proposed was tested for simulations and produced result efficient results. Shamim Yousefi, Hadis Karimipour, Farnaz Derakhshan[17] gave a precise survey on methods of aggregation of data in IoT. The author has discussed the difficulty in designing the data aggregation model. The author has even discussed the prediction of IoT trends in the coming time. Sirajudheen, Rekha, Shobha[18] taking into consideration the fundamental viewpoint and different data in WSN, an appraisal of different sorts of aggregation of data approaches and conventions were introduced in this methodology. In this composition, different methodology has been portrayed for the decrease of redundant data and mainly in aggregating of data. Fouad H. Awad, Awad, Murtadha, M. Hamad[19] has proposed a neural processor-based k-means clustering technique for the clustering data for mobile devices. It showed twice the development in speed with respect to smartphones compared to laptop processors. Additionally, when compared to parallel and distributed k-means algorithms the identification of the clustering speed has twice improved. Mohana S.D., S P Shiva Prakash, Kirill Krinkin[20] in this work, the method for feature selection is proposed based on certain rules and is successful in establishing the relationship for the classification of the services by using R-ANN. It provides an 89.62% accuracy for the services such as quality of air, weather, status of light, parking, and presence of people in SIoT environment when compared to the previously established model. Seyed Omid Mohammadi, Ahmad Kalhor, and Hossein Bodaghi[21] author have introduced an algorithm called K-Splits. It is an advancement of the hierarchical algorithm. The K-Splits algorithm was faster than other clustering algorithms. Hence, the K Splits algorithm is used to identify the correct location of centroids and further feed them to the K-Means algorithm as the initial point for better results. Gurumoorthy.S, Subhash.P, Pérez de Prado.P, Wozniak, M[22] Considering secured and routing of aware of energy in WN a CHS model is built. The energy, distance, and evaluation of trust are considered for selecting the optimal head of the cluster. The result shows that the delivery of packet ratio is 0.98 at 500 rounds for 200 nodes when compared to other methods considered. Panimalar Kathiroli, Kanmani Selvadurai[23] has proposed an algorithm named sparrow search with the combination of evolution of differential algorithm is used to solve the issue related to the efficiency of energy of selection of head of the cluster in WSN. This model has worked best in selecting the best head of the cluster when compared to other algorithms. Mohana S.D., S P Shiva Prakash, Kirill Krinkin[24] in this paper author has considered the research interest in the field of application of smart- city such as recommendation of application, analytics of information about device and discovery of service, personal interest from the community of research. In order to satisfy the application of smart city author has proposed CCNSim SIoT simulator. This simulator provides the basic functionality of a network simulator and AI-based visualization and analytics. This provides communication between the Software and Hardware interface. The proposed model is also capable of offering a configuration environment for the selection and navigation of objects, providing services, recommendation of services and applications, object profiling, and prediction of the behavior of users. The

conducted experiment and result show the capacity and facility provided by the simulator framework proposed are at their best. Ravi, M. Swamy Das, and Karthik Karmakonda[25] has proposed a CRDA schema for IoT network that guarantees the collection, and aggregation of data in an efficient manner and the transfer of data efficiently to other end. The author has initially presented the MSC algorithm to group clusters in IoT sensors which guarantees the transferring of data effectively. In the second phase of aggregation of data, the ISFO algorithm is used to maximize the trust of IoT devices. ROL-DNN algorithm is then used to process ways between the sensors of IoT aggregation of data and transferring. Farooq. O Singh. P, Hedabou. M, Boulila. W, Benjdira. B[26] framework called MLADCF is structured for the management of data for designing IoT devices. The obtained result has helped in reducing energy utilization which led to an extension in the life of the battery of the nodes that are connected. Bing Han Feng Ran, Jiao LI, Limin Yan, Huaming Shen, Ang Li[27] proposed an EH-WSNS model for data transmission in WSN. The experimental results showed that the proposed model can increase the lifetime for the maximum for WSN when compared to standard routing protocol.

1.2 Problem Statement

The Social Internet of Things (SIoT) is an emerging technology that enables devices and objects to communicate with each other in order to provide various services to users. One of the major challenges in SIoT is the efficient management of data generated by these devices, particularly in terms of data aggregation and storage. This is particularly challenging in SIoT applications that involve relationships between devices or objects, as the data generated by these devices can be highly interdependent. Therefore, there is a need for an efficient mechanism for aggregating data in SIoT applications with relationship dependencies.

