Resource Optimized Ensemble Gradient Boosting Classifier for Traffic Aware Big Data Analytics

C.R. Durga Devi, Ph.D Research Scholar, Department of Computer Science, NGM College, Pollachi, India.

E-mail:deviswe@gmail.com

Dr.R. Manicka Chezian, Associate Professor, Department of Computer Science, NGM College, Pollachi, India.

E-mail:chezian_r@yahoo.co.in

Abstract--- Big data analytics is the process of vast quantity of data created and collected in a geographically distributed manner. Groups of data stored and processed in a single datacenter is inefficient due to several limitations such as wider bandwidth, high traffic level, high storage capacity and more task completion time of big data processing. As the quantity of data increases, it provides less efficient to handle large volume of data in a multiple datacenter with available resource utilization. A new optimization framework reduces the inter datacenter traffic of MapReduce jobs during geo-distributed big data. However, its needs to be utilize the optimal resources while reducing the traffic over geo-dispersed locations. To reduce the traffic occurrence level during the big data distribution in cloud environment, Resource Optimized Traffic Aware Gradient Boosting Classification (ROTAGBC) Technique is developed. In cloud, number of incoming tasks is send from the users. The task assigner assigns the priority to incoming tasks based on certain request parameters task size, starting time, finishing time, bandwidth, memory capacity. Based on priority level, Gradient Boosting Classification method is functional to classify the tasks based on their priority level. The resource optimized gradient boosting classifier constructs a prediction model in structure of an ensemble of weak decision trees prediction. The gradient boost ensemble classifier categorizes number of tasks either immediate or reserved for distributing to a multiple datacenters in cloud. The gradient boost data classifier combines all base classifier into strong classifier to present final results with higher classification accuracy. Experimental assessment is carried out on the different factors such as classification accuracy, space complexity, task completion time and resource utilization rate with respect to number of tasks. The results showed that the presented ROTAGBC technique is better in case of classification accuracy, space complexity, task completion time and resource utilization rate. Based on the observations, ROTAGBC technique is more efficient than the other existing methods for geo-data distribution over multiple data centers.

Keywords--- Big Data Analytics, Cloud, Geo-Distributed Big Data, Datacenters, Priority, Gradient Boosting Classification, Resource Optimization, Task Distribution.

I. Introduction

In cloud, big data analytics is the process of large volume of data with multiple and independent sources. By means of cloud computing technology, a number of geo data are distributed effectively and capably to maintain different aspects of problem such as traffic and resource optimization. Classification of incoming tasks is widely used for reducing the traffic occurrences while distributing large volume of data over the data center in different locations. With the appearance of large volume of datasets, the conventional approaches failed to generate desirable results in terms of traffic as well as resource optimization. Therefore, a challenges lies in the big cloud data analytics are optimal resource utilization for geo data distribution across multiple data centers.

Chance-constrained optimization technique was developed in [1] for reducing the inter-DC traffic created through MapReduce jobs on geo-distributed big data. However, it failed to minimize energy consumption while distributing the geo data with less space and time complexity. The Lyapunov framework was developed in [2] for processing cost aware big data analytics namely process cost, storage cost, bandwidth cost, latency lost and migration cost across geo-distributed datacenters. However, the cost of energy was not addressed as well as classification was not performed for efficient big data analytics. Conceptual framework based big data analytics (CF4BDA) was developed in [3] to implement Cloud-based BDA applications. But, traffic aware big data analytics was not performed. In recent days, big data analytics is the process of organizing and analyzing data to obtain useful information in cloud services. In [4], a various typical methods were developed for wide area analytics with

Received: 5 Feb 2018/Accepted: 15 Mar 2018

geographically distributed data centers. But, energy optimization was not performed during big Data analytics. A two-dimensional markov chain was developed in [5] for considering both data transmission and cost minimization problem in Geo-Distributed Data Centers. But, the classification of the data was not performed in order to reduce the traffic level. The distributed data analytics approach was introduced in [6] for handling large and high volume synchrophasor data in wide-area measurement systems. However, the application of data analytics was not easily developed with simple configuration. A flexible data analytics (FlexAnalytics) framework was developed in [7] to analyze a number of potential data-analytics along the I/O path. However, data distribution across multiple data centers remained unsolved. An effective design and performance of G-Hadoop, a MapReduce framework was introduced in [8] to allow large scale data distributed across multiple clusters. But, it failed to handle complex cloud resources. In [9], a MapReduce-based framework was developed to distribute the data through a cluster of computing elements. However, resource utilization during the data distribution remained unsolved. A linguistic fuzzy rule-based classification system (Chi-FRBCS) was developed in [10] for big data classification. However, it failed to use base classifier in a MapReduce scheme.

The certain issues are identified from above said existing methods such as lack of classification, more space and time complexity, lack of resource utilization, failed to perform data storage, high network traffic and so on. In order to overcome such kind of issue, Resource Optimized Traffic Aware Gradient Boosting Classification (ROTAGBC) technique is developed. The major contribution of the paper is described as follows.

- Resource Optimized Traffic Aware Gradient Boosting Classification (ROTAGBC) technique is introduced for traffic aware geo distributed big data analytics. Initially, the task assigner performs priority task classification of incoming tasks using gradient Boosting ensemble classifier. The priority is assigned to each incoming tasks based on certain parameters. Based on priority assignments, task assigner performs efficient task classification for reducing the traffic occurrences level.
- Gradient Boosting ensemble classifier uses decision tree as base classifier. The decision tree is constructed with number of nodes and it classified the number of incoming tasks. These base learners are combined to make a strong classifier which classifies the number of tasks as either immediate or reserved. This in turn improves classification accuracy.
- Task assigner distributing the classified tasks over the multiple data centers at different locations with minimum task completion time and optimal resource utilization. This helps to reduce the traffic occurrence level. Finally, the required services are received by end users in an effective manner. This helps to improve resource utilization rate during the data distribution.

The paper is organized as follows, Section 2 introduces reviews on related works for big data distribution. section 3 briefly discuss ROTAGBC technique with neat diagram. Section 4 discusses the experimental evaluation. Section 5 presents result analysis of ROTAGBC technique with various parameters. Concluding remark of the paper is presented in Section 6.

II. Related Work

A linguistic cost-sensitive fuzzy rule-based classification technique was developed in [11] for handling imbalanced big data obtaining higher accuracy with minimum processing time. However, it failed to provide the best solution when dealing with big data. ROTAGBC technique provides accurate priority task classification results during big data analytics.

