An Efficient Way for Data Mining via Overlay-based Networking for Enhanced Service

Pushpanjali
M.Tech, CSE, BMSCE, Bangalore, India

Jyothi S. Nayak
Associate Prof., Dept.of CSE, BMSCE, Bangalore, India

ABSTRACT
Big data generated from various aspects like online transactions, social websites, logs and search queries is increasing rapidly and thus the demand for data mining has risen as a noteworthy zone. An overlay- based parallel information mining executes completely dispersed information administration and handles processing by utilizing the overlay system, which can achieve high flexibility. The talk incorporates a survey of best in class systems and stages for preparing and overseeing huge information and also the endeavours expected on enormous information mining. Nonetheless, the overlay - based parallel mining structural planning is not fit for achieving data mining administrations if there is an occurrence of the physical system disturbance that is created due to switch/correspondence line breakdowns on the grounds that various hubs are expelled from the overlay system. To get the estimated arrangement and better results, the proposed framework utilizes K-medoids algorithm for cluster formation and overlay based system. Proposed work gives enhancement in terms Energy Consumption in data gathering, reduced delay and Node Coverage.

Keywords
Parallel Data Mining, Big Data, Overlay-based, versatility, Kmeans, K-medoids, physical network disruption.

1. INTRODUCTION
Big data is a large volume of complex data sets that are generated dynamically from different sources. Processing or management of such data becomes complex using traditional technologies. Data generated from sensor networks is large in volume. Wireless sensor nodes monitor and poll various sensor data, which is then transmitted to a host server for further analysis and management. The data collected in a host is not only stored safely for archiving but it will also be analysed to generate abstracted patterns or to associate meaning patterns for decision assistance. Data management and processing for wireless sensor networks (WSNs) has become active research in several areas of computer science engineering such as the distributed systems [1] [2].

Data Mining is an investigative procedure intended to investigate information (business sector related information).The goal of data mining process is to find patterns, once these patterns are found they can further be used to make certain decisions for development of their business models. The objective of information mining is expectation and predictive information mining is the most well-known sort of information mining and one that has the most direct business applications. The procedure of information mining comprises of three stages: (1) the starting investigation, (2) model building or example distinguishing proof with acceptance/check, and (3) organization of the data.

Stage 1: Exploration. This stage ordinarily begins with information arrangement which may include cleaning information, information transformation, selecting subsets of data records and performing some preliminary feature selection determination operations in case of information sets with vast quantities of variables ("fields"), to convey the quantity of variables to a manageable range. This first phase of the procedure of information mining may include anyplace between a basic decision of direct predictors for a regression model, to expand exploratory investigations utilizing a wide mixed bag of graphical and statistical techniques. Keeping in mind that the end goal is to recognize the most important and relevant variables and nature of data based on problem.

Stage 2: Model building and validation. This stage includes considering different models and picking the best one in view of their predictive performance (i.e., clarifying the variability being referred to and delivering stable results crosswise over examples). This may sound like a basic operation, yet actually, in some cases it includes an exceptionally complicated process. There is a mixed bag of systems created to accomplish objective that is, applying diverse models to the same information set and afterward contrasting their execution with pick the best. These strategies, which are frequently viewed as the centre of prescient information mining include: Bagging (Voting, Averaging), Boosting, Stacking and Meta-Learning.

Stage 3: Deployment. That last stage includes utilizing the model chosen as best in the past stage and applying it to new information, keeping in mind that the end goal is to create or predict expected outcome of the normal result.

The idea of Data Mining is turning out to be progressively main stream as a business data administration device where it is required to uncover learning structures. As of late, there has been expanded enthusiasm for growing new diagnostic systems particularly intended to deliver the issues applicable to business Data Mining (e.g., Classification Trees), yet Data Mining is still in view of the reasonable standards of insights including the conventional Exploratory Data Analysis (EDA)[3].

