DOI: 10.21917/ijct.2019.0302

ENSEMBLED ADABOOST MACHINE LEARNING ALGORITHM WITH NONLINEAR REGRESSION TREE FOR ENERGY AWARE DATA GATHERING IN WSN

S.S. Aravinth, A. Arul Prasath and B. Arunkumar
Department of Computer Science Engineering, Dhirajlal Gandhi College of Technology, India

Abstract
Data gathering is a process of collecting more number of data from distributed sensor nodes and sends these data to sink node. During the data gathering, energy consumption (EC) is a major concern for enhancing the network lifetime (NL). Several WSN architectures have been developed to resolve this problem In order to improve the data gathering efficiency, AdaBoost Nonlinear Regression Tree Classification (ABNRTC) technique is developed. ABNRTC technique improves the data gathering with minimal EC. Initially, energy of each senor nodes is measured. After that, mobile sink node gathers the sensed information from the high energy sensor nodes with minimum delay. Then the mobile sink node classifies the collected data packet using nonlinear regression tree based on their relationship among the sensor nodes in WSN. The relationship between the data packets are measured using population Pearson product-moment correlation coefficient. AdaBoost algorithm is a boosting technique for grouping the several weak nonlinear regression tree classifiers to make a single final output of boosted classifier. Finally, the classified data packets are sent to base stations (BS). Simulation of ABNRTC technique is carried out with different parameters such as EC, NL, data gathering efficiency, delay, classification accuracy (CA), false positive rate (FPR) and classification time (CT).

Keywords:
Wireless sensor network, Nonlinear Regression Tree, Pearson Product-Moment Population Correlation Coefficient, AdaBoost Algorithm

1. INTRODUCTION

WSN contains number of low power, low cost sensor nodes to monitor physical and environmental circumstances. These monitoring data are collectively sent to BS. The sink node gathers data either directly or indirectly through intermediate nodes. Recently, several approaches developed for data aggregation.

A Distributed Data Gathering Approach (DDGA) was introduced in [1] to attain optimal data gathering. But, it does not classify the data packet. A two-step approach was designed in [2] for mobile data collection. However, data gathering delay was more. A structure-free and energy balanced data aggregation protocol (SFEB) was designed in [3]. Though, the aggregated data was not classified. Multi-tier clustering scheme was developed in [4] to collect the data with less EC. But, the performance of EC was lower. Contact-Aware ETX (CA-ETX) was introduced in [5]. But, the performance of reliability was not improved. In [6], a tree-cluster-based data-gathering algorithm (TCBDGA) was designed. Though, data loss was not considered.

An anchor selection based on tradeoff between neighbor amount and residual energy (AS-NAE) technique was introduced in [7]. But data collection efficiency was not increased. An energy efficient structure-free data aggregation and delivery (ESDAD) protocol was designed in [8]. However, it does not have efficient reliability in the sensing field. A distributed data compression framework was introduced in [9]. But the classification was not performed. A Cell-based Path Scheduling (CPS) algorithm was designed in [10]. But, the energy of sensor nodes was not measured. In order to overcome above said issues, an ABNRTC is developed.

The main contributions of the paper is described as follows,

- ABNRTC technique is introduced for achieving energy efficient data gathering.
- The mobile sink node gathers the sensed data packet from high energy sensor nodes.
- After that, mobile sink node classifies the gathered data packet using AdaBoost nonlinear regression tree classifier.
- Initially, the weak learner is a decision tree which is constructed based on relationship between the data packets.
- The relationship is measured using population Pearson product-moment correlation coefficient.
- Then the error value of each weak learner is measured and assigned the weight value.
- The classifier with minimum error is trained with strong classifier.
- This helps to improve the data packet CA with minimum time as well as false positive rate. Finally, the classified data packets are transmitted to BS.

The rest of paper is structured as follows: section 2 provides description about ABNRTC with neat diagram. Section 3 describes the simulation settings and result discussions are described in section 4. Section 5 reviews the related works. Section 6 provides the conclusion.

2. ADABOOST NONLINEAR REGRESSION TREE CLASSIFICATION TECHNIQUE FOR DATA AGGREGATION IN WSN

In WSN, many sensor nodes has ability to sense the data and processed with minimum energy and forward the sensed data packets to mobile sink nodes through wireless transceiver. In order to obtain efficient data gathering, the sensor nodes and mobile sink node is deployed to maximize the amount of data collection and balances the EC. The major concern in the design of a WSN application is improving data gathering efficiency and minimizes the EC. Data gathering in WSN contains minimum delay, EC and packet loss rate. With this objective, ABNRTC technique is introduced. The following system model is used for improving ABNRTC technique quality.

2.1 SYSTEM MODEL

The system model is described to enhance the quality of the proposed ABNRTC technique. Let us consider, a wireless sensor network contains rectangular sensing area and arranged in graph $G = (V,E)$. Here vertices V represents a number of sensor nodes denoted as $V = S_1,S_2,\ldots,S_n$ and a set E of edges i.e. connection between the nodes in a sensing rectangular area. Let NN_n is the set of neighboring node. The number of data packets $DP_1,DP_2,\ldots,DP_n$ are monitored by sensor nodes and these data are collected by mobile sink node MS_A. Based on the above said system model, the proposed ABNRTC is designed.

2.2 ADABOOST NONLINEAR REGRESSION TREE CLASSIFICATION

The major concern of sensor nodes is the energy problem since the mobile nodes have restricted energy resources. Therefore, AdaBoost Nonlinear Regression Tree classification technique is developed to perform data aggregation as well as data classification. The number of sensor nodes distributed over the network to monitor and sense the information from the various objects. ABNRTC uses mobile sink node which act as data aggregator to gathers the sensed data from the sensor nodes with minimum EC. The Fig.1 shows flow process of ABNRTC technique to improve data gathering in WSN. Initially, energy of each node is measured. Then the low energy node sends the monitored information to their neighboring high energy sensor node. The higher energy sensor node receives the sensed data from lower energy sensor nodes.