A system-level stability evaluation approach (SDSR) was introduced in [12] with reliable energy function and avoids computation complexity. But, traffic aware big data analytics was not performed. Therefore, an ROTAGBC technique is introduced for big data analytics with less traffic occurrences level using gradient boost ensemble classifier.

Big Data Analytics for dynamic energy management were developed in [13] used smart grid data mining with accurate and efficient power consumption. But, it failed to handle the real-time monitoring system with optimal resource utilization. ROTAGBC technique effectively improves the big data analytics with efficient resource utilization for distributing the incoming tasks to multiple datacenters. An efficient random forest classifier was introduced in [14] to deal with imbalanced datasets in the big data circumstances. However, the performance of classification was not increased at a required level. Therefore, an ROTAGBC technique improves the priority based classification accuracy using gradient boosting ensemble classifier. A Round Robin Load Balancer Algorithm and Throttled load balancer algorithm was designed in [15] to improve data distribution with minimum processing time. But it has more energy for data distributions. ROTAGBC technique consumes minimum energy utilized for

Received: 5 Feb 2018/Accepted: 15 Mar 2018

distributing the big data in cloud. An online lazy migration (OLM) algorithm and randomized fixed horizon control (RFHC) algorithm was developed in [16] for geo-dispersed big data aggregation and processing as well transmission. But, it considered worst case computational complexity. ROTAGBC technique minimizes both space and time complexity during task distribution in cloud.

A hybrid electrical and optical networking structural design was presented in [17] for data centers hosting cloud computing and big data applications. However, an efficient classification was not performed. Therefore, an ROTAGBC technique performs efficient classification using ensemble classifier to distribute number of incoming tasks across multiple data centers. Energy-aware cloud computing data centers according to the workload distribution was presented in [18]. However, the performance of network storage was not performed. ROTAGBC technique reduces the space complexity during the task distribution. An energy-efficient task scheduling technique was developed in [19] for scheduling the tasks to different datacenter in cloud. But, traffic aware data distributions remained unaddressed. An ROTAGBC technique performs efficient tasks distributions with minimum traffic level. An ICP: Data Mining Package tool was introduced in [20] to perform classification processes on large volume of data. This tool includes four classification algorithms namely Decision Trees, Naïve Bayes, Random Forest and Support Vector Machines (SVM). However, resource optimized classification was not performed. Therefore, an efficient ROTAGBC technique introduces a resource optimized classifier for big data analytics.

As a result, Resource Optimized Traffic Aware Gradient Boosting Classification (ROTAGBC) technique reduces the traffic occurrence level with minimal energy consumption for geo-data distribution over multiple data centre.

III. Methodology

Resource Optimized Traffic Aware Gradient Boosting Classifier for Big Data Analytics

The architecture diagram of ROTAGBC technique is shown in figure 1.

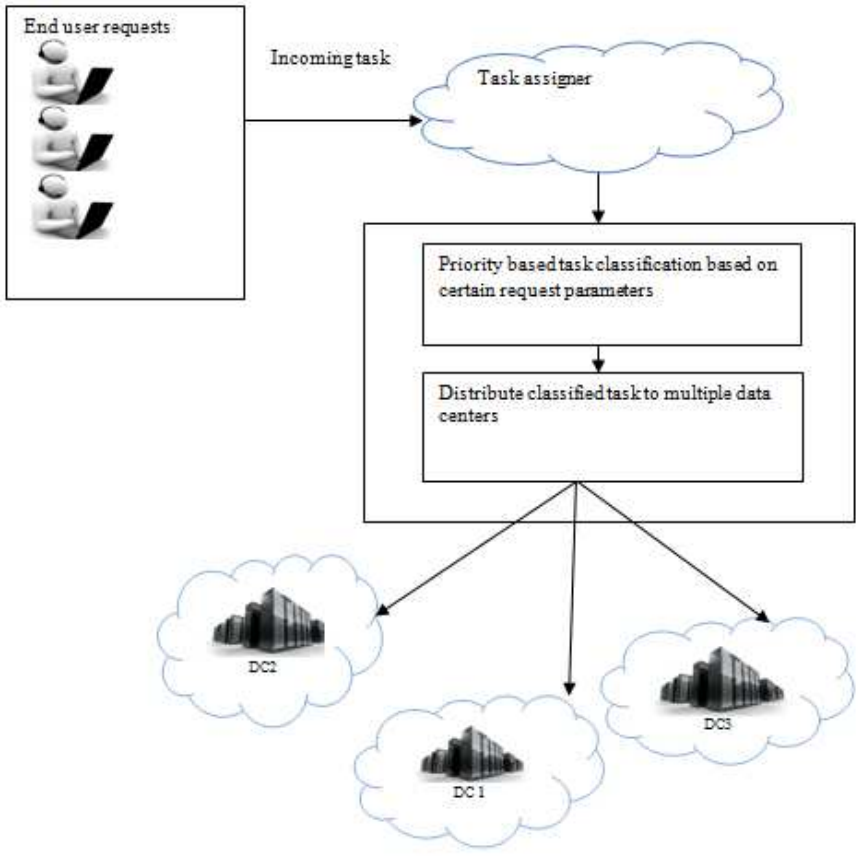

Figure 1: Architecture Diagram of Resource Optimized Traffic Aware Gradient Boosting Classification Technique

Received: 5 Feb 2018/Accepted: 15 Mar 2018

In cloud, big data analytics is the strategy for analyzing large volume of data. To analyze such a large degree of data, big data analytics is performed to reduce the traffic occurrence level with minimal energy consumption for geo-data distribution over multiple data center.

Cloud architecture consists of multiple data centers which are the significant resources to provide various services through virtual machines connected to the end user. These data centers (DCs) are located at different regions.

But the rapid increment of different global services, there is a significant need for big data analytics across multiple data centers (DCs) located in various regions. During the geo data distribution, traffic refers to the amount of data moving across a cloud data center at a specified point of time.

While handling the traffic across several data center in a cloud, end users' service requests are considered as job (i.e. task) and it is allocated to a virtual machine (VM).

During the data distribution, typical job parameters of virtual machine are number of cores required, CPU, memory, and bandwidth of a system required to perform the task in cloud. However, due to increasing demand of cloud computing, number of tasks affects the system load and performance.