Domain experts collect, describe, and explore the data and they also recognize quality problems of the data as shown in figure 1. In the data exploration phase, the traditional data analysis methods such as statistics are used for preparing the data by selecting tables, tuples and attributes. The meaning of the data remains same. During modeling phase, Data mining experts recognize and apply various mining functions because different mining algorithms for the same type of data mining problem produces various outcomes. The data mining experts must evaluate each model. In this phase a frequent interchange with the domain experts from the data preparation stage is crucial. The modeling phase and the evaluation phase are run together. They can be repeated number of times to change parameters until optimal values is accomplished. Once the final phase of modeling is done, a model of high quality has been constructed. In Evaluation phase, Data mining experts estimate this model, if the model does not fulfill their expectations, they rebuild the

model by going back to the modeling phase and changing its parameters until optimal values are accomplished.

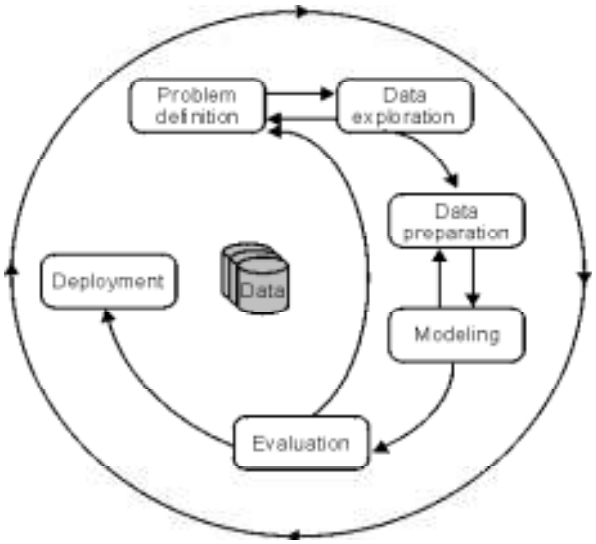

Figure.1 Basic flow of the Data Mining Process [4]

Lastly Deployment, Data mining experts uses the mining outcomes by exporting the results into database tables or into other applications like spreadsheets. Intelligent Mine (IM) Modeling helps to select the input data, exploring the data, transforming the data, and mining the data. With IM Visualization, can display the data mining results to analyze and interpret them [7].

Data extraction can be done effectively using neighbor node selection scheme [14]. Rest of the paper organized as follows: In Section 2, this paper introduces basic Data Mining Algorithms and Techniques. In Section 3, Data Gathering Process in wireless Sensor Networks. In Section 4, study on comparison of kmeans and kmedoid algorithm. The proposed algorithm kmedoid for our envisioned overlay-based parallel data mining architecture is discussed, as well as the experimental works. Conclusions and feature work are presented in Section 5.

2. DATA MINING ALGORITHMS AND TECHNIQUES

A *data mining algorithm* is a collection of heuristics and calculations that create a model from the data. The mining model that an algorithm produces from the data can take various customs. Choosing the best algorithm to use it for a particular analytical task can be a challenging. Different algorithms produce different results for same task and some algorithms can yield more than one type of results. Some of the data mining techniques are,

2.1 Classification

Classification is the commonly used data mining technique, which makes use of group of pre-classified examples to develop a model which can categorize the population of records at huge. Fraud detection and credit risk tenders are well suited to this type of analysis. This approach applies decision tree or neural network-based classification algorithms frequently. The data classification process includes learning and classification. During learning phase, the training data sets are analyzed by classification algorithm and during classification phase, the test data are used to evaluate the accuracy of the classification rules. If the accurateness is acceptable the rules can be applied to the new data records. For example, a fraud detection application would contain complete histories of both fraudulent and valid actions determined on a record-by-record basis. The classifier-training algorithm makes use of these pre-classified examples to determine the set of parameters those are essential for proper judgment and encoding these parameters into a model called a classifier.

Types of classification models:

- Decision tree induction based classification
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Associations

2.2 Clustering

Clustering is the identification of related classes of objects. A group of clusters that describe how the circumstances in a dataset are related. By using these techniques these can further be classified as dense and parse regions in object space and can discover whole distribution pattern and correlations among data attributes. Classification approaches can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example consider, to form group of customers based on purchasing patterns, to categories genes with comparable functionality.

Different clustering methods

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

2.3 Predication

Regression technique can be adapted for predication. Regression analysis can be used to form the relationship between one or more independent and dependent variables. In data mining process, independent variables are attributes which are already known and response variables are what need to calculate. Unfortunately, several real-world problems are not easy for prediction. For instance, sales volumes, stock prices, and product failure rates are all very tough to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression or decision trees) may be essential to forecast future values. The similar data model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks can also be used to create both classification and regression models.

