International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

EARMGA and Apriori Algorithm's Performance Evaluation for Association Rule Mining

Sandeep Pratap Singh, Dharani Kumar Talapula

Abstract: ***Association rule mining techniques are important part of data mining to derive relationship between attributes of large databases. Association related rule mining have evolved huge interest among researchers as many challenging problems can be solved using them. Numerous algorithms have been discovered for deriving association rules effectively. It has been evaluated that not all algorithms can give similar results in all scenarios, so decoding these merits becomes important. In this paper two association rule mining algorithms were analyzed, one is popular Apriori algorithm and the other is EARMGA (Evolutionary Association Rules Mining with Genetic Algorithm). Comparison of these two algorithms were experimentally performed based on different datasets and different parameters like Number of rules generated, Average support, Average Confidence, Covered records were detailed.***

Index Terms**: Data mining, Apriori algorithm, EARMGA, Association rule mining**

I. INTRODUCTION

Association rule mining is one of the most interesting field of research in data mining. It was first introduced by Agrawal et.al. [1]. Its objective is to generate useful relations, pattern discovery, associations among sets of items in the transaction databases ie., generally large datasets with few prior information for an occurrence feature. There can be numerous areas of applications for the purpose of knowledge discovery using association rules, which have effectively solved the real world problems like biomedical, telecommunication networks, stock control, and market basket analysis risk management etc.

Association rule is the implication of format $M \longrightarrow N$, means M implies N where $M, N \subset I$ sets of items termed as itemset and $M \cap N = \phi$. I is the set of distinct item. Since the consideration is large datasets and probably even be high-dimensional are usually huge, so user requires association rules which are useful and interesting to collaborate association with in the data and their correlation. For this purpose, there are concepts of *support* and *confidence.* To determine the usefulness, interestingness rules such as threshold of support (minimum support) and threshold of confidence (minimum confidence) were taken into account. called frequent itemsets.

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

Sandeep Pratap Singh*, Department of Virtualization, School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India. Email: sandeep102209@gmail.com

Dharani Kumar Talapula, Department of Virtualization, School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India. Email: tdharani@rediff.com

These obtained frequent or large itemsets produce useful association rules with minimum confidence threshold. Let one of the large itemset is B_r, thus B_r = {IS_1 ,IS_2 … IS_r }association rules are generated with Br itemset as follow: the first one is { IS_1 ,IS_2 … IS_r }→{IS_k}, now by comparing with the threshold confidence, check if this rule is interesting or not. Similarly, other rules are obtained by inserting the last item of predecessor to the consequent. Subsequently the confidence's of the newly obtained rules are verified to determine the interestingness and usefulness among them[2].

Association rule mining problem is well-explored research field. There are numerous algorithms and techniques available to solve various rule mining problem [4], [5], [6], [7], [8], [9], [10], [19], [20]. Each of them have their own merits and demerits. Among these algorithms, Apriori and EARMGA algorithms are compared on different parameters with different datasets. KEEL Software Tool has been used for this purpose[16][18]

The article is organized as follows. Section II explains Apriori algorithm. Section III describes EARMGA algorithm. Section IV gives details of experiments performed. Section V presents the results of experiment and Section VI concludes the paper.

II. APRIORI ALGORITHM

Agrawal and Srikant [3] were the first to propose Apriori Algorithm, which was great advancement in the area of association rule mining. It was developed as an improvement over first algorithm AIS Agrawal et al.[1] for generating association rules with *confidence* and all *frequent* itemsets. AIS was simple method requiring lots of passes over the database, resulting in too many candidate itemsets among them, most of the itemsets are infrequent. This results in wastage of space and time of computation by computing irrelevant itemsets. Apriori is much improved algorithm which uses different steps for candidate generation and pruning of itemsets. Apriori property mentions that all nonempty subsets of frequent itemsets be deduced by candidate generation and they should be frequent.