Therefore, an efficient machine learning technique namely Resource Optimized Traffic Aware Gradient Boosting Classification (ROTAGBC) is developed to efficiently solve the formulated data center traffic minimization problem during the data distribution

Figure 1 shows architecture diagram of ROGBDC technique with the objective of reducing traffic occurrences level and resource optimization during big data analytics for distributing the number of incoming tasks into various data centers at different locations.

In the proposed ROGBDC technique, the hosts are considered as servers that include computational power where the VMs are deployed. ROGBDC technique considers the data center that contains 'n' number of hosts (i.e. server).

Each host contains number of cores which are assigned to the VMs based on its types. The incoming tasks are scheduled to these VMs and it completes the job in parallel on a host with different finishing time. In ROGBDC technique, the processing of multiple users requests are considered as tasks. Next, the incoming tasks are received by the task assigner.

Followed by, the task assigner perform priority based task classification using gradient boosting ensemble classification technique in order to reduce the traffic among the multiple data centers in cloud. The priority is assigned to the task based on certain job request parameters are task size (i.e. file size), starting time, finishing time, bandwidth, memory capacity.

Based on priority level, the task is classified as immediate or reserved. Finally, the classified tasks are distributed over the multiple data center for providing efficient services to the end user. This helps to reduce traffic occurrence and resource optimization for geo-data distribution over multiple data centre in big data analytics. The brief explanation about the ROGBDC technique is presented in forth coming sections.

Priority based Gradient Boosting Task Classification

The task assigner starts the classification of incoming tasks for multiple geo-distributed datacenters. In cloud big data analytics, number of users generates a different tasks having various size, different deadline (i.e. ending time), arrived time, sending time.

Based on task and their time constrains, an amount of memory utilization, bandwidth utilization of the processor is very important during the task distribution over the multiple data centers in cloud.

In order to obtain an efficient Priority task scheduling approach, the priority tasks classification is most significant in cloud. Therefore, the proposed ROGBDC technique performs efficient task classification based on their priority assignments. The block diagram of priority based gradient boosting task classification is shown in figure 2.

Received: 5 Feb 2018/Accepted: 15 Mar 2018

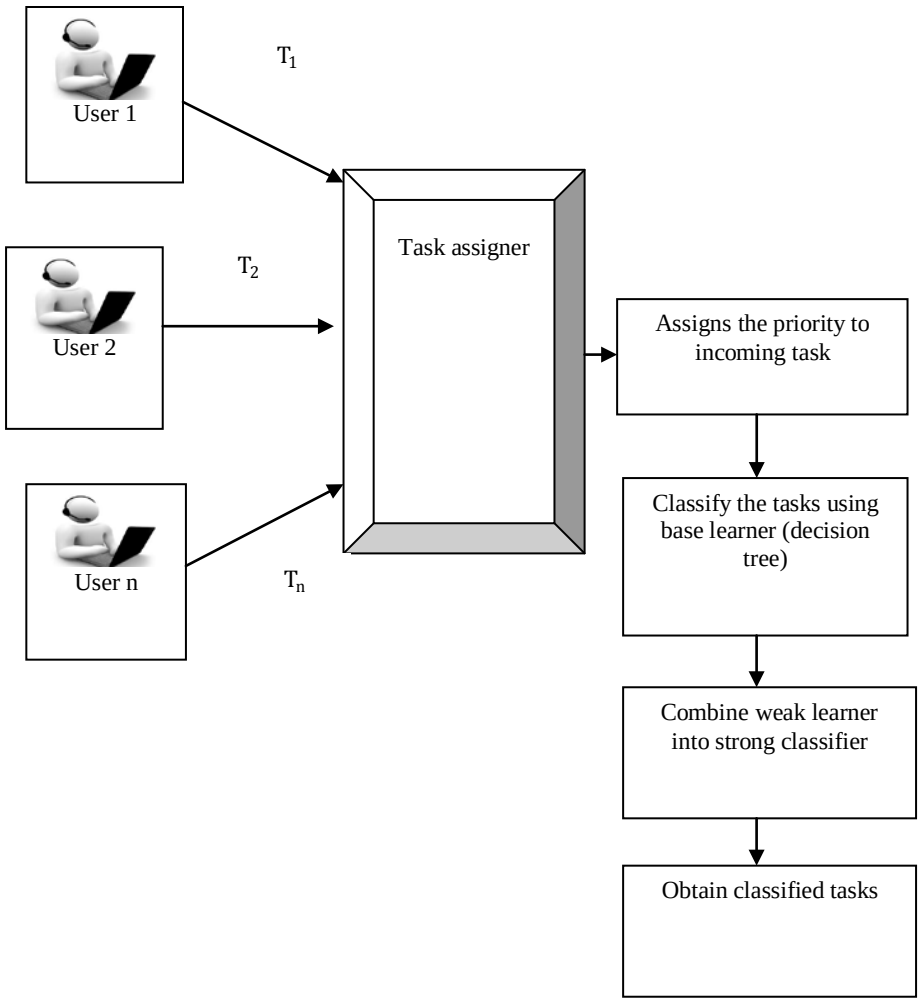

Figure 2: Block Diagram of Priority based Gradient Boosting Task Classifications

Figure 2 shows the priority based Gradient Boosting task Classifications with number of user tasks to improve classification accuracy. Let number of users is denoted as $U = u_1, u_2, .., u_n$ who generates a request as tasks$T_1, T_2, T_3, \ldots . T_n$. In ROGBDC technique, the number of user requested tasks is send to the task assigner for performing effective classification. Therefore, the classifications of tasks are performed based on the priority assignments. The certain job request parameters are requested file size, task size (i.e. file size), starting time, finishing time, bandwidth, and memory capacity. The file size is a measure of how much task a system contains as well as storage. T_S Denotes a Starting time and T_f represents a finishing time, BW is the bandwidth and is defined as the time series of a file transfer. Usually a time series is specified in seconds which is measured as follows,

$$BW = \frac{Maximum \;\; amount \;\; of file \;\; transfer}{Time \;\; in \;\; seconds} \quad (1)$$

From (1), BW denotes a bandwidth SC is represents a storage capacity and it is measured in megabytes (MB). Based on above said parameters, the tasks are prioritized. The task which has less dead line (i.e. finishing time), less size, less bandwidth task utilization, and less storage capacity reside at first priority. Similarly, tasks having highest deadline (i.e. more finishing time), highest length or size which utilizes larger bandwidth and storage capacity which resides at the low priority. Based on the priority assignments, the incoming tasks are classified as immediate task or reserved task using Gradient Boosting classifier. Gradient boosting is a machine learning technique used for classification and prediction purposes. Gradient boosting classifier is used when the incoming tasks are largely populated. It is also an ensemble classifier which means combines the entire weak learner to make a strong classifier and provides the final classification results. In ROGBDC technique, decision tree is used as a weak classifier to construct a strong classifier using gradient boosting technique. The big data is a term that contains a large volume of tasks in cloud environment.

