

Elimination of Data Redundancy and Latency Improving in Wireless Sensor Networks

Vivek Chandran K. C, PG Scholar

Department of Computer Science & Engineering, Malabar
Institute of Technology
Anjarakandy, Kannur (Dt), Kerala

Nikesh , Faculty

Department of Computer Science & Engineering, Malabar
Institute of Technology
Anjarakandy, Kannur (Dt), Kerala

Abstract— The data aggregation and energy consumption of the sensor nodes play an important role in Working of WSN. So it is necessary to minimize the data redundancy and continuous monitoring operation results in more energy consumption of the sensor nodes. To resolve this problem we built an aggregate tree for given size of sensor network. SVM method was applied on tree to eliminate the redundant data. The LSH code is generated in each node and code are sent to supervisor node. The supervisor node find out sensor node with same data select only sensor node to send actual data Supervisor node contain detail about distance between the node and redundant data. By using a protocol ‘Dynamic Energy Efficient Latency Improving Protocol We find out shortest distance between each nodes by adding all distance between each nodes. Actual data transfer only after confirming the shortest path .The overhead of Network traffic is reduced, resulting in improvement of energy efficiency and latency and perform better in all the scenarios for different size network.

Keywords: Wireless Sensor Network , Supporting Vector Machine , Local Sensitive Hashing, Distance metric

I. INTRODUCTION

In the past few years a lot of inventions have been made in the field of wireless sensor network. Nowadays, Wireless Sensor Network (WSN) has become essential in almost all applications of monitoring or control, due to its many advantages. The ‘WSN’ is one of the important persistent networks which sense the environmental situation through various sensing parameters. The WSN is made up of thousands of sensor nodes including source node, sink node and base station to communicate with the outside world. Every sensor node has limited range transceiver, low power embedded .This sensor nodes are used to exchange data or information between two. The nodes sense the data and pass the sensed data to the central server. Here we focus on data aggregation problem and latency improving in wireless network.

The deployment of large number of sensors over sensing region increases the data accuracy. The sensors deployed on the nearby region sense the phenomena leading to produce lot of redundant data. When multiple nodes sense the same similar packets to near one this lead to waste of resources in sensor node. This problem can be overcome by data aggregation technique. Data aggregation tries to collect

the most critical data from the sensor and make it accessible to the sink in an energy efficient manner. The method will find out node that send real data and eliminate other. After removing redundant data by using a DEELIP we find out a shortest path to send the data it is done by adding all distance between the node then only it send the data.

II. RELATED WORK

Ali Kashif Bashir et al. have proposed an energy efficient in network RFID Data Filtering Scheme (EFID) that divides the node into clusters. Every cluster head filters the data of its member nodes and send it towards the base station .Inter-cluster data is being filtered at neighboring nodes along the route. For this ,they use a clustering mechanism where cluster heads eliminate duplicate data and forward filtered data towards the base station. The main drawback of this method is that the node cannot eliminate the all duplication by itself.

SuatOzdemir have proposed fault tolerant data aggregation scheme that eliminates the false data sent by malfunctioning and compromised sensor nodes .To conserve energy while eliminating false data ,an in network outlier detection technique that is based on locality sensitive hashing scheme is used. This work uses Euclidean distance to compute the distance between sensor node data sets. It is also observed that if sensor data are highly correlated FTDA can eliminate redundant data transmission and reduce the overall data transmission in the network.

Hamed Yousefi have proposed structure free real time data aggregation protocol RAG, using two mechanisms for temporal and spatial convergence of packets-judiciously waiting policy takes advantage of the available slack for efficient aggregation by dealing packets on their way to the sink as long as their deadline are not missed .A real time data aware any casting policy at MAC layer makes the routing decisions on the fly for efficiency aggregation of data as there is no pre constructed structure. The paper does not address all aspects. There is also scope for developing an aggregation aware real time routing protocol in mobile WSNs on the basis of the proposed work.

Nurcan Tezcan have proposed effective redundant node elimination method that considers even the smallest overlapping regions to establish a coverage set .Further an extensions scheme is presented that finds the minimum number of sensors among the coverage set,where the network connectivity is guaranteed

III. PROBLEM IDENTIFICATION

The WSN is made up of thousands of sensor nodes including source node, sink node and base station to communicate with the outside world. Every sensor node has limited range transceiver, low power embedded processor, small memory and limited battery. In WSN, all sensor nodes are battery operated, and these batteries are non-rechargeable. In most of the applications, due to complicated deployment of sensor nodes, it is difficult to replace the battery. An energy efficient operation of the 'WSN' is very necessary for continuous monitoring application and prolong lifetime of the network. Also, it is obligatory to have modest processing time to avoid delay in sensitive applications like Industrial control, disaster monitoring, military surveillance and remote patient monitoring, etc.Main problem that we focus on this paper are Low power embedded processor,Small memory and limited battery,Impact of simply forwarding Data from sensor.