1.3 System Model

In SIoT, the information and services gathered in the network are being used by the user with the help of various devices D_n, where the number n = 1, 2, n, based on the total number of applications A_p. Where p = 1, 2, 3…n, provides the independent services S respectively. The devices are heterogeneous in nature and each device has its device attributes D_A. The devices in network N, form social relationships R_S, with each other based on the relationship factors R_F and relationship criteria R_C, forming the SIoT between the devices D. Therefore, relationship management requires communication between the corresponding devices by exchanging the information. Hence, the system model is

$$\sum_1^{n,p}(D_n,\ A_p,\ S)|N : R_S\ for\ all\ (D_A,R_F,R_C) \tag{1}$$

Device D is actively participating w.r.t to corresponding digital information which has the device attributes D_A. Such as device D_C, device type D_T, device ID, D_{ID}, device manufacturer ID DM_{ID}, device brand ID DB_{ID}, device location D_L.

With the help of these attributes, the devices try to form relationships R_S, with each other based on criteria and factors between the devices. The data aggregation model is necessary for mutual information exchange in SIoT. Hence, the objective function is

$$AM = \sum_1^n f\,(D_A,R_F,R_C) \tag{2}$$

Subjected to $(R_F, R_C) = R_S$ and $D_A = D_n and\ A_p$

From Equation (2) the objective function has R_F and R_C have clustered the devices with respect to the relationship conditions and relationship factors. The device attributes D_A, namely the type of the device, protocols of the device, the ID of the device, locations of the device, and manufacturing ID of the device are the features which is responsible for selecting the head of the cluster using the decision model in the SIoT. Therefore, using Equation (2) proposed a solution to the decision model D_m. Hence, the decision model initially used k-means to cluster the device data then the decision tree algorithm for selecting the head of the cluster and finally aggregates the data using the compression technique Huffman coding. Hence, the model for data aggregation is

$$AM\ =\ DAgg(Dm:\ Ap,\ DA,\ RS,\ RF,\ RC) \tag{3}$$

Where D_m is the decision model using the k-means K_M, clustering technique to device data iteratively that tries to partition the data into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. The k-means equation is,

$$K_M = \sum_{j=1}^{k}\sum_{i=1}^{n}|Xij - Cj|2 \tag{4}$$

Where X= A_p,D_A, S, R_S, R_F, R_C, and C are cluster center data points. The decision tree DT has calculated the entropy and information gain to obtain the decision tree for selecting the cluster head.

The entropy E, is obtained through the purity and impurity of the group of observation, the entropy equation is,

$$E = \sum_{i=1}^{N} pilog2\ pi \tag{5}$$

Where i is the cluster class.

The information gain G, features of the cluster data which provides the cluster head selection and sub-cluster selection using the entropy of each group of the clusters of data. The information gained helps to split the nodes in the decision tree. The information gain equation is

$$G = E(CH) - E(CM) \tag{6}$$

Where (CM) is the cluster member in the node of the cluster head CH, E (*cluster member*) is the average entropy of the sub-nodes the clusters.

1.3.1 Relationship Cluster Identification

Let X = {x1, x2, ..., xn} be the set of n devices or objects in the SIoT network. The goal of relationship cluster identification is to group devices with relationship dependencies into clusters.

Let R = {r1, r2, ..., rm} be the set of m relationship clusters. Each relationship cluster is defined as a subset of devices that have a relationship dependency. The relationship cluster identification is done using a clustering algorithm such as K-means.

1.3.2 Cluster Head Selection

The cluster head selection goal is to identify the most relevant cluster head for each relationship cluster. Let h(r) be the cluster head for relationship cluster r. The Decision Tree(DT) is used to select the cluster head. The DT models the distribution of data in each relationship cluster and selects the device with the highest probability of being the cluster head. The probability of each device being the cluster head is given by the equation:

$$P(i|r) = exp(-||xi - \mu(r)||2/2\sigma2)/Z \tag{7}$$

where x_i is the data generated by device i in relationship cluster r, μ(r) is the mean of the data in relationship cluster r, σ^2 is the variance of the data in relationship cluster r, and Z is the normalization factor. The device with the highest probability of being the cluster head is selected as h(r).

1.3.3 Data Compression

Data aggregation DA_{gg} is the method of data collection with compression techniques that holds the data of corresponding devices and applications to the users of the SIoT. The cluster head is less than the cluster member, and the compression of data is done on the device data. The bandwidth information is deduced based on the type of protocol the Huffman technique is used by the device. The Huffman code equation is,

$$Pn = (P1,\ P2 \ldots Pn-2,\ Pn-1,\ Pn) \tag{8}$$

Where, without loss of generality, pi $\geq$ pi + 1, is constructed from the Huffman code of the (n $-$ 1) element probability distribution,

$$P\,n - 1 = P1,\ P\,2 \ldots P\,n - 2,\ P\,n - 1 + P\,n \tag{9}$$

Where the n-element probability of distribution works by creating a binary tree, where the nodes store the characters of the data, initially all nodes in the tree are considered leaf nodes, these leaf nodes contain the characters of the data, and the nodes also contain the frequency of occurrence of the character and link to the child node. Here as a followed convention, the left child of the tree is represented by bit '0', whereas the right child with '1'. The priority queue is used to store the nodes that retrieve the node with the least frequency when popped. The process of Huffman data compression starts with creating a leaf node of a binary tree, provided there is more than one node in the priority queue, removing two nodes of the high frequency from the priority queue, and in the next step creating a new internal tree node with the recently popped two nodes as child node with a frequency that is equal to the sum of the two nodes frequency, next step is to add the newly created node to the priority queue,

the remaining tree node is the root node and the resulting tree is Huffman tree. Hence, this model is used for Data Aggregation purposes in the SIoT.