Received: 5 Feb 2018/Accepted: 15 Mar 2018

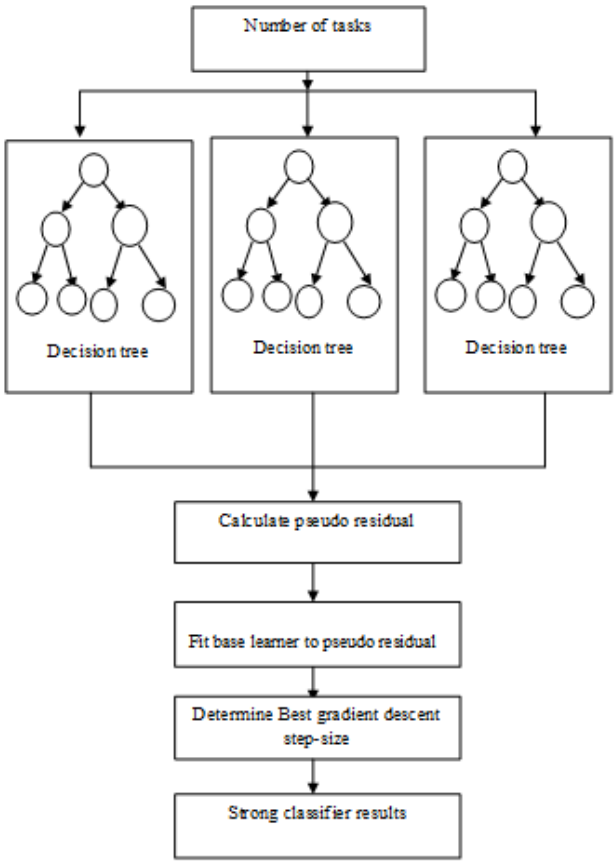

Figure 3: Flow Processing Diagram of Gradient Boosting Classification

As shown in figure 3, flow processing diagram of gradient boosting classification with decision tree classifier is described. From the figure, ROGBDC technique considers number of tasks as input for performing the classification. Initially, decision tree classification for task is measured based on priority assignments. A Decision tree is constructed with root nodes, branch nodes, and leaf nodes. Each internal node represents a number of tasks and the branch node indicates a result of decision tree classification. The leaf node of decision tree represents a class labels.

Let us consider training sets of prediction model is represented as $(X_1, Y_1), (X_2, Y_2) \ldots . (X_n, Y_n)$ Here $X_1, X_2, X_3 \ldots . . X_n$ represents a number of tasks whereas $Y_1, Y_2, Y_3 \ldots . . Y_n$ denotes a prediction output results and it is fit into the loss function $F_l(X)$. Therefore, the prediction results of the classifier is described as follows,

$$Y_i = F_l(X) + H_{mn}(X_i) \quad (2)$$

From (2), $F_l(X)$ represents a loss function of base learner (i.e. decision tree) where $h_m(X_i)$ denotes a decision tree classifier output. By applying gradient boost classification, loss function is defined as the squared error of difference between actual and predicted value. Therefore, the loss function is expressed as follows,

$$F_l(X) = \left(Y_i - H_m(X_i)\right)^2 \ (3)$$

From (3),Y_i represents a actual output value and$h_m(X_i)$ predicted output value of base classifier. Therefore, the output of the weighted sum functions of all the base decision tree classifiers used to construct a strong classifier. Therefore, a strong classifier output $H_m(X_i)$ is formulated as follows

$$H_{mn}(X) = h_{m1}(X) + h_{m2}(X) + \cdots + h_{mn}(X) \ (4)$$

From (4),$h_{m1}(X)$ denotes a first classifier output, $h_{m2}(X)$ denotes an output of second classifier and $h_{mn}(X)$ represents an output of 'n' classifier. After that, base learner output $h_{mn}(x)$ is fit into pseudo-residuals with training data.

The pseudo-residuals (R_s) function is calculated as,

Received: 5 Feb 2018/Accepted: 15 Mar 2018

$$R_s = -\left[\frac{\partial(Y_i, F_l(X))}{\partial F_l(X_i)}\right] \qquad where\ i = 1,2,3 \ldots n \quad (5)$$

From (5) $'R_s'$ represents a pseudo-residuals. Then fit a base decision tree to an input training data. Therefore, output of the strong predictive classifier $H_{mn}(X)$ with the number of tasks are measured the sum of individual base classifier output. It is measured as follows,

$$H_{mn}(X) = \sum_{i=1}^{n}\{X_i, (Y_i - F_l(X_i)\} \quad (6)$$

The output of individual classifier is described as follows,

$$h_{m1}(X) = (X_1, (Y_1 - F_l(X_1)) \quad (7)$$

Similarly, the second weak classifier is formulated as follows,

$$h_{m2}(X) = (X_2, (Y_2 - F_l(X_2)) \quad (8)$$

As a result, the final weak classifier $h_{mn}(X)$ is formulated as follows,

$$h_{mn}(X) = (X_n, (Y_n - F_l(X_n)) \quad (9)$$

The gradient boosting classifier combines a weak decision tree classifier output. The best gradient output is used for classifying the task. The best gradient descent step-size (ρ_B) is determined by using following mathematical equations,

$$\rho_B = arg\min_{\rho} \sum_{i=1}^{n}[Y, F_{l-1}(X_i) + \rho H_{mn}(X)] \quad (10)$$

Finally, update the predictive model to classify the tasks based on priority assignments,

$$Y_i = \sum_{i=1}^{n} F_{l-1}(X_i) + \rho_B H_{mn}(X) \qquad (11)$$

Equation (11) $'Y_i'$ denotes a target strong classifier output. As a result, a strong classifier output is used to classify the tasks as immediate or reserved. The final output of the strong classifier provides a positive result (i.e. '1') indicates the user requested task is classified as immediate task which has high priority. Otherwise, it is said to be a reserved tasks which has low priority. Therefore, an ensemble of weak classifier is used to make strong classifiers and performs the better classification in order to improve classification accuracy.