Different regression methods

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression.

2.4 Association rule

Association and correlation is usually used to find frequent item set among large data sets. These type of findings help

businesses to make certain decisions, such as catalogue design, customer shopping behavior analysis etc. The number of possible Association Rules for a given dataset is generally Very large and a high proportion of the rules are usually of little value.

Different association rules

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

2.5 Neural networks

Neural network is a set of connected input/output units, each connection has a weight present with it. During the learning phase of this, network learns by adjusting weights so as to be able to predict the correct class labels for the input tuples. Neural networks have the amazing capability to derive meaning from complicated or imprecise data and this can be used to extract patterns and identify trends that are too difficult to be observed by either humans or other computer systems. These networks are well suitable for continuous valued inputs and outputs. For example consider, handwritten character reformation and many real world business problems. Neural networks are best in identifying patterns or trends in data and well suited for prediction or forecasting needs [5] [6].

Different neural networks

- Back Propagation
- Radial Basis Function Networks
- Wavelet Neural Networks

3. DATA GATHERING PROCESS IN WIRELESS SENSOR NETWORKS

A multidisciplinary research zone such as wireless sensor networks [8, 9, 10, 11, 12] where close cooperation between application users, application domain professionals, hardware developers and software developers is necessary to implement efficient systems. Wireless Sensor Network generates a huge amount of data that has to be aggregated at various levels Wireless sensor networks consist of small nodes which exhibit the properties like sensing, computation, and communication capabilities. A sensor network is a collection of sensor nodes cumulated in an ad-hoc manner. Wireless sensor networks have restricted computational power, memory and battery power, which leads to increased complication for application developers. In WSN data gathering is an effective way to save the limited resources. The main goal is gathering data in an energy efficient manner so that network lifetime is enhanced [13].

The parameter being used to control the overall energy and battery power are rationalized to provide best conceivable solution. Accuracy is taken into consideration as distance varies from cluster to cluster but can be used to provide a variety of research application, mobile communication and location tracking system. Bandwidth, memory, signal strength, time, battery power etc. have been utilized to study the performance of a sensor network, its efficiency can be improved by reducing the cost of cluster development. Sensor nodes are beneficial in disaster, war zone and several new technologies like mobile technology, laser technology etc. where the data has to be transferred accurately and in a fraction of time where each node is responsible for the extraction and transfer of data such that the data to be exchanged cannot be lost on its way to the receiver.

Overlay-based parallel data mining is one of architectures that improve the service availability against server break-downs. The architecture shown in fig.2, all the servers execute both management and processing functions and it also shows mapping and reduction processes in the overlay-based parallel data mining. The overlay network is constructed by all servers and utilized to find processing nodes, similar to the master nodes in the conventional architecture. This architecture will provide the service even if some nodes are removed from the overlay network [14].

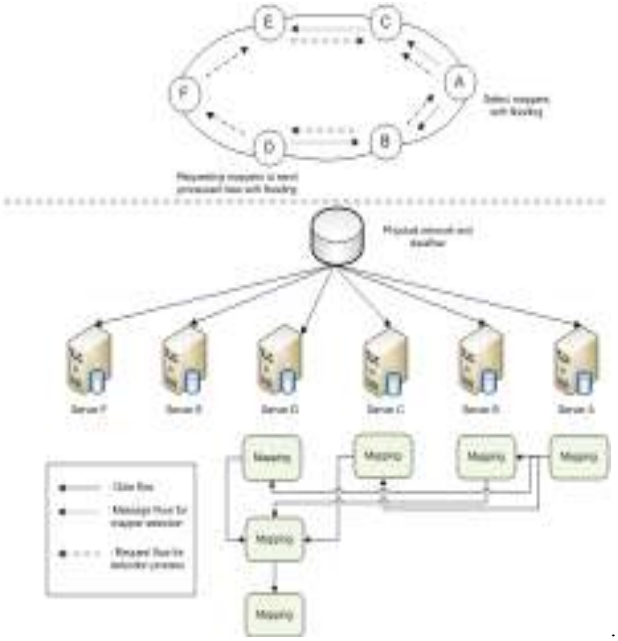

Figure.2 Mapping and reduction processes in an overlay-based parallel data mining architecture