Let D(r) = {d1, d2, ..., dk} be the devices generated data in relationship cluster r. The data compression is done using Huffman coding, which assigns a variable-length code to each data point in D(r). The compressed data for each device i in relationship cluster r is given by:

$$C(di) = Huffman(D(r)) \tag{10}$$

1.3.4 Cluster-Based Aggregation

The goal of cluster-based aggregation is to aggregate the compressed data from each device within the relationship cluster by the selected cluster head. The aggregated data is then transmitted to the sink node. Let A(r) be the aggregated data for relationship cluster r. The cluster-based aggregation is done using the following equation:

$$A(r) = 1/k * sum\,(i = 1\,to\,k)\,C(di) \tag{11}$$

where $C(d_i)$ is the compressed data point for device i in relationship cluster r. The proposed model aims to provide an efficient mechanism for managing data in SIoT applications with relationship dependencies, thus enabling more effective use of the SIoT technology. The model minimizes the amount of data transmitted and stored by SIoT devices while still maintaining high accuracy in data aggregation, as shown by the objective function.

2 Proposed method

Figure 1 shows the design of the proposed model. The proposed methodology for Relationship-cluster head selection and Data compression cluster-based aggregation model SIoT is based on the concept of relationship identification between data elements and grouping them into clusters for efficient data transmission and storage. In this approach, the selection of the relationship cluster head is based on a combination of factors such as proximity, similarity, and connectivity. The head is responsible for managing the cluster and performing aggregation and compression of data before transmission to the cloud. The compression is achieved using a Huffman coding algorithm, which assigns variable-length codes to the clusters based on their frequency of occurrence. The novelty of this proposed approach lies in its integration of relationship-based clustering, head selection, and Huffman coding for data compression in SIoT. This approach not only reduces the data transmitted amount and the associated consumption of energy but also improves the data accuracy and reliability due to the aggregation performed by the relationship cluster head. Compared to existing literature, this approach offers a more comprehensive and efficient solution to data aggregation and compression in SIoT. The proposed technique combines clustering, head selection, and Huffman coding to achieve optimal compression and transmission of data, thus addressing the challenges of energy consumption, bandwidth limitations, and data accuracy in SIoT. The methodology design for Relationship-cluster head selection and Data compression enabled cluster-based aggregation model for the Social Internet of Things (SIoT) involves several steps. Firstly, the relationships between data elements are identified using clustering using the K-Means algorithm. This step is crucial for efficient data transmission and storage as it groups related data elements into clusters, reducing the amount of data that needs to be transmitted to the cloud. Next, cluster heads are selected based on their ability to maintain relationships between data elements and ensure data consistency within the cluster using the Decision tree algorithm. The selected cluster heads act as intermediaries between the SIoT devices and the cloud and only transmit the compressed data to reduce energy consumption and bandwidth usage. The compressed data is then encoded using Huffman coding to further reduce its size before transmission to the cloud. In the cloud, the compressed data is decoded to retrieve the original data.

2.1 Algorithm

In this section, the algorithm used to implement the proposed model is presented. It makes use of the equations defined in the proposed model to train the model.

Algorithm: Clustering and D_{Agg} Calculation
Input: Features – A, D_A, R, R_F, R_c
Output: D_{Agg}
Step 1: Compute Features() // Compute features for clustering using Equation (4).
Step 2: Calculate CH() // Calculate the CH in step 1.
Step 3: Evaluate Clusters() // Evaluate each cluster using Equation (5).
Step 4: Calculate Entropy() // Obtain the Entropy E to gain the information G using Equation (6).