Input : Number of user tasks $T_1, T_2, T_2, \ldots . T_n$

Output : Improved classification accuracy

Step 1: **Begin**

Step 2: **For** each incoming tasks

Step 3: Task assigner assigns the priority based on file size, start time, task finishing time, bandwidth utilization, storage capacity

Step 4: Construct decision tree as base learner

Step 5: Measure loss function $F_l(X)$ of base classifier using (3)

Step 6: Calculate pseudo-residuals function (R_s)using (5)

Step 7: Fit a base classifier $h_{mn}(X)$ to pseudo-residuals with number of tasks using (6)

Step 8: Find the best gradient descent step-size (ρ_B) using (10)

Step 9: Update the model as strong classifier using (11)

Step 10: **If** (Y_i results is '1' **) then**

Step 11: The task is classified as immediate

Step 12: **else**

Step 13: The task is classified as reserved

Step 14: End if

Step 15: End for

Step 16: End

Algorithm 1 Priority based Gradient Boosting Task Classification

Algorithm 1 describes a Priority based gradient boosting task classification with decision tree for efficient task classification to distribute the number of tasks over multiple data center in cloud. For each incoming task, priority is assigned based on the certain job request parameters. After that, the decision tree is constructed for performing the

Received: 5 Feb 2018/Accepted: 15 Mar 2018

classification based on priority assignments. Subsequently, the gradient boosting classifier measures loss function of base decision classifier in order to increase classification accuracy. After that, the output of base decision tree classifier is fit into pseudo-residuals. The best gradient descent step size is determined to create a strong classifier. Finally, the output of strong classifier provides the better priority task classification results. This helps to improve the classification accuracy.

Traffic Aware Data Distribution in Cloud

After classifying the tasks, the task assigner distributes the tasks to multiple data centers according to their priority. In ROGBDC technique, task with high priority is served before a task with low priority. Based on the priority task classification, the immediate task has high priority than the other reserved tasks. Therefore, the high priority tasks are handled first than the low priority tasks. The task assigner checks if the resources (i.e. energy, bandwidth and memory) are available in a host at a requested time for handling the immediate tasks and then the task is allocated to that particular datacenter. This helps to reduce the traffic among the data centers in cloud environments. Traffic refers to a number of tasks distributed across a multiple data center at a specified point of time. As a result, the load is increased during the task distribution and the task completion time gets increased. This increases the more space and time complexity in data distribution. In order to reduce traffic occurrences during task distribution over the data centers, the incoming tasks are classified and send to data center. Thus reduces the workload among the multiple data centers in cloud. As a result, resource utilization and traffic data distribution are obtained. The data distribution is shown in figure 4

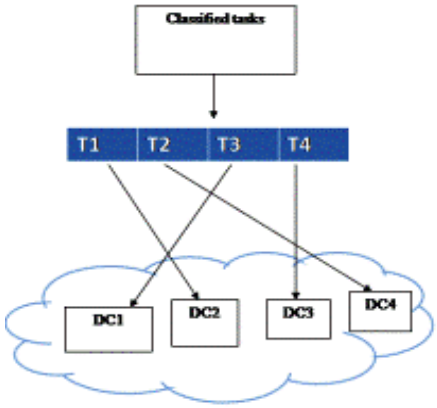

Figure 4: Traffic Aware Task Distributions

Figure 4 shows the Traffic aware task distributions. The classified tasks distribute over a multiple data centers at different locations in cloud. A datacenter has number of hosts (i.e. servers) and the ROGBDC technique shows the host name of the node executing this operation. From the figure, let us consider T_1 is immediate tasks and the task assigner distributed to that task to the datacenter 2 (DC2).

Followed by, the reserved and other immediate tasks are allocated into a datacenter with efficient resource is available in a host. This helps to reduce the traffic occurrences and optimal resource utilization across the multiple data centers. Therefore, energy consumption is used to calculate the energy efficiency for tasks distribution across multiple data center. From the data center, the end user receives specified services after job completion. This helps to minimize the time complexity and improve resource utilization rate for traffic aware data distribution across multiple datacenter in cloud computing.

IV. Experimental Settings

The proposed Resource Optimized Traffic Aware Gradient Boosting Classification (ROTAGBC) technique is experimented using java language and Cloudsim simulator used for implementation. The experimental evaluation is conducted for big data analytics using Personal Cloud Datasets (http://cloudspaces.eu/results/datasets). This dataset is used for Active Personal Cloud Measurement. It takes two arguments such as provider and test type. Based on test type, the script executes one of the following files namely load_ and_ transfer and service _variability. The objective

Received: 5 Feb 2018/Accepted: 15 Mar 2018

of file transfer is continuously performed at each node based on file size. By using Personal Cloud Datasets, some measurement traces are collected. The dataset contains 17 columns field to perform load and transfer test. The columns are row_id, account_id, file size, operation_time_start, operation_time_end, time zone (not used), operation_id, operation type, bandwidth trace, node_ip, node_name, quoto_start, quoto_end, quoto_total (storage capacity), capped (not used), failed and failure info. Based on the above column fields, some of the fields are file size, operation_time_start, operation_time_end, bandwidth trace, node_ip, node_name, quoto_total (storage capacity) taken for perform resource optimized load transfer to multiple datacenter in cloud.

Based on the above said parameters the incoming user requested tasks are classified and distributes a workload among multiple datacenter for reducing the traffic occurrences in cloud environments. Experimental evaluation ROTAGBC Technique is conducted on different factors such as classification accuracy, space complexity and task completion time and resource utilization rate.

V. Results and Discussion

Results and discussion of ROTAGBC technique is described in this section. The ROTAGBC technique compared against existing Chance-constrained optimization technique [1] Lyapunov framework [2]. The performance analysis is conducted on the various parameters such as classification accuracy, space complexity and task completion time and resource utilization rate with respect to number of incoming tasks in cloud. Experimental results are analyzed with the help of following table and graph.

Impact of Classification Accuracy

Classification accuracy is defined as the ratio of number of incoming tasks is correctly classified to the total number of tasks. It is measured as follows,

$$CA = \frac{No.of\ tasks\ that\ are\ correctly\ \ classified}{n} * 100 \quad (12)$$

From (12), Where CA denotes classification accuracy and 'n' denotes a number of tasks. Classification accuracy is measured in terms of percentage (%).