IV. PROPOSED MODEL

There is lot of data redundancy in WSN due to spatial and temporal redundancy. So it is necessary to reduce the data redundancy by adopting suitable aggregation techniques. To resolve this problem, supporting vector machine (SVM) based redundancy elimination for data aggregation in WSN (SRDE) has been proposed in this work

In this work aggregation tree is build for the given size of the given network. Then, SVM method is applied on the tree to eliminate the redundant data. Locality Sensitive Hashing (LSH) is used to minimize the data redundancy and to eliminate the false data based on similarity[18,19]. During each session, the LSH code is generated on the latest data readings of sensor nodes. The size of LSH code b is very small compared to the data size of last m readings. The LSH codes are sent to the aggregation supervisor node. Aggregation supervisor maintains redundancy count for similar LSH code . The aggregation supervisor finds sensor nodes that have same data and selects only one sensor node among them to send actual data. Aggregation supervisor also eliminates the outliers and it did not accept the data sent from any other than selected node. The benefit of this approach is that it minimizes the redundancy and eliminate the false data, thus improving the overall performance of the WSN.

WORKING MODEL SUPPORT VECTOR MACHINE

A Support Vector Machine (SVM) is a concept in statics and computer science for a set of related supervised learning methods that analyze data and recognize patterns. SVM is used for classification and regression analysis . The basic SVM takes a set of input data and predicts for each given input, which of two possible classes forms the output , making it a non probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or other . An SVM model is a representation of example as point in space ,mapped so that the example of the separate categories are divided by a clear gap that is as wide as possible. New example are then mapped into that same space and predicated to belong to a category based on which side of the gap they fall on.In the classification of redundant data.

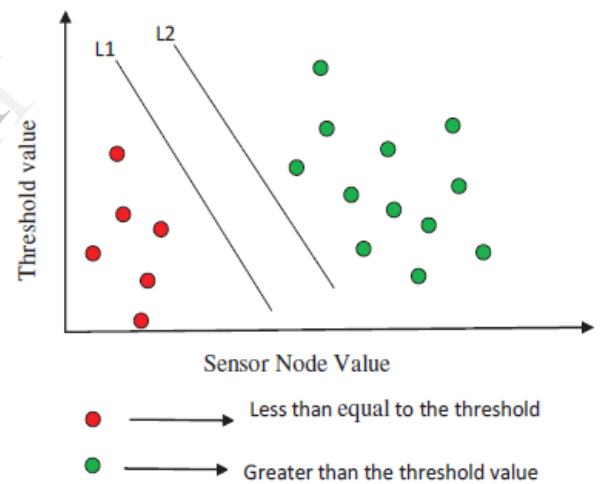


Fig. 1. Classification of Redundant data

In the field of machine learning, the purpose of statistical classification is to use characteristics of data or data or objects, to identify ,which class (category)it fits into.A linear classifier realizes this by making a classification decision based on the value of a linear combination of the distinctiveness .An object characteristics are also known as feature values,and are typically presented to the machine in a vector called a feature vector .A linear classifier splits a high dimensional input space with a hyper plane :all points on one side of the hyper plane are classified as "yes" while the others are classified as "no". The nodes generating the redundant data are classified by the linear

classifier ,using threshold value ,and this is shown as L1 and L2 in figure.1

Consider the following equation:

$$f(x) = T * c + N \tag{1}$$

T is the threshold value and N is the node IDc is the constant Then L1 and L2 are two hyperplanes.

$$R = \begin{cases} L1 = f(x) \leq 0 \\ L2 = f(x) > 0 \end{cases} \tag{2}$$

In equation (2),L1 is the plane where all the values are less than $f(x)$ or equal to zero and L2 is the plane where all the values are greater than $f(x)$.SVM perform two functions ,one is classification and the other is correlated data elimination.

Classification: The aggregation supervisor will use SVM technique to classify the data from the nodes based on the similarity estimation using Locality Sensitive Hashing

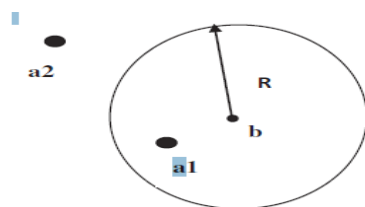


Fig. 2. Data Points and Radius in R-near neighborhood

In this figure(2) ,a,b are the data points and R is radius .Let D denote a set of data points and it belongs to the N dimensional space R^n . a_i denotes the i th coordinate of A . a and b are some data points in R^n .The value of radius R decides the neighborhood to which the data point belong .If the distance between a and b is less than R ,then a is an R near neighbor of b .