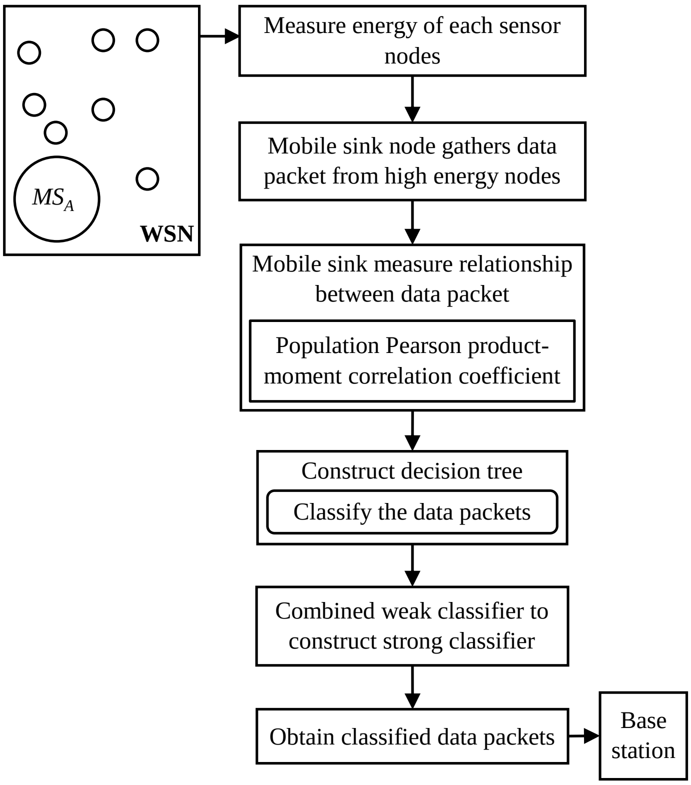

Fig.1. Flow Process of ABNRTC

After that the mobile sink node (MS_A) gathers sensed data from the higher energy nodes rather than all the nodes in network. The mobile sink node travels across sensing area and directly gathers the data. Then the MS_A node finds the relationship between the gathered data packets for classification. The nonlinear regression tree is constructed to classify the data packet based on their relationship. Population Pearson product-moment correlation is used for finding the relationship between the data packets. Regression tree is a decision tree. Here the weak learner as nonlinear regression tree which is trained with AdaBoost. AdaBoost is a boosting technique for grouping the several weak classifiers to construct a strong classifier. Finally, mobile sink sends the classified data to.

2.2.1 Node Energy Calculation:

In WSN, energy in data aggregation plays a major role to extend NL. Energy of the sensor node is measured with the product of power and time in seconds. After that, the low energy sensor node sends the data packet to neighboring node. The low energy sensor node detects the nearest higher energy node through distance calculation. The nodes and their neighboring node distance is measured as follows,

$$d(S_1, S_2) = \sqrt{(u_1 - v_1)^2 - (u_2 - v_2)^2} \tag{1}$$

From Eq.(1), d denotes a distance between the sensor nodes and current coordinate for s_1 is (u_1,v_1) and for s_2 is (u_2,v_2). Therefore, the node with minimum distance is selected as neighboring node with higher energy. The mobile MS_A node gathers the data from higher energy sensor nodes. ABNRTC technique selects the nodes based on their energy and guarantee data collection. Therefore, the energy is a significant resource during data collection for improving the NL. In addition, the energy of the sensor nodes is measured based on amount of energy consumed for data packet transmission and reception.

$$E = E(DP_{TX} + DP_{RX}) \tag{2}$$

From Eq.(2), E denotes energy of data packet transmission DP_{TX} and data packet reception DP_{RX}. From that, the EC of the entire nodes in WSN is measured using following mathematical formula,

$$EC = E(S)\text{*No. of sensor nodes} \tag{3}$$

From Eq.(3), $E(S)$ denotes an amount of energy consumed by single sensor nodes which is multiplied with number of sensor nodes in WSN. The Fig.2 shows the energy efficient data gathering in WSN where the green color node indicates higher energy node which collects the information from low energy sensor nodes node.

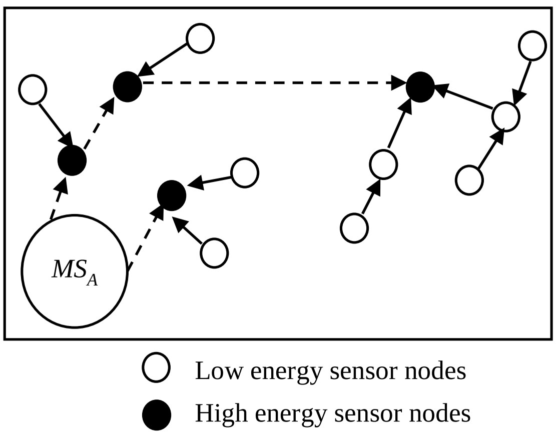

Fig.2. Energy efficient data gathering in WSN

As shown in Fig.2, the low energy sensor nodes are indicated in orange color. Then the mobile sink MS_A node gathers the sensed information from higher energy nodes thus minimizes the data loss. It is mathematically formulated as follows,

$$MS_A(DP) \rightarrow E_H(S) \tag{4}$$

From Eq.(4), MS_A denotes a mobile sink node collects the data packet (DP) from high energy sensor nodes $E_H(S)$. This helps to reduce the EC and improving the NL. The mobile sink node travelling paths are indicates dotted line. Therefore, the mobile sink node gathers the data packet with minimum delay for achieving high reliability. Then the gathered data packets are classified and send to BS. Algorithm 1 describes an energy efficient data gathering in WSN.