When a data processing request is injected, a node that received the request (node A in the Figure. 2) executes a reception function by using the overlay network [14, 15]. In other words, the node finds mappers by using flooding message, where mappers are randomly selected (nodes B, C, and D in the Fig. 2). The mapper that finished the mapping process initially (node D in the Fig. 2) becomes a reducer, and it requests to other mappers to transmit the processed data to itself, where the request message can be forwarded by using flooding scheme. After receiving the processed data from mappers, the reducer starts executing the reduction process and outputs the analysed result.

In this architecture, since the connectivity of overlay network dramatically affects the service availability of data mining, there are numerous works, which tackled the connectivity issue from the sundry viewpoints, i.e., context- cognizant, graph theory predicated, and intricate network theory predicated overlay network construction schemes [16]. These works make overlay networks that are tolerant to minute-scale server breakdowns but do not consider the sizably voluminous- scale server breakdowns, i.e., physical network disruption. Therefore, this paper develops an overlay-predicated parallel data mining architecture that is tolerant to physical network disruption so that data mining is available at anytime, anywhere [17].

4. PROPOSED METHOD

In this paper, mainly focused on efficient big data mining techniques from heterogeneous wireless sensor networks. When dealing with the heterogeneous sensor nodes, there will be a

problem of data inaccuracy. To overcome all the problems, we implemented effective cluster based technique.

4.1 System Architecture

Figure 3 depicts the block diagram of proposed system. The proposed architecture capable of providing service in case of route/node failures by making use of overlay-based network. The architecture consists of following modules,

4.1.1 Network Initialisation

In network initialization, a network terrain with specific area and finite number of sensor nodes is created. The parameters like Received Signal Strength Indication (RSSI), Time to live (TTL), Multicast Routing Information cost (MRIC), bandwidth, battery consumption have been used to determine the number of nodes that would be considered in a cluster.

4.1.2 Cluster formation

For cluster formation we have used k-medoids clustering algorithm. K-medoids method uses medoids to represent the cluster instead of using centroid. A medoid is the most centrally positioned data object of a cluster, which is selected as cluster head (CH). Here, k data objects are selected randomly as medoids to represent k cluster and remaining all data objects are placed in a cluster having medoid nearest (or most similar) to that data object. After dealing out all data objects, new medoid is calculated which can represent cluster in an improved way and the entire process is repeated and all these data objects are bound to the clusters based on the new medoids. In every iteration, medoids change their position, this process is continued until no any medoid remained to move. As outcome, *k* clusters are found representing a set of data objects.Our method of constructing clusters in distributed fashion performs over traditional LEACH-like algorithms in terms of traffic balance and execution cost, thereby leading to a longer system lifetime.

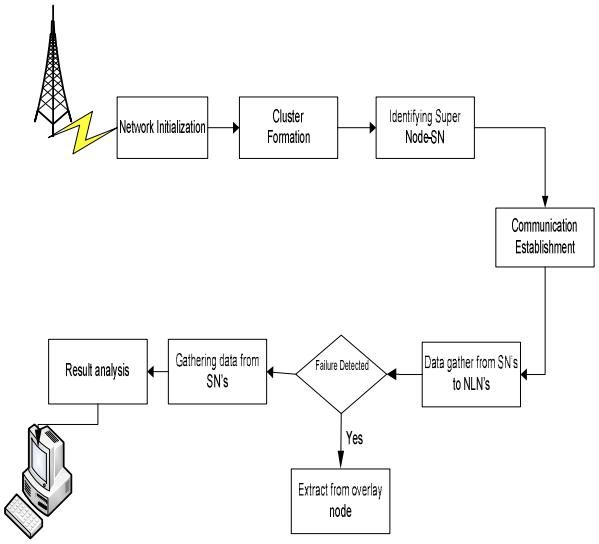

Figure.3 Secure and efficient way for big data gathering in densely distributed WSN

4.1.3 Communication establishment

In this, each cluster head is having communication among nodes of the same cluster and cluster head or super nodes of other clusters. CH will be responsible for administration of all other nodes inside respective cluster and collecting the data from the nodes inside the cluster and transferring the data to the neighboring cluster head for further information exchange.