Table 1: Tabulation for Classification Accuracy

Number of tasks	Classification accuracy (%)		
	ROTAGBC	Chance-constrained optimization technique	Lyapunov framework
10	80	71	60
20	82	73	63
30	84	75	66
40	86	77	69
50	88	79	72
60	90	82	75
70	91	84	78
80	93	86	80
90	95	88	83
100	97	90	85

Table 1 describes experimental results of classification accuracy with respect to number tasks in cloud. The number of incoming tasks for experimental consideration is varied from 10 to 100. As shown in table value, classification accuracy is considerably improved than the other existing methods Chance-constrained optimization technique [1] Lyapunov framework [2]. This is because of ROTAGBC technique performs priority task classification using gradient boost ensemble classifier. The performance result of classification accuracy is shown in figure 5.

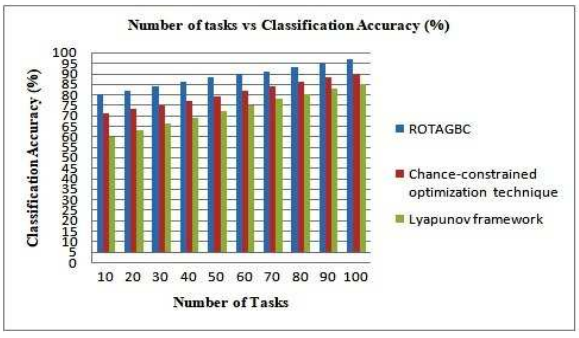

Figure 5: Performance Results of Classification Accuracy

Received: 5 Feb 2018/Accepted: 15 Mar 2018

Figure 5 shows the performance analysis of classification accuracy with respect to number of tasks. As shown in figure, the classification accuracy is increased using ROTAGBC technique compared to existing methods. This is because, ROTAGBC technique classifies the incoming tasks based on their priority level to obtain traffic aware data distribution to a multiple datacenter in cloud environments. The user requests are arrived in cloud and it is considered as tasks. The task assigner receives these incoming tasks and performs priority task classification. The priority is assigned to a tasks based on certain parameters task size, bandwidth, storage capacity, task dead line (i.e. finishing time) and starting time. The prioritized tasks are classified using gradient boost ensemble classification techniques. Initially, decision tree is used as a base learner which classifies the tasks as immediate or reserved tasks with minimum time based on their priority level. The base learners are combined and construct a strong classifier by measuring a loss function to increase classification accuracy. The output of base learner is fit into pseudo-residuals. Finally, gradient descent step size is determined to make a strong classifier in an efficient manner. This process is repeated until all the tasks are correctly classified for reducing the traffic occurrences among the multiple datacenters at different locations. As a result, classification accuracy of ROTAGBC technique is considerably increased by 10% and 22% when compared to existing Chance-constrained optimization technique [1] Lyapunov framework [2] respectively.

Impact of Space Complexity

Space complexity is defined as an amount of storage capacity required to store the classified tasks. The space complexity is measured in terms of Mega bytes (MB). The space complexity is measured as follows,

$$SC = n * space(storing\ classified\ task) \quad (13)$$

From (13) 'SC' denotes a space complexity. Lower the space complexity, more efficient the method is said to be.

Table 2: Tabulation for Space Complexity

Number of tasks	Space complexity (MB)		
	ROTAGBC	Chance-constrained optimization technique	Lyapunov framework
10	15	25	33
20	18	28	37
30	23	33	42
40	29	35	48
50	32	38	52
60	36	43	55
70	40	48	57
80	42	53	62
90	48	58	65
100	50	62	68

Table 2 describes a performance result of space complexity with respect to number of incoming user tasks. The space complexity is an amount of memory space required for an algorithm to store the classified tasks. From the table value, the results of space complexity using ROTAGBC technique is significantly reduced when compared to existing Chance-constrained optimization technique [1] Lyapunov framework [2] respectively. Performance resultant graph of space complexity is shown in figure 6.

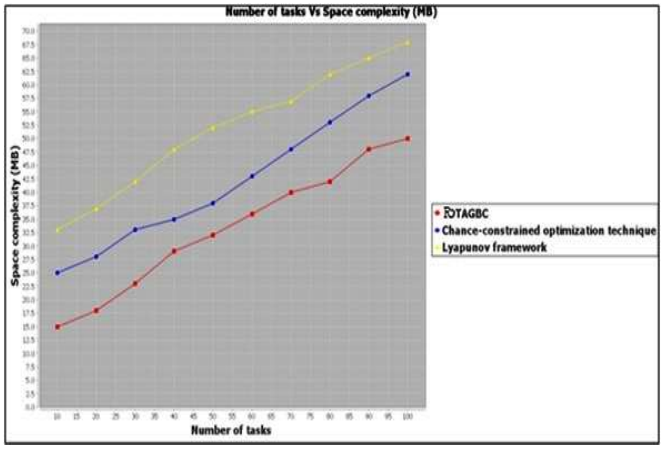

Figure 6: Performance Results of Space Complexity

Received: 5 Feb 2018/Accepted: 15 Mar 2018

Figure 6 depicts the performance results of space complexity versus the number of tasks distributed in multiple data centers in cloud. From the figure, it is clearly evident that the space complexity is reduced using ROTAGBC technique than the existing methods. The proposed ROTAGBC technique has number of tasks as input and each task has various sizes. The storage capacity is significant parameters in the cloud for obtaining efficient job scheduling. Therefore, the storage capacity is reduced by performing the priority task classification. Gradient boosting ensemble classifier is used for classifying the number of user tasks either immediate or reserved. The immediate tasks having high priority whereas reserved tasks have low priority. Then these classified tasks are stored for distributing a different datacenter to minimize the traffic and workload. Therefore, the ROTAGBC technique reduces the space complexity in traffic aware data distribution. Let us consider, number of tasks are 10, an amount of storage space for ROTAGBC technique is 15MB, whereas 25MB and 33MB using Chance-constrained optimization technique [1] Lyapunov framework [2] respectively. This shows ROTAGBC technique utilizes the minimum storage space for storing the different tasks. As a result, space complexity of ROTAGBC technique is significantly reduced by 23% and 38% when compared to existing Chance-constrained optimization technique [1] Lyapunov framework [2] respectively.

Impact of Task Completion Time

Task completion time is defined as an amount of time required to distribute an incoming task into a particular data center with minimum traffic occurrences. The mathematical formula for task completion time is described as follows,

$$TCT = n * time(distributing\ a\ task) \quad (14)$$

From (14), where 'TCT denotes a task completion time and it is measured in terms of milliseconds (ms).