Initially in the first pass large 1-itemset(*r*-itemset refers to *r* no. of items in an itemset) is generated by just counting the occurrences of item. Apriori algorithm uses two phases to calculate large frequent itemsets from the database. First phase is to generate candidate itemsets using *apriori-generation* function, second phase is to test the support count of the generated itemset by scanning the

database. In the initial scan of database, support count of each item is computed by pruning of itemsets below the minimum support count will result in large 1-itemset. For quick computation of support count, efficient generation of candidates C_r is computed with the help of *subset* function. After (*r*-1)th pass over the database and by joining the *r*-1 frequent itemset's candidate *r*-itemsets. Candidate itemsets are pruned according to apriori property by which all *r*-itemset candidate check its sub(*r*-1) itemsets. If any of its sub itemsets are not frequent then candidate *r*-itemsets will be pruned out. All frequent itemsets are deduced by executing processes iteratively until any of candidates itemsets or frequent itemsets becomes empty. For details of *apriori-generation* and *subset, GenerateRules* functions refer [3]. Algorithm of Apriori is as follows:

Input: In database D
Minimum Support Threshold = α
Minimum Confidence Threshold = β
Output:
A_t Association rules from D
Steps:
Initially r =1,B_r= large *r*-itemsets;
r=2
while($B_{r-1} \neq \emptyset$) do begin
C_r =*apriori-generation*(B_{r-1}); // It will generate new candidate itemsets from B_{r-1}.
for all transactions T $\in$ D
do begin
C_t=*subset*(C_r,T); //candidate itemsets contained in T.
for all candidates C $\in C_t$ do
S_Count(C)=S_Count(C)+1; // increments support count of C by 1
end
B_r={C $\in C_t$ |S_Count(C) $\geq \alpha$ }
End
$B_f = \bigcup_r B_r$;
A_t= *GeneratRules*(B_f ,β)
r++;
end

III. EARMGA ALGORITHM

EARMGA is developed by YanZhangZhang et.al for the advancement over ARMGA [10]. ARMGA was designed only for Boolean association rule mining. EARMGA is an expanded form of ARMGA which is developed for generalized association rules [11][12].Genetic algorithm is mainly used for the efficient searching, particularly when deterministic searching techniques were less effective. Due to searching requirement of large size of database, Genetic algorithm follows the course of natural evaluation of species with genetic evolution techniques like inheritance, selection, crossover and mutation.

EARMGA generates the most interesting rules with the help of fitness function. Fitness function measures the interestingness which is given by relative confidence in EARMGA. In this method user do not need to mention the minimum support threshold. In traditional association rule mining, rules are generated based on comparison of support and confidence with minimum support and minimum confidence respectively. Unlike traditional rule mining in EARMGA confidence ($M \rightarrow N$) (association $M \rightarrow N$ is a *r*-rule if $M \cup N$ is a *r*-itemset) should be greater than or equal to support(*N*). Therefore, *interestingness measure* as relative confidence is given as:

$$rel_{conf} = \frac{support(M \cup N) - support(M)support(N)}{\big(1 - support(N)\big)support(M)}$$

Association rule mining task by EARMGA can be stated as : Input database D and rule of length r. Now search for the interesting generalized association r-rules with their relative confidence optimized by genetic algorithm. Before designing the algorithm, single chromosome for each rule is generated by encoding each association rule [10].

A. Algorithm Design

We have *popu,sp'* as chromosome population and selection probability respectively. Unwanted chromosomes are discarded using *select*(*popu,sp'*)function on the basis of fitness values. It takes *popu* as input and gives a new selected population as output. Function *crossover* (*popu*,*cp'*) also produces new chromosome according to crossover probability *cp*'. With the given population, *popu* function *mutate*(*popu,mp'*) changes the chromosome population *popu* on the basis of mutation probability *mp'* and fitness values of chromosomes. Now population is initialized with help of fuction *initialize*(*s'*) where *s'* is seed chromosome. All these functions integrate to form EARMGA algorithm. Fitness function used is a relative confidence of association rules to increase efficiency, Generalized FP-tree given by Han et al. can be use to implement EARMGA algorithm[7][10]. EARMGA algorithm is as follows:

population EARMGA (*s',sp',cp',mp'*)
begin
for(*i*←0,*popu*[*k*]← *initialize*(*s'*); *terminate* (*popu*[*k*]); *k*++)
do begin
popu_t ← *Ø*;
popu [*k+1*] ← *select*(*popu* [*k*] ,*sp'*);
popu_t ← *crossover*(*popu* [*k+1*],*cp'*);
popu[*k+1*] ← *popu* [*k+1*] ∪ *mutate*(*popu_t,mp'*);
end
return *popu* [*k*];
end