4.1.4 Gathering data from Super Nodes

User can request to respected group head that is SN's, for extracting the useful information or overlay node if in case of one SN's fails. It means we will use overlay based concept to avoid breakage of entire network that leads to increased delay.

4.2 Comparison of k-means and k-medoid:

4.2.1 Kmeans

K-means is the simplest clustering method comes under unsupervised learning algorithms. It follows a simple and easy way to classify a given database D of n objects into a set of clusters (i.e. k clusters). The input to this algorithm is k and task is to partition a set of n objects into k clusters so that the resulting intra cluster similarity is high but the inter cluster similarity is low. K-means algorithm required to find the centroids, where the coordinate of the centroid is the means of the coordinates of the data objects within the cluster and assigns every object to the nearest centroid. Since the original database may have tens of thousands of records, such calculations will be slow [21].

The algorithm has following steps:
Input : '*k*', the number of clusters to be partitioned; '*n*', the number of objects.
Output*:* A set of '*k*' clusters based on given similarity function.

- We place k points into the space represented by the objects that are being clustered. Initial group centroids are represented by these points.
- Assign each object to the group that has the closest centroid.
- After the assignment of all objects, recalculate the positions of the k centroids.
- Repeat Steps 2 and 3 until the centroids no longer move.

Limitations

- K-means clustering is sensitive to presence of outliers and a set of objects closest to a centroid may be empty, in which case centroids cannot be updated.
- Appropriate only when the mean of a cluster is defined; not applicable to categorical data.
- Unable to handle noisy data and outliers.

4.2.2 Proposed K-medoids algorithm

K-medoids clustering algorithm, where medoids are considered instead of centroids. It is less sensitive to outliers compared with the K-means clustering. A medoid is the most centrally located data object in a cluster. Initially k data objects are selected randomly as medoids, which are used to represent k clusters and remaining all data objects are placed in a cluster having medoid nearest to that data object. After processing all data objects, new medoid is determined which can represent cluster in a better way and the entire process is repeated. Again all data objects are placed into the clusters based on the new medoids. In every iteration, medoids change their location step by step. This process is continued until no any medoid move. As a result, *k* clusters are found representing a set of n data objects [22].

The proposed algorithm calculates the distance matrix once and uses it for finding new medoids at every iterative step. Rather than calculating the mean of the items in each cluster, a medoid is selected for each cluster at each iteration. These Medoids for each cluster are calculated by finding object *i* within the cluster that minimizes

$$\sum_{j \in C_i} d(i,j)$$

Where *Ci* is the cluster containing data object *i* and *d* (*i; j*) is the distance between objects *i* and *j*.
The k-medoids algorithm can be summarized as follows:

1. Choose *k of n data* objects at random to be the initial cluster medoids
2. Assign each data object to the cluster associated with the nearby medoid based on distance metric i.e. Euclidean distance metric.

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{i.k} - x_{j.k})^2}$$

3. Recalculate position of new medoid of each cluster, which is the object minimizing the total distance to other objects in its cluster.
4. Repeat the Steps 2 and 3 until all the medoids become fixed.

Strengths

- The algorithm has outstanding feature that it requires the distance between every pairs of objects only once and uses this distance at every iterative step.
- More robust compared to k-means clustering algorithm in the occurrence of noise and outliers.
- A medoid rule helps to usefully describe the cluster, which is less subjective to outliers and other extreme values than a mean.
- Kmedoids is effective algorithm compared to kmeans, as it finds the more accurate centers of the clusters, which can results in better communication between wireless sensor nodes, reduced energy consumption and lower packet delay. This leads to better maintainability of the system.