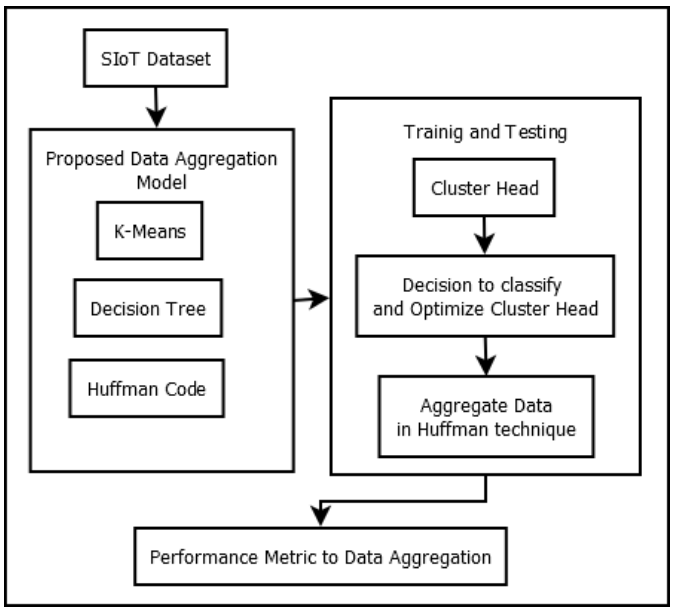

Fig 1. Proposed Model Design

Step 5: Compress Data() // Compress the data using Equation (7).
Step 6: Choose Compressed Data() // Choose the compressed data from step 5.
Step 7: Calculate DAgg() // Calculate the D_{Agg} using Equation (8).
Step 8: Send To DeviceD() // Send D_{Agg} to the corresponding device D.

Input features A, D_A, R, R_F, and R_c are utilized to generate the output D_{Agg}. The clustering process Cluster Head (CH) is calculated in step 2 to assess the quality of the clusters formed in step 1. Step 3 evaluates each cluster using Equation (5), and step 4 obtains the entropy (E) to gain information (G) through Equation (6). In step 6, data compression is performed using Equation (7), and the resulting compressed data is chosen in step 7. The D_{Agg} score is computed in step 7 using Equation (8), and finally, the D_{Agg} value is transmitted to the corresponding device D in step 8. The method employed for data aggregation, utilizing clustering techniques, contributes significantly to the enhancement of SIoT performance. This involves the application of machine learning approaches to cluster the SIoT devices based on relevant features.

3 Result and Discussion

The proposed Cluster-based Social Internet of Things using the K-Means algorithm was implemented and tested on a dataset consisting of social IoT devices, and the features used in the dataset are Device Id, User Id, Manufacture ID, Brand ID, Latitude and Longitude, Device class, Device Mobility, Application, Service, Protocol, and some of the other attributes that are necessary for experimentation: Data holds the device data that has been generated by the device, Memory is the total storage capacity of a device and Data Transfer rate that signifies the rate at which a device can transfer the data. Table 1 shows the result of object profiling. It consists of features such as Id of the owner (Owner id), the name of the device (Device Name), the brand of the device (Device Brand), where the device is located (Location), the landmark (Landmark), the temperature of the location (Temperature) and pressure value (Pressure), Humidity value of the location (Humidity), Point of interest, the status of the weather (Weather Status), the present status of people(People Presence), People Present Status, Light Status, people's backward movement (Backward Movement), people's forward movement (Forward Movement), Internet connection (Connection Type), Connection device brand (Device Brand), Connection Device Type (Device Type), distance between device and connection(Distance), protocol type (protocol). The algorithm performance was evaluated in terms of cluster purity, the average distance to the centroid, processing time, and scalability. Simulation Parameter Specification: CPU: 7th Gen Intel (2.40 GHz), RAM: 4 GB, operating system: Windows, Mac, Programming: Python 3.7, scipy: 1.6.3, pandas: 1.2.3, numpy: 1.20.1, matplotlib:3.4.1, PyQt5: 5.15.4, AI libraries: joblib 1.0.1, scikit-learn: 0.24.2, sko: 0.5.7

In the context of SIoT, silhouette scores for the number of clusters, Number of the distance between Train and Test K-Means, BIC Scores for the number of clusters, and Gradient of BIC scores for the number of clusters equations are used to evaluate the

Table 1. Object Profiling

Index	People Presence	People Present Status	Light Status	Backward Movement	Forward Movement	Connection Type	Device Brand	Device Type	Distance	Protocol
0	No	Nobody	On	-47	159	0	1	1	30	2
1	Yes	Alone	On	-152	33	1	1	0	60	1
2	NO	Nobody	On	-60	20	1	0	1	30	1
3	No	Pair	Off	-104	197	1	1	1	80	1
4	Yes	Group	On	-63	165	0	1	0	30	0

performance of the clustering algorithm. The Silhouette score is used to find the cluster's optimal number by testing different values of k and selecting which has the highest score. The distance between the train and test K-Means is used to evaluate the generalization ability of the model. The BIC score is used for comparison of the goodness of fit of different models with different numbers of clusters. The optimal number of clusters is found by calculating the gradient of BIC scores with respect to the number of clusters and selecting the k value that maximizes the gradient. Overall, these equations provide useful metrics for evaluating the performance of clustering algorithms in the Social Internet of Things.