The LSH algorithm is mainly dependent on the locality sensitive hash function. Let H be a family of hash functions mapping R to some universe

H is locality sensitive if for any two points a, b element of R ,it satisfies the following condition:

$$H = \begin{cases} \text{if } |a-b| \leq cR \text{ then } D_{r_H} [h(a) = h(b)] \geq a_1 \\ \text{if } |a-b| \geq cR \text{ then } D_{r_H} [h(a) = h(b)] \leq a_2 \end{cases} \tag{3}$$

In equation (3), c is a constant. In the above equation $A1 = 1 - (R/n)$ and $A2 = 1 - (cR/n)$.

An LSH family must satisfy $A1 > A2$.using equation(3),the

LSH family can determine whether the two data points are in the R near neighborhood of each other.

We need an LSH algorithm that can work on vectors, in order to use the LSH technique on sensor node data sets.

B.LSH Algorithm:

LSH function is useful for finding similar objects in a dataset .The LSH algorithm is required to work on sensor data sets. In a random hyper plane based LSH algorithm for vector is proposed .Consider the vectors $p, q \in R^n$ and define a hash function as

$$h_r(p) = \begin{cases} 1 \text{ if } p \cdot r \geq 0 \\ 0 \text{ if } p \cdot r < 0 \end{cases} \tag{4}$$

Then vector p and q are

$$P_r [h_r(p)] = h_r(q) = 1 - \frac{\theta(p, q)}{\pi} \tag{5}$$

$$\theta(p, q) = \Pi * (1 - p_r) \tag{6}$$

The random hyper plane based hash function can measure the similarity between any pairs of sets. Similarity in terms of angle between the two vector is determined. Hash function measures the similarity in terms of angle between two vectors. For the resource constrained sensor node, finding the angle between two vectors is not an important task. So here following methods described in [19], is used, rewriting the equation in terms of Hamming distance.

$$D_h(LSH_p, LSH_q) = l * (1 - p_r) \tag{7}$$

In the equation (7) $LSH(p)$, $LSH(q)$ are the codes of the vectors p and q respectively.

$D_h(LSH_p, LSH_q)$ is the hamming distance between LSH_p, LSH_q . Each LSH code length of 1 bit and which is smaller than original vectors p and q . Using the equation (4) ,the above equation is rewritten as

$$D_h(LSH_p, LSH_q) = b * \frac{\theta(p, q)}{\Pi} \tag{8}$$

Using the equation (8) , sensor nodes are enabled to measure the similarities of their data sets. By using simple bit comparisons, the sensor nodes will measure the similarities of their data sets. The equation (9) is used to compute the

threshold value. The threshold value is used to measure similarities of data.

$$TH(D_h) = b * \frac{TH}{\prod} \quad (9)$$

A. LSH code generation

The LSH code are generated and this codes are used to reduce the amount of data transmission assume each node sensed the value is n bits and each sensor nodes has a data vector of size mxn bits. If mxn data bits are fully sent to a data aggregator then sensor nodes battery life came down drastically . LSH code can represent sensor data using less number of bits thereby reducing amount of data transmission. When a node applies LSH algorithm to its data, it will obtain the 1 bit LSh code where . there is a trade off between the value of (m*n) and b . When(m*n) , and b values are close to each other the outlier detection ability of protocol increases. We can compute the probability P that LSH code of data vector u and v are equal or not by using the equation (4). The probability of successful similarity test can be expressed by following equation (10).

$$P_{similar} = \sum_i^{TH(D_h)} \begin{bmatrix} 1 \\ i \end{bmatrix} p^{l-i} * (1-p)^i \quad (10)$$

In the above equation (10) , the probability similarity between two vectors is determined.

B. Correlated Data Elimination

Each sensor node in a sensor sends its LSH code to each data aggregation supervisor

Upon request . Each sensor nodes sends its LSH code along with its unique sensor node ID . The aggregation supervisor verify the LSH code of the sensor node pair using the equation (8) and(9). There will be two cases , one is the LSH code matching and the other is not in matching condition. The data aggregator looks for the following two cases: In the first case when LSH code are exactly same the data aggregate supervisor discover the sensor nodes that have same data.The redundant data can be eliminated .The aggregation supervisor has to find those sensor nodes that send the same data and eliminate the redundant data. If the data aggregation gets the same LSH code from two or more nodes ,then it will select data from only one sensor node .In the second case ,the LSH codes are different ,so that the data aggregation will receive the LSH code of the sensor nodes. In this case the LSH codes are totally different ,so the data aggregator will take data from all the sensor node.

TABLE I REDUNDANT DATA COUNT

SID	LSH Code of Node	Redundant Data Count
1
2

In the first case ,a table is maintained ,to take count of redundant data packet .