Table 3: Tabulation for Task Completion Time

Number of tasks	Task completion time (ms)		
	ROTAGBC	Chance-constrained optimization technique	Lyapunov framework
10	25	35	42
20	31	39	47
30	36	42	53
40	40	48	58
50	45	53	62
60	50	58	66
70	52	62	70
80	55	66	74
90	58	70	78
100	63	74	82

The comparative performance analysis of task completion time is shown in table 3. An amount of time for distributing the number of incoming tasks to different center is reduced using ROTAGBC technique than the existing methods. Three different methods and their experiential results are presented in table 3. While considering the 10 user tasks, ROTAGBC technique uses the task completion time of 25ms whereas Chance-constrained optimization technique [1] Lyapunov framework [2] achieves 35 ms and 42 ms respectively. This shows the significant improvement of ROTAGBC technique than existing methods.

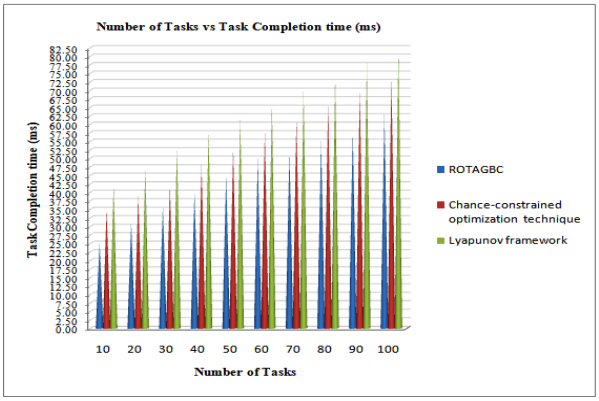

Figure 7: Performance Result of Task Completion Time

Received: 5 Feb 2018/Accepted: 15 Mar 2018

Figure 7 shows the performance results of task completion time versus number of tasks in cloud. As shown in figure, task completion time is considerably reduced using ROTAGBC technique than the existing methods. This is because, incoming tasks are defined in the form of tasks. These tasks are received by the task assigner in cloud environment. The task assigner allocates the priority to incoming tasks based on the certain parameters. The size of task is minimum and less finishing time, bandwidth utilization, storage capacity. Based on these constrains, the tasks are prioritized. After that, the gradient boosting classifier is applied to perform priority based task classification.

The gradient boost ensemble classifier provides the strong classification output results of incoming tasks through the weak decision tree classifier. Based on classification, the incoming tasks are identified as immediate or reserved. Then the task assigner distributes the high priority task (i.e. immediate task) to particular datacenter. This helps to minimize the time for task distribution across the multiple data centers in cloud with available resources. As a result, ROTAGBC technique improves the geo data distribution in an effective way with minimum task completion time. Therefore, task completion time is considerably reduced by 17% and 29% when compared to Chance-constrained optimization technique [1] Lyapunov framework [2] respectively.

Impact of Resource Utilization Rate

Resource utilization rate is measured as the ratio of number of resources utilized by task distribution over datacenters in cloud to the total available resources in cloud. The formula for resource utilization is expressed as follows,

$$RUR = \frac{cloud\ \ resources\ \ \ utilized}{Total\ \ available\ \ \ resources} * 100 \quad (15)$$

From (15), where $'RUR'$ denotes resource utilization rate and it is measured in terms of percentage (%).

Table 4: Tabulation for Resource Utilization Rate

Number of tasks	Resource utilization rate (%)		
	ROTAGBC	Chance-constrained optimization technique	Lyapunov framework
10	82	72	60
20	84	75	64
30	88	78	67
40	90	80	70
50	92	83	73
60	93	85	75
70	94	88	78
80	95	90	81
90	96	92	83
100	97	93	84

Table 4 describes experimental results of resource utilization rate for distributing the number of incoming tasks into a multiple datacenters at different locations. Resource utilization is a significant parameter technology for utilizing the available cloud resources to distribute the tasks to datacenters As a result, Resource utilization rate using proposed ROTAGBC technique is improved when compared to exiting Chance-constrained optimization technique [1] and Lyapunov framework [2] respectively.

Figure 8 depicts performance result of resource utilization rate versus number of tasks in the range of 10-100.

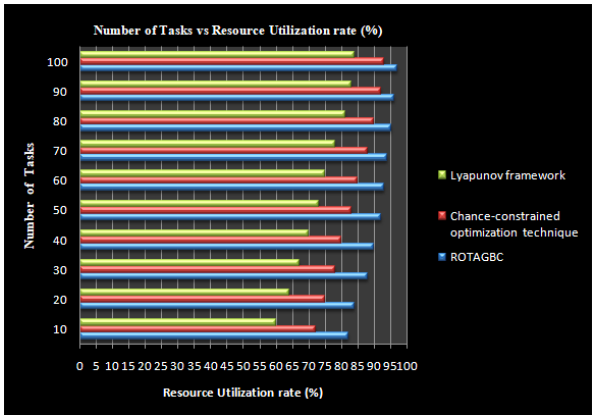

Figure 8: Performance Result of Resource Utilization Rate

Received: 5 Feb 2018/Accepted: 15 Mar 2018

As shown in figure, the proposed ROTAGBC technique increases the resource utilization rate when compared to existing methods. While varying the number of tasks, amount of resources utilized for computing the number of incoming tasks by utilizing a memory usage, energy consumption. This is because, ROTAGBC technique minimizes the amount of traffic while distributing the number of tasks to multiple data centers in cloud computing. This is done by the tasks which has higher priority is stored in queue for distributing the data center at a different locations. Moreover, the dynamic incoming tasks are scheduled to multiple datacenter with optimized resources for offering the services from cloud server. The ROTAGBC technique improves the resource utilization rate by 10% and 21 % when compared to Chance-constrained optimization technique [1] and Lyapunov framework [2] respectively.

VI. Conclusion

An efficient machine learning technique called Resource Optimized Traffic Aware Gradient Boosting Classification (ROTAGBC) is developed for geo-data distribution over multiple data centre with optimal resource utilization and minimum traffic level. Initially, the user requests are considered as tasks. Processing the number of tasks is often called as big data for distributing the workload across multiple data center. The incoming tasks are prioritized with job request parameters and the classification is carried out using gradient boosting classifier in structure of an ensemble of base decision trees classifier. The gradient boost ensemble classifier categorizes number of tasks as immediate or reserved. This helps to improve classification accuracy. After classification, the tasks are distributed to the multiple datacenters with optimal resource utilization and traffic occurrence level. Experimental evaluation of ROTAGBC technique is carried out with certain parameters such as classification accuracy, space complexity, task completion time and resource utilization rate. The results analysis of ROTAGBC technique improves classification accuracy and resource utilization rate with minimum space and task completion time than the state-of-art methods.