3.1 Evaluation metrics

Silhouette Score: The silhouette score measures how well each data point fits into its assigned cluster. It is calculated as follows:

$$silhouette\ score\ =\ (b(i)\ -\ a(i))\ /\ max\{a(i),\ b(i)\} \tag{12}$$

where a(i) is the average distance from data point i to all other data points within the same cluster, and b(i) is the minimum average distance from data point i to all data points in any other cluster.

Number of Distance between Train and Test K-Means: The distance between train and test K-Means is calculated as the log-likelihood ratio of the two models. It is expressed as follows:

distance = log likelihood of the test data under the train model - log likelihood of the test data under the test model

BIC Score: The Bayesian Information Criterion (BIC) score is a measure of the goodness of fit of a statistical model. It is calculated as follows:

$$BIC\ =\ -2\ *\ \log\ likelihood\ +\ p\ *\ \log(n) \tag{13}$$

where log likelihood is the logarithm of the maximum likelihood of the model, p is the number of parameters in the model, and n is the number of data points.

Gradient of BIC Scores: The gradient of BIC scores with respect to the number of clusters can be used to find the optimal number of clusters. It is calculated as follows:

$$d(BIC)/d(k)\ =\ -2\ *\ \log\ likelihood\ +\ p\ *\ \log(n)\ -\ p\_k\ *\ \log(n) \tag{14}$$

where p_k is the number of parameters for the k-th model, and k is the number of clusters.

3.2 Results

The above graph (Figure 2) represents the Three Principal Component Cluster which consists of values P1 and P2

Figure 3 shows the Silhouette scores for the number of clusters formed using the proposed model. It can be noticed that the score lies between 0.36 and 0.44 indicating that the object matches with its own cluster in the proposed model.

Figure 4 shows the distance variation between the clusters during training and testing of the K-Means. It can be observed that the number of clusters increases as the distance between the objects increases. This is due to the effect of object availability within the proposed model.

$$BIC\ =\ -2*\ LL\ +\ \log(N)\ *\ K \tag{15}$$

Figure 5 shows the BIC scores obtained for the number of clusters. The BIC scores are calculated using the Equation (15).