Algorithm 1 Energy Efficient Data Gathering

Input: Number of sensor nodes $S_1, S_2, \ldots, S_n$, Number of data packets $DP_1, DP_2, \ldots, DP_n$, mobile sink node MS_A.

Output: Improve data aggregation

Step 1: Begin

Step 2: For each sensor nodes

Step 3: Measure energy using Eq.(2)

Step 4: Low energy node select nearest high energy sensor node using Eq.(1)

Step 5: Mobile sink gathers data from high energy node using Eq.(4)

Step 6: End for

Step 7: End

Initially, energy of each sensor nodes is measured. Then the low energy sensor nodes send their data to nearest high energy nodes. Then the mobile sink node visits high energy nodes to collects the sensed data packets. This helps to reduce the EC in data gathering and maximize the NL. As a result, ABNRTC technique increases data gathering efficiency with minimum delay.

2.2.2 Adaboost Nonlinear Regression Tree based Data Packet Classification:

After data gathering, the mobile sink node performs data packet classification using ensemble classification technique called adaboost nonlinear regression tree. Nonlinear regression tree is often called as decision trees in which sensed data are classified by a nonlinear combination of the data packets among the sensor nodes. The regression analysis is a statistical process for finding the relationship between data packets using population Pearson product-moment correlation. It is used to measure the correlation between the data packets. It has a value between the ranges from +1 to −1, where 1 denotes a positive correlation, 0 indicates a no correlation, and −1 indicates a negative correlation between the data packets. Pearson's correlation is the ratio of covariance of two data packets and the product of their standard deviations. Hence the name involves a product moment correlations. The sensor nodes monitor the data from the different environmental conditions such as temperature, pressure, humidity and so on. In WSN, each sensor node monitors the number of data packets hence the population of data packets is generated. Among the data population, the relationship between the data packets are more efficient to improve the classification. Let us consider the number of data packets $DP_1, DP_2, DP_3, \ldots, DP_n$. The relationship between the data packets are measured as follows,

$$\rho(DP_1, DP_2) = \frac{c(DP_1, DP_2)}{\sigma DP_1, \sigma DP_2} \tag{5}$$

From the Eq.(5), $\rho(DP_1, DP_2)$ denotes the population Pearson product-moment correlation coefficient and C denotes a covariance between data packets DP_1 and DP_2. σDP_1, σDP_2 represents a standard deviation of two data packets DP_1 and DP_2. The Eq.(5), is derived and obtaining the final correlation results,

$$\rho(DP_1, DP_2) = \frac{\sum DP_1, DP_2 - \frac{\sum DP_1 * \sum DP_2}{n}}{\sqrt{\sum DP_1^2 - \frac{\left(\sum DP_1\right)^2}{n}} * \sqrt{\sum DP_2^2 - \frac{\left(\sum DP_2\right)^2}{n}}} \tag{6}$$

From Eq.(6), n denotes a number of data packets, DP_1 and DP_2 are two data packets, $\sum DP_1 * DP_2$ refers the sum of cross product of DP_1 and DP_2, $\sum DP_1$ is the sum of DP_1, $\sum DP_2$ is the sum of DP_2, $\sum DP_1^2$ is the sum of squared score of DP_1 and $\sum DP_2^2$ is the sum of squared score of data packet 2. The population Pearson product-moment correlation provides positive correlation +1 which means the relationship between the data is identified. The negative correlation provides false relationship between the data.

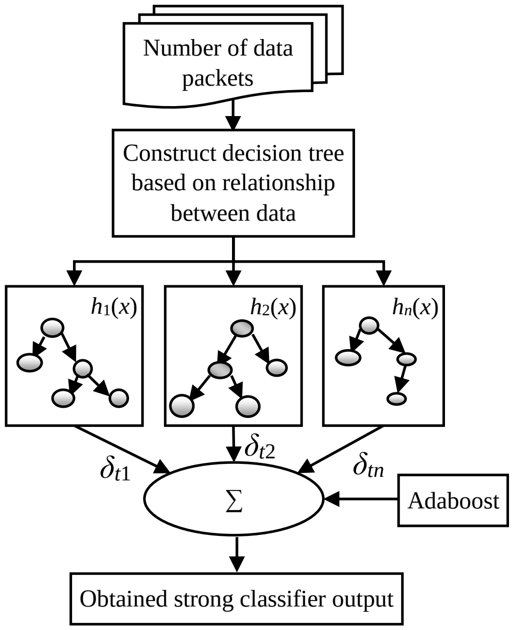

Fig.3. Adaboost Nonlinear Regression Tree classifier

Based on the relationship between the data, the decision tree is constructed to classify the data packets. The nonlinear regression tree is a decision tree to classify the data based on their correlation. A regression tree is a decision based classifier that uses a tree model for classifying the data packets. The Fig.3 shows the flow process of adaboost nonlinear regression tree to obtain the strong classification results. The decision tree is constructed where internal node represents a "test" on a data packet based on their relationship and branch node represents the outcome of the test, and the leaf node provides the output class label which is obtained after classifying all data packets. The path from root node to leaf node denotes a classification rules. AdaBoost is an adaptive one where subsequent weak classifiers are combined. A weak classifier is a base classifier that provides some classification error. In ABNRTC technique, the classified data packets from base learner are tweaked using Adaboost algorithm.