If the redundant packet are more than that of threshold value,then warning msg are send to the center node.

In table (1) the LSH codes of nodes are compared .If LSH code is same as another LSH code ,then count is increment by 1.The threshold value is define by equation (9)

Algorithm:

1. Start
2. Build the Aggregation Tree.
3. Generate the LSH code for each node in the aggregated tree
4. Check for redundant data.
5. If (two LSH codes match)
6. {
 - a. Increase the redundant data count.
 - b. Keep the data of only one sensor node to send the data that have same LSH code.
7. }
8. Else
9. Go to the next level;
10. If (redundant data count > Threshold value)
 - a. Send a warning message to the sender of the LSH code.
11. Exit

Algorithm 1: LSH code generation

.. After find out the node with out Data Redundancy we try to reduce the transmission delay we have considered a new concept of the 'Distance Metric' (DM) Here, the distance means the air distance between sink to particular sensor node. The total path distances from sensor nodes to sink node are calculated by adding all distances between nodes in that routing path. Actual data transferred only after confirming the shortest path. Since the path is the shortest, the transmission time can reduce and so that delay can be minimize In this way with all possible combination of path distance, a 'Distance Metric' is obtained. The total path distances from sensor nodes to sink node are calculated by adding all distances between nodes in that routing path. Actual data transferred only after checking the shortest

path. Since the path is the shortest, the transmission time can reduce and so that delay can be minimize

In this way with all possible combination of path distance, a

- All the sensor nodes are static during data collection process.
- All nodes possess the same ratings of transmission ,processing power and storage capacity.
- The geographical positions of node (X, Y) are known through GPS technique

If geographical positions are known then the aerial distance (d) between any two nodes can be estimated by using well-known Euclidean distance expression,

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

After find out the distance between the nodes if it is shortest way it will allow to send the data .By this method we can avoid the main problem of WSN .

V. ADVANTAGES/DISADVANTAGES

Advantages;

- 1) Performance improve
- 2) Improve the packet delivery ratio
- 3) Low value of Drop
- 4) Overcome Delay

Disadvantages;

Presently, in this we first build aggregation tree for the given size of sensor network. Then SVM method was applied on tree to eliminate the redundant data.LSH is used to remove the problem if the LSH code produced will have any problem in calculation it will make all the working of the process

VI. FUTURE SCOPE

The 'WSN' is one of the important persistent networks which sense the environmental situation through various sensing parameters .Our proposed scheme, it can reduce the problem overhead in the sensor.

To increase the merits of our research work, we plan to investigate the following issues in our future research

- ❑ In the future, this protocol can be physically deployed in indigenous low cost sensor node for Industrial control applications.
- ❑ In the future, we want to predict a model that helps to get the minimum threshold for a particular area

VII. CONCLUSION

The SVM based method along with DEELIP idea will help to improve the working of Wireless Sensor Network. It provides very low end to end delay and efficient battery utilization. It is the outstanding achievement of this research work. The benefit of this approach is that it minimizes the redundancy and eliminate false data to improve the overall performance of WSN. Various terminologies, principle of operation of a wsn and recent developments have been discussed. The life of the battery is an important parameter which decides the area of application of the wsn. The Locality Sensitive Hashing is used to eliminate data redundancy and false data based on similarity .During the process LSH codes are sent to the supervisor node. The aggregation supervisor finds sensor node that have the same data using LSH code and select only one sensor node among them to send the actual data .After that by using 'DEELIP' protocol, we check that node selected to send is a best path so for that it calculates the distance between nodes and also updates the distance between the sink to that node. So that sink node sent their recent position in announcement message to all nodes. At hop level one, it will update this between sink to their node, so the sink node has to send their recent one position in our announcement message. This method finally results in low transmission delay. The WSN perform better in all the different sized networks and varying data rate.

REFERENCES

- 1 John A.Stankovic,"Wireless Sensor Network",University of Virginia,June 19,2006.
- 2 C.Y.Chong and S.P.K umar,"Sensor networks:Evolution,Opportunities ,and challengers," ,IEEE Proceedings,VOL,91,NO.8,PP,1247-1254,AUGUST 2003.
- 3 F.L.LEWIS ,,"Wireless Sensor Networks",Smart Environments:Technologies,protocols ,and Application ed.D.J Cook and S.K.Das,John Wiley,New York.
- 4 I.F.Akyildiz,W.Su,Y.Sankarasubramaniam,and E.Cayirci,Wireless sensor network: a survey",Elsevier,Computer Networks 38
- 5 Cinzia Cappiello and Fabio A.Schreiber,"Experiment and analysis of quality and energy aware data aggregation approaches in WSN.