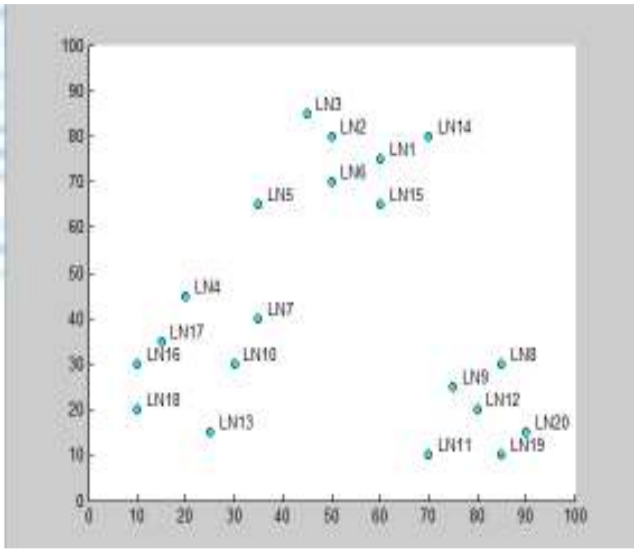

Figure 4. Initial nodes on network area of 100x100

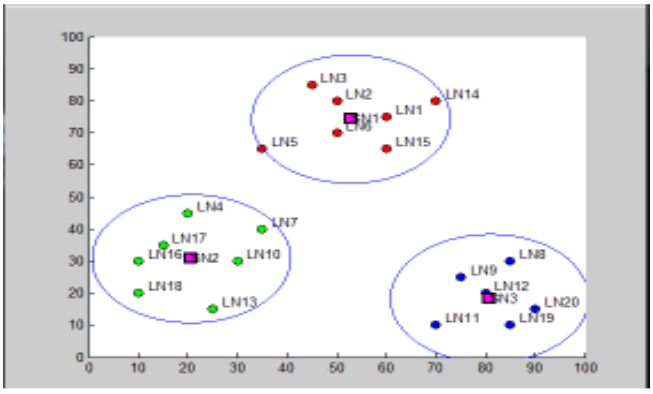

Figure 5. Super node selection based on kmeans algorithm

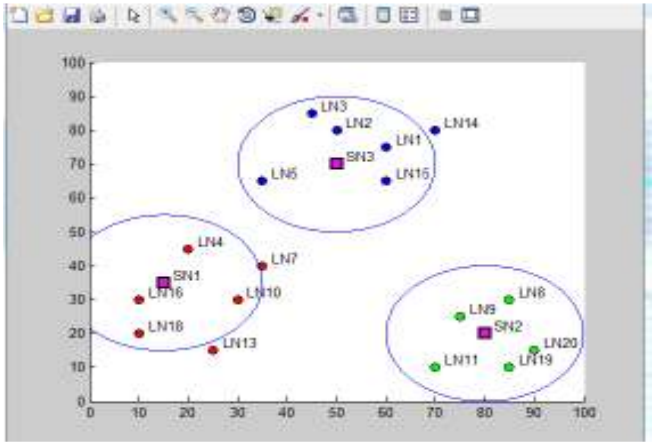

Figure 6.Super node selection based on kmedoid.

4.3 Experimental Setup

The MATLAB R2010a software is installed on windows operating system for this application. Our performance tests were accomplished on network area of 100X100, initialised with certain nodes. The nodes are placed on an area, each node is having an initial energy of 100joules.MainToolBox contains all the modules of the project such as Network Initialisation, Cluster Formation, Identifying Super Nodes, Communication Establishment and Data Gathering.

4.3.1 Result Analysis

We conclude that our results will gives the better performance in terms of available nodes, delay and energy consumption.

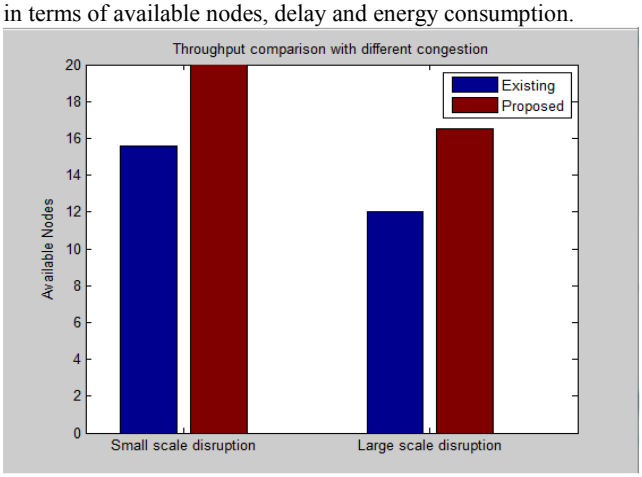

Figure 7. number of available nodes in different physical network disruptions

While the number of available nodes in the existing network represents the **lower value**, the proposed network achieves

maximum number of available nodes regardless of the physical network disruption scenarios.