Let us consider the training set $\{(DP_1,Y_1),(DP_2,Y_2),\ldots,(DP_n,Y_i)\}$ where DP denotes an input data packets and Y represents a final classification output class. The Nonlinear Regression Tree classifier function provides a class output $Y_i = \{-1,+1\}$, where $Y_i = -1$ denotes an incorrectly classified data packets and $Y_i = +1$ denotes a data packets are correctly classified. The set of weak hypothesis (i.e. weak classifier) is $\{h_1(x),h_2(x),h_3(x),\ldots,h_n(x)\}$. Therefore, a decision tree is a classifier which performs recursive classification to classify data packets. The adaboost classifier is applied to construct a strong classifier. The summation of weak classifier provides the output hypothesis for each data packets to obtain strong classifier output which is mathematically denoted as,

$$H_f(DP_i) = \sum_{i=1}^{n} h_i(x) \tag{7}$$

From Eq.(7), $H_f(DP_i)$ denotes a strong classifier output and $h_i(x)$ denotes a number of base learners. A weight value is assigned to each data packets in the training set. The decision tree classifies the data packets with high weights. At first, the uniform weight for all data packet is assigned as follows,

$$w_t(DP_i)=1/n \tag{8}$$

From Eq.(8) where w_t represents weight and n denotes a number of data packets. For each iteration, weak classifiers $h_i(x)$ is determined using weight value. The weighted sum error of the base learner for misclassified data packets are measured as follows,

$$\beta_E = \sum_{i=1}^{n} w_t(i) \tag{9}$$

From Eq.(9), β_E denotes a weighted sum error for misclassified data packets. Based on the error, the best weak learner is selected and it trained with strong classifier is formulated as follows,

$$h_i(x) = \arg\min \beta_E \tag{10}$$

After that, the weight of base learner is updated as follows,

$$w_t(i+1) = \frac{w_t(i)}{z_n} e^{-(\delta_t Y_i h_i(x))} \tag{11}$$

From Eq.(11) where $w_t(i+1)$ represents an updated weight of base learner and z_n represents a normalization factor. Therefore the weight of base learner is measured based on error value as follows,

$$\delta_t = \log_{10}\left(\frac{1-\beta_E}{\beta_E}\right) \tag{12}$$

The updated weight is normalized as follows,

$$w_t(i+1)=1 \tag{13}$$

Then the results of all base classifiers are combined into a strong classifier to improve the classification performance. The strong classifier output is obtained as follows,

$$H_f(DP_i) = sign\left[\sum_{i=1}^{n} \delta_t h_i(x)\right] \tag{14}$$

From Eq.(14), $H_f(DP_i)$ denotes the final strong classifier output. The output of strong classifier provides the two output results for improving the CA. From Eq.(14), sign indicates positive and negative results of the final output classifier. The output of strong classifier is described as follows,

$$H_f(DP_i) = \begin{cases} +1 & DP\ is\ correctly\ classified \\ -1 & DP\ is\ incorrectly\ classified \end{cases} \tag{15}$$

From Eq.(15), where the final strong classifier output $H_f(DP_i) = +1$ represents the number of data packets is correctly classified whereas -1 denotes data packets are incorrectly classified. After that, the mobile sink node sends the classified data to BS for further processing.

Algorithm 2: Adaboost Nonlinear Regression Tree classifier

Input: $\{(DP_1,Y_1),(DP_2,Y_2),\ldots,(DP_n,Y_i)\}$ is a set of training sets

Output: Improve CA

Step 1: Begin

Step 2: For each data packet

Step 3: Measure relationship between data packet using Eq.(5)-Eq.(6)

Step 4: If correlation coefficient ρ=+1 then

Step 5: Positive correlation between data packets

Step 6: else if correlation coefficient ρ=-1 then

Step 7: Negative correlation between data packets

Step 8: else

Step 9: no correlation between data packets

Step 10: end if

Step 11: Construct regression decision tree to classify the data packet based on correlation

Step 12: Initialize similar weight data packet using Eq.(8)

Step 13: Calculate weighted sum error of base classifier using Eq.(9)

Step 14: Select best weak learner with minimum weight using Eq.(10)

Step 15: Train a best weak learner for classifying the data packet using Eq.(14)

Step 16: if strong classifier output result Y_i = +1 then

Step 17: Data packets are correctly classified

Step 18: else

Step 19: data packets are incorrectly classified

Step 20: End if

Step 21: End for

Step 22: End

Algorithm 2 describes an Adaboost nonlinear regression tree classifier to improve the data packet classification. Initially, the nonlinear regression tree is constructed based on relationship between the data packets. The relationship between the data packets are measured using population Pearson product-moment correlation. The correlation coefficients provide the results as positive correlation, negative correlation and no correlation. After that, the decision tree is constructed based on the correlation between the data packets. The best weak classifier is determined with minimum error. Then the weak classifiers are combined and make a strong classifier based on the weight value. This in turn improves CA. As a result, ABNRTC technique improves data classification and data gathering with minimum EC and high reliability.

3. SIMULATION SETTINGS

An efficient ABNRTC technique is implemented in NS2.34 network simulator. Totally 500 sensor nodes are considered for simulation in a square area of A^2 (1500m*1500m). The Random Waypoint model is used as mobility model. The number of data packets is varied from 9 to 90. The sensor nodes speed is varied from 0 to 20m/sec. The simulation time is set as 300 sec. In order to improve the data gathering in WSN, the DSR protocol is used. The Table.1 provides the simulation parameters.

Table.1. Simulation parameters

Simulation Parameters	Values
Simulator	NS2.34
Network area	1500m×1500m
No. of sensor nodes	50,100,150,200,250,300,350,400,450,500
Mobility model	Random Way point model
No. of data packets	9,18,27,36,45,54,63,72,81,90
Sensor nodes speed	0-20m/s
Simulation time	300sec
Number of runs	10
Protocol	Dynamic Source Routing (DSR)

The performance results of different parameters with diffident methods namely ABNRTC technique and existing [1]-[4] are described in following section.