References

[1] Li, P., Guo, S., Miyazaki, T., Liao, X., Jin, H., Zomaya, A.Y. and Wang, K. Traffic-aware geo-distributed big data analytics with predictable job completion time. *IEEE Transactions on Parallel and Distributed Systems* **28** (6) (2017) 1785-1796.

[2] Xiao, W., Bao, W., Zhu, X. and Liu, L. Cost-Aware Big Data Processing Across Geo-Distributed Datacenters. *IEEE Transactions on Parallel and Distributed Systems* **28** (11) (2017) 3114-3127.

[3] Lu, Q., Li, Z., Kihl, M., Zhu, L. and Zhang, W. CF4BDA: A Conceptual Framework for Big Data Analytics Applications in the Cloud. *IEEE Access* 3 (2015) 1944-1952.

[4] Ji, S. and Li, B. Wide area analytics for geographically distributed datacenters. *Tsinghua Science and Technology* **21** (2) (2016) 125-135.

[5] Zeng, D., Gu, L. and Guo, S. Cost minimization for big data processing in geo-distributed data centers. *Cloud networking for big data*, 2015, 59-78.

[6] Zhou, D., Guo, J., Zhang, Y., Chai, J., Liu, H., Liu, Y. and Liu, Y. Distributed data analytics platform for wide-area synchrophasor measurement systems. *IEEE Transactions on Smart Grid* **7** (5) (2016) 2397-2405.

[7] Zou, H., Yu, Y., Tang, W. and Chen, H.W.M. FlexAnalytics: a flexible data analytics framework for big data applications with I/O performance improvement. *Big Data Research* **1** (2014) 4-13.

[8] Wang, L., Tao, J., Ranjan, R., Marten, H., Streit, A., Chen, J., & Chen, D. G-Hadoop: MapReduce across distributed data centers for data-intensive computing. *Future Generation Computer Systems* **29** (3) (2013) 739-750.

[9] Triguero, I., Peralta, D., Bacardit, J., García, S. and Herrera, F. MRPR: a MapReduce solution for prototype reduction in big data classification. *Neurocomputing* **150** (2015) 331-345.

[10] Del Rio, S., Lopez, V., Benitez, J.M. and Herrera, F. A mapreduce approach to address big data classification problems based on the fusion of linguistic fuzzy rules. *International Journal of Computational Intelligence Systems* **8** (3) (2015) 422-437.

[11] López, V., Del Río, S., Benítez, J.M. and Herrera, F. Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data. *Fuzzy Sets and Systems* **258** (2015) 5-38.

[12] Wang, K., Li, H., Feng, Y. and Tian, G. Big data analytics for system stability evaluation strategy in the energy internet. *IEEE Transactions on Industrial Informatics* **13** (4) (2017) 1969-1978.

[13] Diamantoulakis, P.D., Kapinas, V.M. and Karagiannidis, G.K. Big data analytics for dynamic energy management in smart grids. *Big Data Research* **2** (3) (2015) 94-101.

Received: 5 Feb 2018/Accepted: 15 Mar 2018

[14] Del Río, S., López, V., Benítez, J.M. and Herrera, F. On the use of MapReduce for imbalanced big data using Random Forest. *Information Sciences* **285** (2014) 112-137.

[15] Raut, S., Jaiswal, K., Kale, V., Mote, A. and Soudamini, M. Data Distribution Handling on Cloud for Deployment of Big Data. *International Journal on Cloud Computing: Services and Architecture (IJCCSA)* **6** (3) (2016) 15-22.

[16] Zhang, L., Wu, C., Li, Z., Guo, C., Chen, M. and Lau, F.C. Moving big data to the cloud: An online cost-minimizing approach. *IEEE Journal on Selected Areas in Communications* **31** (12) (2013) 2710-2721.

[17] Rahman, M.N. and Esmailpour, A. A hybrid data center architecture for big data. *Big Data Research* **3** (2016) 29-40.

[18] Kliazovich, D., Bouvry, P. and Khan, S.U. GreenCloud: a packet-level simulator of energy-aware cloud computing data centers. *The Journal of Supercomputing* **62** (3) (2012) 1263-1283.

[19] Wang, S., Qian, Z., Yuan, J. and You, I. A dvfs based energy-efficient tasks scheduling in a data center. *IEEE Access* **5** (2017) 13090-13102.

[20] Ayma, V.A., Ferreira, R.S., Happ, P., Oliveira, D., Feitosa, R., Costa, G. and Gamba, P. Classification algorithms for big data analysis, a Map Reduce approach. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences* **40** (3) (2015) 17-21.

Authors Profile

C.R. Durga Devi received her Bsc.Computer Technology from Coimbatore Institute of Technology, Coimbatore, India. she had her Master of Computer Applications from Bharathiar University, Coimbatore, India. she holds Mphil in Computer Science from Bharathiar University, Coimbatore, India. She has 11 years of experience in teaching. She is presently working as an Assistant Professor in NGM college, pollachi. Her research interest includes Data Mining, Big Data Analytics. Now she is pursuing her ph.d Computer Science in Dr.Mahalingam center for research and Development at NGM college, Pollachi.

Dr.R. Manickachezian received his M.Sc., Degree in Applied Science from P.S.G College of Technology, Coimbatore, India in 1987. He completed his M.S. Degree in Software Systems from Birla Institute of Technology and Science, Pilani, Rajasthan, India and Ph.d degree in Computer Science from school of Computer Science and Engineering, Bharathiar University, Coimbatore, India. He served as a faculty of maths and Computer Applications at P.S.G College Of Technology, Coimbatore from 1987 to 1989. presently, he has been working as an Associate Professor of Computer Science in NGM college (autonomous), pollachi under Bharathiar University, Coimbatore, India since 1989. He has published 150 papers in International/National Journal and Conferences. He is a recipient of many awards like best Computer Science Faculty of the year 2015, Best Research Supervisor award, Life Time Achievement award, Desha Mithra Award and best paper award. His research focuses on Network Databases, Data Mining, Distributed Computing, Data Compression, Mobile Computing, Real Time Systems and Bio-Informatics.

Received: 5 Feb 2018/Accepted: 15 Mar 2018