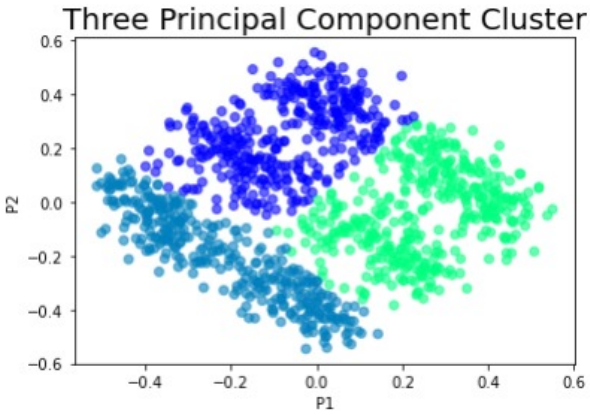

Fig 2. Three Principal Component Cluster

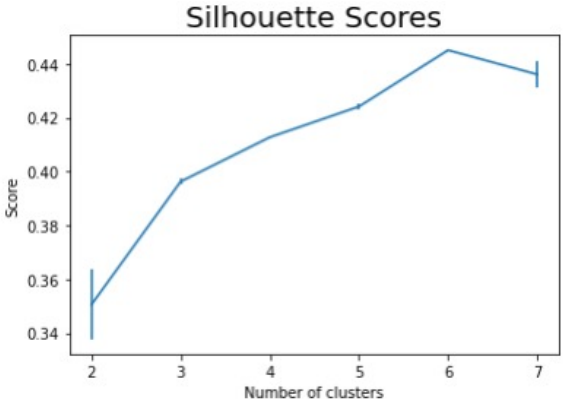

Fig 3. SILHOUETTE scores for the number of clusters

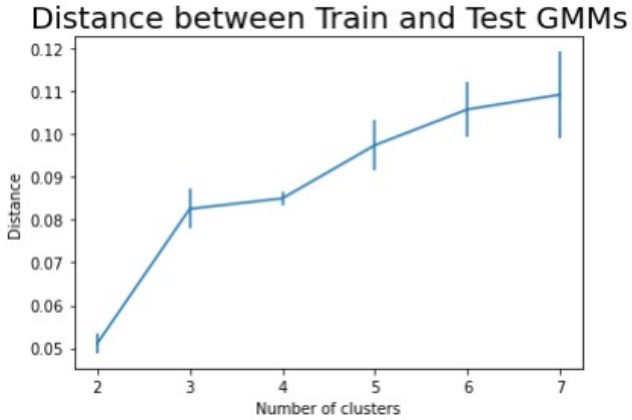

Fig 4. Number of distance between Train and Test K-Means

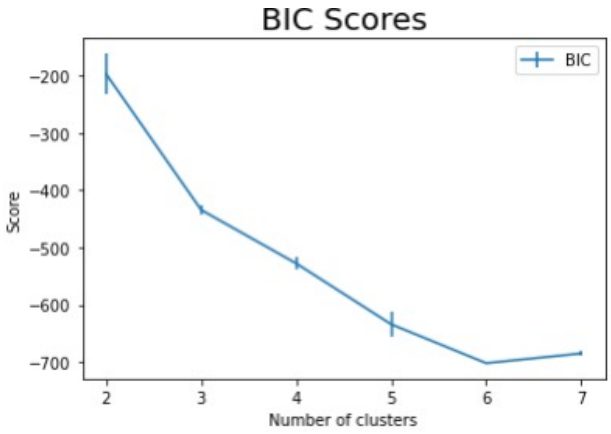

Fig 5. BIC Scores for the number of clusters

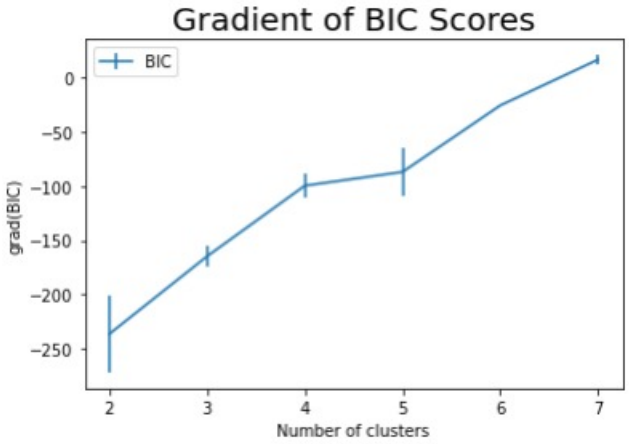

Fig 6. Gradient of BIC scores for the number of clusters

The gradient of the BIC scores for the number of clusters is formed and is shown in Figure 6. From the graph, it is evident that the proposed model is best as it lies between 5 and 7 when clusters are increased.

The result of cluster head selection for each cluster, each cluster has one CH, and the cluster members have a valid relationship with the CH. Samples of CH and relationships between the devices selections are as follows:

{1: [8, 10, 15, 17, 30]}
{9: [11, 14, 20, 21, 23, 24, 27, 28]}
{3: [4, 5, 13, 16, 25]} {18: [22, 26]}
{2: [6, 7, 12, 19, 29]}
{1: [8, 10, 15, 17, 30],
9: [11, 14, 20, 21, 23, 24, 27, 28],
3: [4, 5, 13, 16, 25], 18: [22, 26],
2: [6, 7, 12, 19, 29]}
Only Cluster Heads [1, 9, 3, 18, 2]

The result of data compression denotes the original size and the compressed size of the device data. The sample of data aggregation is like data sent from Device 8 to CH 1 Wait: 0.2

The size of the Original data is: 47
The size of the encoded data is: 18
[[1,'0101110101001011100010010111000110000
0111101111110000001110110001010010 0001000101000000111010000101000100011101 1

011010110001001100001011101010010111 1000100101110001100000011110111111100001010010000010001010000001110100001010001000111011011010110000001001001000111 000000111101111110101 1’],
[8,’1000100001100100011001011101’],
[10,’1000110001100100011001010101010101011100011
000110010001100101001011000110001100 010001100101001011000110001100100011001010010110001100010010100101001011 0001100011001000110010100101’],[15,1010110010000100010000101001111101’]

To compare the results of the proposed approach with existing works, a simulation study was conducted using the CCNSim simulator[23]. The performance of the proposed approach was compared with two existing works: a relationship-based clustering approach without data compression, and a data compression-enabled cluster-based aggregation model without relationship-based clustering. The simulation results showed that the proposed approach achieved a significant reduction in data transmission and storage compared to the existing approaches. Table 2 shows the comparison of the results obtained with the Proposed Relationship cluster head selection and Data Compression cluster-based aggregation model, Relationship-based clustering approach without data compression, and Data compression cluster-based aggregation model without relationship-based clustering. When comparing BIC scores, a lower absolute value indicates a better fit. Therefore, in this case, the BIC score of -5227.080 is better than the other two scores. BIC score of -5227.080: This score indicates a good balance between model fit and complexity. It suggests that the clustering model with this score is a good fit for the data and is not overly complex. BIC score of -7264.630: This score is higher than the previous score, indicating that the clustering model is less well-fitting and more complex than the model with a BIC score of -5227.080.

BIC score of -8867.340: This score is the highest of the three, indicating that the clustering model is the least well-fitting and most complex of the three models. The BIC score of -5227.080 is the best of the three and indicates the best trade-off between model fit and complexity. However, it is worth noting that the optimal BIC score depends on the dataset and proposed clustering algorithm being used.