4. SIMULATION RESULTS AND DISCUSSIONS

The analysis of ABNRTC technique is performed with existing [1]-[4] with the help of table values and simulation results.

4.1 IMPACT OF ENERGY CONSUMPTION

EC is defined as amount of energy consumed by the sensor nodes for gathering the data packets. The EC is measured using Eq.(3) and measured in joule (J).

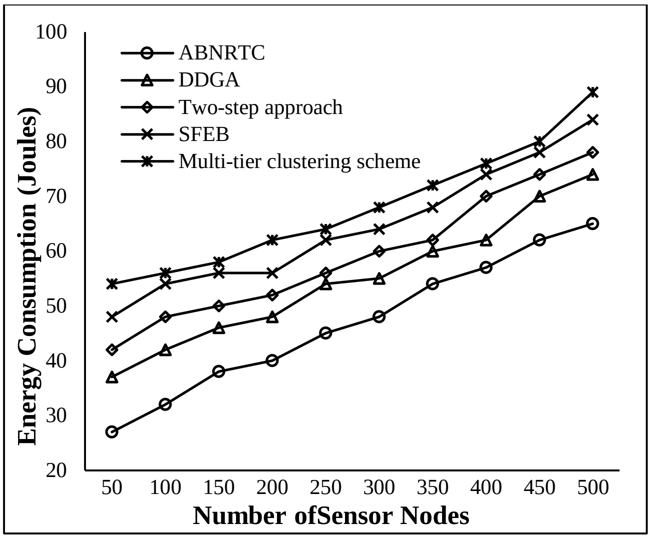

Fig.4. Simulation result of energy consumption

The Fig.4 shows the results of EC versus number of sensor nodes in WSN. EC is a major parameter, since limited energy supply reduces the data gathering efficiency. DSR protocol is used for data gathering to improve performance of networks. As shown in Fig.4, EC is considerably reduced using proposed ABNRTC technique when compared to existing methods. This is because, ABNRTC technique initially measures the energy of each sensor node. The sensor node which has higher energy is selected to collect the sensed data from the nearest low energy sensor nodes. Therefore, the node which utilizes the minimum energy for receiving the data packets from the other nodes is selected. After that, the mobile sink node collects the sensed data packet from the higher energy sensor nodes instead of all the nodes. This is because, low energy nodes not able to hold more data packets in longer duration. As a result, the some of the data packets are lost thereby reducing data aggregation efficiency. Therefore, the proposed ABNRTC technique uses mobile sink for gathering the data from the high energy sensor nodes. Due to this process, EC in data aggregation is reduced. As a result, EC of ABNRTC technique considerably reduced by 16%, 24%, 29% and 33% as compared to existing [1]-[4] respectively.

4.2 IMPACT OF NETWORK LIFETIME

NL is defined as the numbers of energy efficient nodes are selected for data gathering in WSN. It is measured in terms of percentage and measured as follows,

$$NL = \frac{(\text{No. of selected energy efficient sensor nodes})}{n} * 100 \quad (16)$$

From Eq.(16), NL denotes a network lifetime and 'n' denotes a number of sensor nodes in network. The Fig.5 illustrates a simulation result of NL versus number of senor nodes. As shown in the Fig.5, ABNRTC technique gives better performance in NL than the other existing methods. This shows the significant improvement of simulation results with energy efficient sensor nodes.

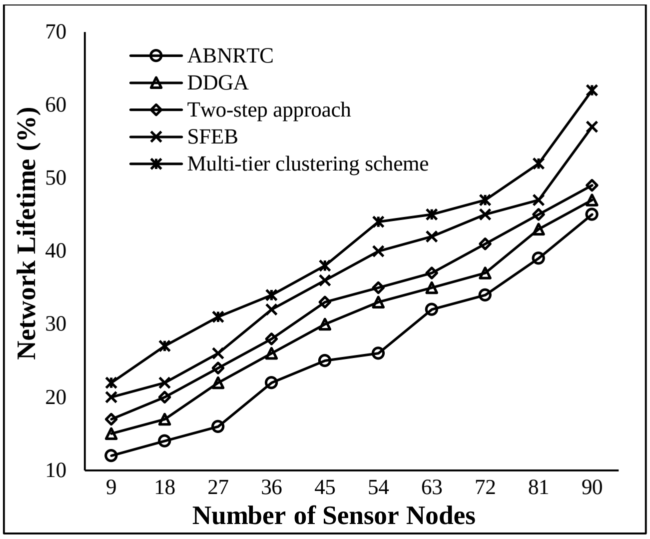

Fig.5. Simulation result of network lifetime

The sensor nodes with higher energy and less energy utilization are selected in data aggregation process. In general, sensor nodes are distributed in an environment and used batteries as energy source for transmitting and receiving the data packets from other sensor nodes. The insufficient energy of sensor nodes degrades the network performance. This also failed to prolong NL. In ABNRTC technique, low energy sensor node transmits sensed data packets to nearest high energy sensor nodes thus increase the NL. In addition, the minimum EC of the sensor nodes increases the NL. Let us consider the simulation results of NL, the proposed ABNRTC improves the NL with 80% while considering the 50 sensor nodes. Similarly, the NL of existing method namely [1]-[4] are obtained by 74%, 71%, 68% and 62% respectively. Hence, the proposed ABNRTC technique enhances the NL by 7%, 11%, 16% and 22% when compared to existing [1]-[4] .