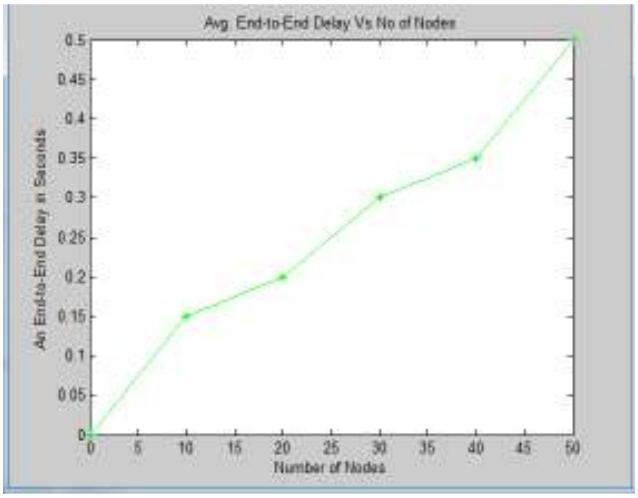

Figure 8. Slight increase in Delay as number of nodes increases

Here we are reducing the delay by providing support from overlay Nodes, Here we are achieving least delay that to in terms of seconds.

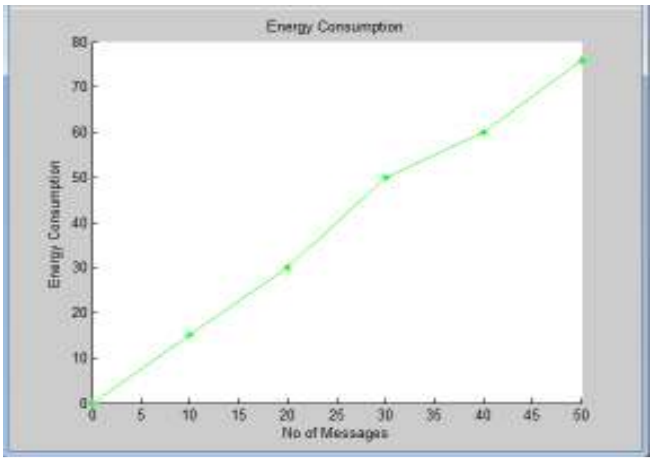

Figure 9. Reduced Energy Consumption

From the above graph it has been concluded that as the number of nodes increases energy consumption also increases slightly and here we are reducing consumed energy by using energy efficient algorithm Kmedoids, so here reduced the energy consumption to least.

5. CONCLUSION

In an Overlay-based data mining architecture, which completely distributes management and processing functions by using overlay based network technologies that can potentially provide scalable data mining in large-scale network. However, but physical network disruption is one of the major issue and that goes on decreases the service availability of data mining and hence increased energy consumption and delay. To overcome with these issues it is important to propose new scheme, hence we proposed Overlay network based on neighbor selection and task allocation schemes. Along with that in order to reduce energy consumption, in proposed system K-Medoids algorithm was used to create optimal number of clusters and based on minimal distance algorithm the service ability of every network was enhanced efficiently for better performance. The results obtained from our work demonstrated the effectiveness of our proposed system in terms of energy consumption, end to end delay and number of alive nodes. Thus, our proposed schemes can be Wireless Sensor Networks are important for monitoring the changes in the environmental conditions so that preliminary cautions can be taken to deal with the problem. Moreover Cluster based Wireless Sensor Network are hierarchical networks where the data extraction from super nodes can do effectively and overlay usage also gives better results in case of decreased service ability from neighbor nodes.

6. ACKNOWLEDGMENT

The work reported in this paper is supported by the college through the TECHNICAL EDUCATION QUALITY IMPROVEMENT PROGRAMME [TEQIP-II] of the MHRD, Government of India.

7. REFERENCES

[1] Data Mining with Big Data, vol. 26, no. 1, IEEE, 2014.