Table 2. Result comparison

Evaluation Metrics	Proposed Relationship cluster head selection and Data Compression enabled cluster based aggregation model	Relationship-based clustering approach without data compression	Data compression enabled cluster-based aggregation model without relationship-based clustering
Silhouette Score	0.459	0.626	0.58
Number of Distance between Train and Test K-Means	-1148.951	-203.345	-467.89
BIC Score	-5227.08	-726.08	-8981.786
Gradient of BIC Scores	-96.445	-61.415	-94.244

4 Conclusion

The proposed approach of Relationship-cluster head selection and Data compression cluster-based aggregation model for SIoT using Huffman coding is a novel and effective solution for efficient data transmission and storage in SIoT applications. The integration of relationship-based clustering and Huffman coding for data compression, as well as the selection of cluster heads based on their ability to maintain relationships between data elements and ensure data consistency within the cluster, provides a comprehensive solution for SIoT applications. The simulation results demonstrate that the proposed approach achieves significant reductions in energy consumption and bandwidth usage compared to existing approaches. The strength of this research lies in its innovative approach that combines several techniques for efficient data transmission and storage in SIoT. It provides a comprehensive solution that takes into account the selection of cluster heads, data consistency, and data compression, which are important considerations in SIoT. The weaknesses of this research include the assumptions made during the simulation study, which may not accurately reflect real-world scenarios. Additionally, further research is needed to evaluate the proposed approach in different SIoT applications. Thus, future research focuses on extending the proposed approach to consider dynamic cluster head selection and adaptive data compression techniques. Also, the proposed approach will be extended to incorporate machine learning algorithms for intelligent cluster head selection and dynamic compression rate adjustment, which can further improve the efficiency of SIoT applications.

5 Acknowledgement

This work was carried out under the "Development program of ETU "LETI" within the framework of the program of strategic academic leadership" priority-20230.