4.3 IMPACT OF DATA GATHERING EFFICIENCY

Data gathering efficiency is measured in terms of data packet loss rate which is defined as ratio of number of data packets are lost to the total number of data packets. It is measured as follows,

$$DGE = \frac{\left(\text{Number of data packets lost}\right)}{\text{Number of data packets}} * 100 \tag{17}$$

From Eq.(17) DGE represents data gathering efficiency. It is measured in percentage (%). The Table.2 describes a data gathering efficiency with respect to number of data packets varied from 9 to 90. The data gathering efficiency using ABNRTC technique is improved when compared to existing methods.

While gathering the data from the high energy sensor nodes, the data loss rate is reduced thus increases the data gathering efficiency. In simulation results, while considering 9 data packets, 2 data packets gets lost using proposed ABNRTC technique. As a result, 22% of data packet lost is obtained using ABNRTC technique. Whereas 32%, 43%, 54% and 65% of data packet loss rate is obtained using [1]-[4]. Therefore, the efficiency of data gathering is improved using ABNRTC technique while reducing the data packet loss rate by 25%, 41%, 48% and 56% when compared to existing [1]-[4] techniques.

Table.2. Tabulation for data gathering efficiency

Number of data packets	Data gathering efficiency (%)				
	ABNRTC	**DDGA**	**Two-step approach**	**SFEB**	**Multi-tier clustering scheme**
9	22	32	43	54	65
18	24	35	48	57	67
27	26	38	50	59	68
36	30	41	52	61	70
45	33	43	54	63	72
54	35	45	57	65	73
63	37	48	60	67	75
72	39	51	62	68	76
81	42	53	64	70	78
90	45	55	67	71	79

4.4 IMPACT OF DELAY

Delay is defined as difference between expected time and actual time for data gathering in WSN. The mathematical formula for measuring the delay is expressed as,

$$\textit{Delay} = \textit{Expected time} - \textit{Actual time} \tag{18}$$

The Eq.(18) is used for measuring the delay during the data gathering in WSN.

Table.3. Tabulation for Delay

Number of data packets	Delay (ms)				
	ABNRTC	**DDGA**	**Two-step approach**	**SFEB**	**Multi-tier clustering scheme**
9	13	18	22	25	28
18	16	20	24	28	30
27	18	22	28	33	36
36	21	25	30	35	38
45	23	29	33	38	42
54	25	32	37	42	45
63	28	35	40	44	48
72	32	38	44	47	50
81	35	42	47	50	52
90	38	46	50	53	55

The Table.3 shows the simulation results of delay with respect to numbers of data packets varied from 9 to 90. While considering 9 data packets, the delay of ABNRTC technique is 13ms whereas 18ms, 22ms, 25ms and 28ms of delay is obtained using existing methods [1]-[4] respectively. Compared to existing methods, the proposed ABNRTC technique effectively improves the data gathering efficiency with minimum delay. As a result, the delay performance of ABNRTC technique is considerably reduced by 19%, 31%, 38% and 43% when compared to existing [1]-[4] techniques.

4.5 IMPACT OF CLASSIFICATION ACCURACY

CA is defined as the ratio of number of data packets are classified to the total number of data packets. The mathematical formula for CA is described as follows,

$$CA = \frac{(\text{Number of data packets are classified})}{\text{Number of data packets}} * 100 \quad (19)$$

From Eq.(19), the CA denotes classification accuracy which is measured in percentage (%).

Table.4. Tabulation for Classification accuracy

Number of data packets	Classification Accuracy (%)				
	ABNRTC	**DDGA**	**Two-step approach**	**SFEB**	**Multi-tier clustering scheme**
9	88	78	67	56	45
18	89	80	68	62	48
27	90	82	70	65	52
36	91	83	72	67	58
45	92	84	74	69	62
54	93	85	76	72	66
63	94	86	78	73	69
72	95	87	80	75	71
81	96	88	82	77	73
90	97	89	85	80	75

The Table.4 describes CA with respect to number of data packets varied from 9 to 90. The mobile sink node classifies the collected data packets based on correlation measure. Initially the mobile sink node collects the sensed data from the high energy sensor nodes. After that, the collected data packets are categorized using adaboost nonlinear regression tree classifier. The regression based decision tree classifier categorizes the data packet based on their relationship. Therefore, ABNRTC technique measures the relationship using population Pearson product moment correlation. Among the population, the correlation between the data packets is measured. The correlation coefficients provide positive and negative correlation between the data packets. The decision tree is constructed based on the correlation between the data packets. After that, the boosting technique is applied to make a strong classifier by determining a weak learner with minimum error. The weight is assigned for each base classifier based on their error value. Then the weak learners are summed to make a strong classifier. The strong classifier improves the CA of sensed data packets. From the above discussion, the CA is considerably increased using ABNRTC technique by 10%, 23%, 34% and 53% when compared to existing DDGA [1] and two-step approach [2], SFEB [3] and Multi-tier clustering scheme [4] respectively.

4.6 IMPACT OF FALSE POSITIVE RATE

The FPR is defined as the number of data packets are incorrectly classified to the total number of data packets. It is measured in percentage (%) and expressed as follows,

$$FPR = \frac{(\text{Number of data packets are inclassified})}{\text{Number of data packets}} * 100 \quad (20)$$

From Eq.(20), FPR denotes a false positive rate with respect to number of data packets. The Table.5 describes simulation results of FPR versus number of data packets.