[2] Ubiquitous Analytics: Interacting with Big Data Anywhere, Anytime, vol.46, Issue 4, IEEE, 2013.

[3] http://documents.software.dell.com/statistics/textbook/data-mining-techniques

[4] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., et al. (2000), CRISP-DM 1.0, Chikago, IL. SPSS.

[5] Jiawei Han and MichelineKamber (2006), "Data Mining Concepts and Techniques", published by Morgan Kauffman, 2nd ed.

[6] Mrs. Bharati M. Ramageri, Lecturer**,** "DATA MINING TECHNIQUES AND APPLICATIONS", Indian Journal of Computer Science and Engineering, Vol. 1 No. 4 301-305.

[7] Data Mining at a Glance @Source:http://publib.boulder.ibm.com/infocenter/db2luw/v9r5/index.jsp?topic=%2Fcom.ibm.im.easy.doc%2Fc_dm_process.html

[8] Mohammad HosseinAnisi, Abdul Hanan Abdullah, ShukorAbdRazak, "Energy-Efficient DataCollection in Wireless Sensor Networks", Wireless Sensor Network, pp.329-333, October 2011.

[9] Xin Guan, Lin Guan and Xingang Wang, "A Novel Energy Efficient Clustering Technique Based on Virtual Hexagon for Wireless Sensor Networks", Volume 7, Issn 1349-4198, pp. 1891-1904, April,2011.

[10] Shio Kumar Singh, M P Singh, and D K Singh, "Energy Efficient Homogenous Clustering Algorithm for Wireless Sensor Networks", International Journal of Wireless & Mobile Networks (Ijwmn), Vol.2, No.3, August 2010.

[11] D. Kumar, T.C. Aseri, R.B. Pate, "Energy Efficient Clustering and Data Aggregation Protocol for Heterogeneous Wireless Sensor Networks", ISSN 1841-9836, E-Issn 1841-9844 Vol. No. 1, pp. 113-124, (March 2011).

[12] M. Ahmed and S. Vorobyov, "Collaborative Beamforming for Wireless Sensor Networks with Gaussian Distributed Sensor Nodes," Wireless Communications, IEEE Transactions, Vol. 8, No. 2,pp. 638 –643, Feb.2009.

[13] KiranMaraiya, Kamal Kant, Nitin Gupta "Architectural Based Data Aggregation Techniques in Wireless Sensor Network: A Comparative Study", International Journal on

Computer Science and Engineering (IJCSE), Vol. 3 No. 3 Mar 2011

[14] An Overlay-Based Data Mining Architecture Tolerant to Physical Network Disruptions, IEEE Std.10.1109, 2014.

[15] Wenjun Xiao, Mingxin He and Huomin Liang, "Cayley CCC: A Robust P2P Overlay Network with Simple Routing and Small-World Features", Journal of Networks, vol. 6, Issue 9,September 2011

[16] KatsuyaSuto, Hiroki Nishiyama, XueminShen and Nei Kato1, "Designing P2P Networks Tolerant to Attacks and Faults Based on Bimodal Degree Distribution", Journal of Communications, vol 7, Issue 8, August 2012.

[17] Nikita Jain and Vishal Srivastava "Data Mining Techniques: A Survey Paper", International Journal of Research in Engineering and Technology, vol. 2, Issue 11, November 2013.

[18] Distributed Data Mining in Peer-to-Peer Networks, IEEE Std.1089-7801,2006.

[19] DunrenChe, MejdlSafran, and Zhiyong Peng, From Big Data to Big Data Mining: Challenges, Issues, and Opportunities,B. Hong et al. (Eds.): DASFAA Workshops 2013,Springer-Verlag Berlin Heidelberg 2013.

[20] Michael Cardosa, Aameek Singh,Himabindu Pucha and Abhishek Chandra,"Exploiting Spatio-Temporal Tradeoffs for Energy-aware MapReduce in the Cloud", IEEE 4th International Conference on Cloud Computing, 2011.

[21] A. P. Reynolds, G. Richards, and V. J. Rayward-Smith, "The Application of K-medoids and PAM to the Clustering of Rules"

[22] Shalini S Singh, and N C Chauhan, "K-means v/s K-medoids:A Comparative Study".