References

1) Redhu S, Hegde RM. Multi-Sensor Data Fusion for Cluster-based Data Aggregation in IoT Applications. In: 2019 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), 16-19 December 2019, Goa, India. IEEE. 2020. Available from: https://ieeexplore.ieee.org/document/9117970.
2) Ochoa-Zambrano J, Garbajosa J. Social Internet of Things: Architectural Approaches and Challenges. 2020. Available from: https://doi.org/10.48550/arXiv.2002.04566.
3) Roopa MS, Pattar S, Buyya R, Venugopal KR, Iyengar SS, Patnaik LM. Social internet of things (siot): Foundations, thrust areas, systematic review and future directions. *Computer Communications*. 2019;139:32–57. Available from: https://doi.org/10.1016/j.comcom.2019.03.009.
4) Prakash SPS, Nagabhushan TN, Krinkin K. Congestion Avoidance and Delay Minimization in Energy Aware Routing of Dynamic ieee 802.11s WMN: Wireless Mesh Networks Under Mobility Conditions. 2020. Available from: https://doi.org/10.4018/978-1-7998-2460-2.ch010.
5) Aljubairy A, Zhang WE, Sheng QZ, Alhazmi A. SIoTPredict: A Framework for Predicting Relationships in the Social Internet of Things. In: International Conference on Advanced Information Systems Engineering, CAiSE 2020: Advanced Information Systems Engineering;vol. 12127 of Lecture Notes in Computer Science. Springer, Cham. 2020;p. 101–116. Available from: https://doi.org/10.1007/978-3-030-49435-3_7.
6) Puranikmath VI, Harakannanavar SS, Kumar S, Torse D. Comprehensive Study of Data Aggregation Models, Challenges and Security Issues in Wireless Sensor Networks. *International Journal of Computer Network and Information Security*. 2019;11(3):30–39. Available from: https://www.mecs-press.org/ijcnis/ijcnis-v11-n3/IJCNIS-V11-N3-5.pdf.
7) Bongale AM, Nirmala CR, Bongale AM. Energy efficient intracluster data aggregation technique for wireless sensor network. *International Journal of Information Technology*. 2022;14(1):827–835. Available from: https://doi.org/10.1007/s41870-020-00419-7.
8) Dehkordi SA, Farajzadeh K, Rezazadeh J, Farahbakhsh R, Sandrasegaran K, Dehkordi MA. A survey on data aggregation techniques in IoT sensor networks. *Wireless Networks*. 2020;26(2):1243–1263. Available from: https://doi.org/10.1007/s11276-019-02142-z.
9) Jan SRU, Khan R, Jan MA. Energy efficient data aggregation approach for cluster-based wireless sensor networks. *Annals of Telecommunications*. 2021;76:321–329. Available from: https://doi.org/10.1007/s12243-020-00823-x.
10) Kaur M, Munjal A. Data aggregation algorithms for wireless sensor network: A review. *Ad Hoc Networks*. 2020;100:102083. Available from: https://doi.org/10.1016/j.adhoc.2020.102083.
11) Bhindu KM, Yogesh P. An energy efficient cluster-based data aggregation in wireless sensor network. In: 2019 11th International Conference on Advanced Computing (ICoAC), 18-20 December 2019, Chennai, India. IEEE. 2020. Available from: https://ieeexplore.ieee.org/document/9087289.
12) Yao X, Wang J, Shen M, Kong H, Ning H. An improved clustering algorithm and its application in IoT data analysis. *Computer Networks*. 2019;159:63–72. Available from: https://doi.org/10.1016/j.comnet.2019.04.022.
13) Khan AR, Chishti MA. Data Aggregation Mechanisms in the Internet of Things: A Study, Qualitative and Quantitative Analysis. *International Journal of Computing and Digital Systems*. 2020;2(9):289–297. Available from: http://dx.doi.org/10.12785/ijcds/090214.
14) Zhang G, Li Y, Deng X. K-Means Clustering-Based Electrical Equipment Identification for Smart Building Application. *Information*. 2020;11(1):1–18. Available from: https://doi.org/10.3390/info11010027.
15) Khan AR, Chishti MA. Data Aggregation Mechanism in the Internet of Things: A Study, Qualitative and Quantitative Analysis. *International Journal of Computing and Digital Systems* . 2020;9(2):289–297. Available from: http://dx.doi.org/10.12785/ijcds/090214.
16) Abhijith HV, Babu HSR. Intelligent Data Aggregation Framework for Resource Constrained Remote Internet of Things Applications. *International Journal of Advanced Computer Science and Applications*. 2021;12(5):146–151. Available from: https://doi.org/10.14569/IJACSA.2021.0120518.
17) Yousefi S, Karimipour H, Derakhshan F. Data Aggregation Mechanisms on the Internet of Things: A Systematic Literature Review. *Internet of Things*. 2021;15:100427. Available from: https://doi.org/10.1016/j.iot.2021.100427.
18) Sirajudheen M, Rekh AS. An overview on data aggregation in IOT for wireless sensor network. *Turkish Journal of Computer and Mathematics Education*. 2021;12(12):3390–3395. Available from: https://turcomat.org/index.php/turkbilmat/article/view/8060/6309.
19) Awad FH, Hamad MM. Improved k-Means Clustering Algorithm for Big Data Based on Distributed Smartphone Neural Engine Processor. *Electronics*. 2022;11(6):1–20. Available from: https://doi.org/10.3390/electronics11060883.
20) Mohana SD, Prakash SPS, Krinkin K. Service Oriented R-ANN Knowledge Model for Social Internet of Things. *Big Data and Cognitive Computing*. 2022;6(1):1–23. Available from: https://doi.org/10.3390/bdcc6010032.
21) Mohammadi SO, Kalhor A, Bodaghi H. K-Splits: Improved K Means Clustering Algorithm to Automatically Detect the Number of Clusters. In: Computer Networks, Big Data and IoT;vol. 117 of Lecture Notes on Data Engineering and Communications Technologies. Singapore. Springer. 2022;p. 197–213. Available from: https://doi.org/10.1007/978-981-19-0898-9_15.
22) Gurumoorthy S, Subhash P, de Prado RP, Wozniak M. Optimal Cluster Head Selection in WSN with Convolutional Neural Network-Based Energy Level Prediction. *Sensors*. 2022;22(24):1–23. Available from: https://doi.org/10.3390/s22249921.
23) Kathiroli P, Selvadurai K. Energy efficient cluster head selection using improved Sparrow Search Algorithm in Wireless Sensor Networks. *Journal of King Saud University - Computer and Information Sciences*. 2022;34(10, Part A):8564–8575. Available from: https://doi.org/10.1016/j.jksuci.2021.08.031.
24) Mohana SD, Prakash SPS, Krinkin K. CCNSim : An artificial intelligence enabled classification, clustering and navigation simulator for Social Internet of Things. *Engineering Applications of Artificial Intelligence*. 2023;119:105745. Available from: https://doi.org/10.1016/j.engappai.2022.105745.
25) Vani S Badiger, Ganashree TS. Energy efficient new data aggregation method for IoT. *Journal of Critical Thinking*. 2020;7(4):2660–2670. Available from: https://www.jcreview.com/admin/Uploads/Files/61b8506b76d879.18759858.pdf.
26) Farooq O, Singh P, Hedabou M, Boulila W, Benjdira B. Machine Learning Analytic-Based Two-Staged Data Management Framework for Internet of Things. *Sensors*. 2023;23(5):1–29. Available from: https://doi.org/10.3390/s23052427.
27) Han B, Ran F, Li J, Yan L, Shen H, Li A. A Novel Adaptive Cluster Based Routing Protocol for Energy-Harvesting Wireless Sensor Networks. *Sensors*. 2022;22(4):1–16. Available from: https://doi.org/10.3390/s22041564.