Table.5. Tabulation for false positive rate

Number of data packets	False positive rate (%)				
	ABNRTC	**DDGA**	**Two-step approach**	**SFEB**	**Multi-tier clustering scheme**
9	21	33	44	53	58
18	23	34	46	54	60
27	25	36	48	55	62
36	28	38	50	56	64
45	30	40	51	58	66
54	31	42	53	60	67
63	33	43	55	61	68
72	35	45	57	62	69
81	38	47	60	64	70
90	40	49	62	65	72

From the result, the FPR of ABNRTC technique is reduced by combining weak classifier to make a strong classifier. The performance results of FPR is compared with proposed ABNRTC technique and existing DDGA [1] and two-step approach [2], SFEB [3] and Multi-tier clustering scheme [4]. Performance of FPR is reduced by applying ensemble adaboost learning classifier. At first, nonlinear regression tree is used as base classifier to categorize the sensed data packets. During the classification, weighted sum errors for misclassified data packets are measured. The error calculation during the classification is significant for obtaining the accurate classification. After that, the weak learner is selected with minimum error value and it trained to make a strong classifier. The strong classifier efficiently classifies the sensed data packets. Moreover, the ABNRTC technique improves the performance of classification results thus minimizes the false positive rate.

The simulation result of proposed ABNRTC technique and existing methods are illustrated in different colors. While classifying the data packets, the FPR of proposed ABNRTC technique is 21%. The other existing methods namely [1]-[4] obtain the incorrect classified results is 33%, 44%, 53% and 58%. Finally, the FPR results is considerably reduced by 26%, 43%, 49% and 56% when compared to existing [1]-[4], respectively.

4.7 IMPACT OF CLASSIFICATION TIME

The CT is defined as an amount of time required to classify gathered data packet. It is measured in milliseconds (ms) and expressed as follows,

$$CT = \textit{Number of data packets} * \textit{time}\,(\textit{classify the data packets}) \quad (21)$$

From Eq.(21), CT represents a classification time.

The Fig.6 depicts simulation results of CT. CT of ABNRTC technique is reduced when compared to existing methods. This is due to the application of ensemble classifier. An ensemble of Adaboost with nonlinear regression tree classifier improves data packet CA with minimum time. The base classifiers are combined based on weight value to classify the data packets. Let us consider, the number of input data packet is 9, the CT of ABNRTC technique is 11ms whereas 15ms, 18ms, 20ms and 22ms CT of existing [1] - [4] respectively. Therefore, the ABNRTC technique takes minimum time for classification. As a result, CT is significantly reduced by 16%, 23%, 30% and 36% as compared to existing [1] - [4] respectively.

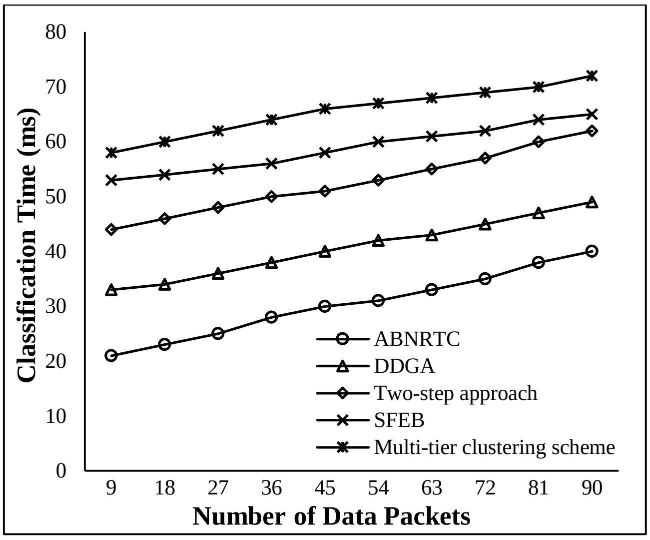

Fig.6. Simulation results of classification time

5. RELATED WORKS

Maximum Lifetime Data Aggregation Tree Scheduling (MLDATS) algorithm was introduced in [11]. However, it failed to provide efficient data aggregation. Multirate combine-skip-substitute (MR-CSS) scheme was introduced in [12]. But, the EC was more.

A cluster-rings approach was introduced in [13]. But the accuracy was not improved. In [14], an energy-efficient data collection approach was developed to enhance the NL. However, it failed to consider mobility of sensor nodes.

Ring-Based Correlation Data Routing (RBCDR) method was developed in [15]. Though the method improves NL but it failed to classify the data packets. A novel Data Routing for In-Network Aggregation, called DRINA was introduced in [16]. But, the correlation between aggregated data was not considered. A LEACH clustering algorithm was designed in [17]. But, the delay was higher. Robust and Dynamic data aggregation MAC protocol was developed in [18]. However, the data packet classification remained unaddressed. A closed form expression was determined in [19] for predicTable.EC of sensors using Greedy approach with less time constraints. However, the sink node has limited information regarding the sensor nodes.

An efficient subtree merging based data collection approach (SMDC) was introduced in [20] for minimizing EC in sensor networks. However, it failed to use environment monitoring application.

Quantized Compressed Sensing (QCS) and distributed compressive sensing theory (DCS) was developed in [21] for data aggregation in sensor networks. However, the data loss rate was higher.

6. CONCLUSION

An efficient ABNRTC is introduced for energy aware data gathering in WSN. The energy of each sensor node is measured to improve data gathering efficiency. The ABNRTC uses mobile sink node for aggregating the sensed data from the higher energy sensor nodes. This helps to reduce EC and improving the data gathering efficiency. After that, the mobile sink node performs data packet classification using AdaBoost Nonlinear Regression Tree classifier. Based on error value, the weight is assigned to weak classifier. The weak classifier classifies the data packet based on correlation. The classifiers with minimum weight are combined to construct strong classifier. The strong classifier efficiently classifies the data packets. Finally, the mobile sink node sends classified data packet to BS. ABNRTC technique increases the data gathering efficiency, NL and CA with minimum EC, delay, CT and FPR.

REFERENCES

[1] Yongmin Zhang, Shibo He and Jiming Chen, "Near Optimal Data Gathering in Rechargeable Sensor Networks with a Mobile Sink", *IEEE Transactions on Mobile Computing*, Vol. 16, No. 6, pp. 1718-1729, 2017.

[2] Cong Wang, Songtao Guo, and Yuanyuan Yang, "An Optimization Framework for Mobile Data Collection in Energy-Harvesting Wireless Sensor Networks", *IEEE Transactions on Mobile Computing*, Vol. 15, No. 12, pp. 2969-2986, 2016.

[3] Chih Min Chao and Tzu Ying Hsiao, "Design of Structure-Free and Energy-Balanced Data Aggregation in Wireless Sensor Networks", *Journal of Network and Computer Applications*, Vol. 37, No. 2, pp. 229-239, 2014.

[4] Jin Tao Meng, Jian Rui Yuan, Sheng Zhong Feng and Yan Jie Wei, "An Energy Efficient Clustering Scheme for Data Aggregation in Wireless Sensor Networks", *Journal of Computer Science and Technology*, Vol. 28, No. 3, pp. 554-557, 2013.

[5] Shusen Yang, Usman Adeel, Yad Tahir and Julie A. McCann, "Practical Opportunistic Data Collection in Wireless Sensor Networks with Mobile Sinks", *IEEE Transactions on Mobile Computing*, Vol. 16, No. 5, pp. 1420-1433, 2017.

[6] Chuan Zhu, Shuai Wu, Guangjie Han, Lei Shu and Hongyi Wu, "A Tree-Cluster-Based Data-Gathering Algorithm for Industrial WSNs With a Mobile Sink", *IEEE Access*, Vol. 3, pp. 381-396, 2015.

[7] Ping Zhong, Ya-Ting Li, Wei-Rong Liu, Gui-Hua Duan, Ying-Wen Chen and Neal Xiong, "Joint Mobile Data Collection and Wireless Energy Transfer in Wireless Rechargeable Sensor Networks", *Sensors*, Vol. 17, No. 8, pp. 1-23, 2017.

[8] Prabhudutta Mohanty and Manas Ranjan Kabat, "Energy Efficient Structure-Free Data Aggregation and Delivery in WSN", *Egyptian Informatics Journal*, Vol. 17, No. 3, pp. 273-284, 2016.

[9] Soobin Lee and Howon Lee, "Energy-Efficient Data Gathering Scheme Based on Broadcast Transmissions in Wireless Sensor Networks", *The Scientific World Journal*, Vol. 2013, pp. 1-17, 2013.

[10] Shouling Ji, Raheem Beyah and Zhipeng Cai, "Snapshot and Continuous Data Collection in Probabilistic Wireless Sensor Networks", *IEEE Transactions on Mobile Computing*, Vol. 13, No. 3, pp. 626-637, 2014.

[11] Ngoc-Tu Nguyena, Bing-Hong Liu, Van-Trung Pham and Yi-Sheng Luo, "On Maximizing the Lifetime for Data Aggregation in Wireless Sensor Networks using Virtual Data Aggregation Trees", *Computer Networks*, Vol. 105, pp. 99-110, 2016.

[12] Liang He, Jianping Pan and Jingdong Xu, "A Progressive Approach to Reducing Data Collection Latency in Wireless Sensor Networks with Mobile Elements", *IEEE Transactions on Mobile Computing*, Vol. 12, No. 7, pp. 1308-1320, 2013.

[13] Soo Hoon Moona, Sunju Park and Seung-jae Han, "Energy Efficient Data Collection in Sink-Centric Wireless Sensor Networks: A Cluster-Ring Approach", *Computer Communications*, Vol. 101, pp. 12-25, 2017.

[14] Ilkyu Ha, Mamurjon Djuraev and Byoungchul Ahn, "An Energy-Efficient Data Collection Method for Wireless Multimedia Sensor Networks", *International Journal of Distributed Sensor Networks*, Vol. 2014, pp. 1-8, 2014.

[15] Liangshan Jiang, Anfeng Liu, Yanling Hu and Zhigang Chen, "Lifetime Maximization through Dynamic Ring-based Routing Scheme for Correlated Data Collecting in WSNs", *Computers and Electrical Engineering*, Vol. 41, pp. 191-215, 2015.

[16] Leandro Aparecido Villas, Azzedine Boukerche, Heitor Soares Ramos, Horacio A.B. Fernandes De Oliveira, Regina Borges de Araujo and Antonio Alfredo Ferreira Loureiro, "DRINA: A Lightweight and Reliable Routing Approach for In-Network Aggregation in Wireless Sensor Networks", *IEEE Transactions on Computers*, Vol. 62, No. 4, pp. 676-689, 2013.

[17] Moulood Heidari, Akbar Nooriemamzade and Hamid Reza Naji, "Effect of using Mobile Sink and Clustering on Energy Consumption in Wireless Sensor Networks", *Turkish Journal of Electrical Engineering and Computer Sciences*, Vol. 23, pp. 1-16, 2015.

[18] Weigang Wu, Jiannong Cao, Hejun Wu and Jingjing Li, "Robust and Dynamic Data Aggregation in Wireless Sensor Networks: A Cross-Layer Approach", *Computer Networks*, Vol. 57, pp. 3929-3940, 2013.

[19] Mihaela Mitici, Jasper Goseling, Maurits de Graaf and Richard J. Boucherie, "Energy-Efficient Data Collection in Wireless Sensor Networks with Time Constraints", *Performance Evaluation*, Vol. 102, pp. 34-52, 2016.

[20] Yu Lasheng, Li Jie and Liu Renjie, "An Effective Data Collection Algorithm for Wireless Sensor Network", *Computing*, Vol. 95, No. 9, pp. 723-738, 2013.

[21] Wei Wang, Dan Wang and Yu Jiang, "Energy Efficient Distributed Compressed Data Gathering for Sensor Networks", *Ad Hoc Networks*, Vol. 58, pp. 112-118, 2017.