IV. EXPERIMENT PERFORMED

In the experiment section, Apriori algorithm is compared with EARMGA algorithm on 5 datasets and different parameters. For performing experiment we have used KEEL software tool [16][18].

A. Dataset Used

In our experiment we have 5 datasets namely Wine, Forest Fire, Breast Cancer Wisconsin (original), Mushroom, Computer Hardware.

All datasets as mentioned: Wine dataset, Forest Fire, Breast Cancer, Mushroom, Computer Hardware datasets were taken from UCI repository.

Before using the data it is first converted into KEEL format to be compatible with the KEEL Software Tool[13], [14]. Datasets size and instances are shown in Table 1.

Table 1. Datasets size and attributes

Dataset	No.of attributes	No.of instances
Wine	13	178
Forest Fire	13	517
Breast Cancer Wisconsin(original)	10	699
Mushroom	22	8124
Computer Hardware	9	209

B. Comparison Parameters

In this section we will state the parameters on which both Apriori and EARMGA are compared on above mentioned datasets. The parameters are as follows:

Execution Time. It is the time consumed in generating the association rules. Algorithm having lesser execution time is better.

Number of Association rule generated:

It is the number of association rules after the execution of the experiment on a particular algorithm. Algorithm generating lesser number of rules is better, because information is obtained with lesser number of rules .This results in reducing time and space complexity.

Average Support:

It refers to the frequency of itemsets in transactions. More support will have the larger occurances of itemsets in the transactions. For eg. 0.2 % support of an item means only 0.2% of transaction has this item. So algorithm with greater average support is preferred.

Average Confidence:

It refers to the strength of association rules. For eg. if confidence of association rule $M \rightarrow N$ is 90% means 90% of transactions that contain M also contain N with it together.

Covered Records:

It means the percentage of record in dataset accessed by the algorithm at a particular condition.

C. Experimental Setup

In this section experimental setup details are elaborated. Parameters of both Apriori and EARMGA algorithm that are set in KEEL software tool are also detailed. Parameters set in KEEL for Apriori algorithm are shown in Table 2 and parameters set in KEEL for EARMGA algorithm are shown in Table 3

Table 2. Parameters set in Apriori algorithm

Parameter Descriptor	Value
No. of partitions	4
for numeric data	
Minimum Support	0.3
Minimum Confidence	0.8

Table 3. Parameters set in EARMGA algorithm

Parameter Descriptor	Value
Fixed Length of Association Rules	2
Population Size	30
Number of Generations	100
Probability of Selection	0.75
Probability of Crossover	0.7
Probability of Mutation	0.1
Number of Partitions for Numeric Attributes	4
Minimum Support	0.3
Minimum Confidence	0.8

In experiment setup we first select the dataset on which experiment is to be performed. Some of the datasets selected have some missing values, to solve this some preprocessing methods are applied using Fuzzy K-means implementation[15]. A clean dataset is produced by removing rows containing missing values. It will preprocess data by means of k fuzzy cluster of data and inserting the values on the basis of closeness degree of each instance to each cluster. After this association rule mining algorithms is applied with the parameter values shown in Table 2 and Table 3.

V. Results Obtained

After the complete setup of values of parameters in each algorithm, experiments are carried out for each datasets on both algorithms. Following are the results obtained as shown in Table 4 and Table 5. Some acronyms are used for datasets like Wi for wine, B.C. for breast cancer, F.F for forest wire, Mush for Mushroom, ComH for computer hardware dataset.

Table 4 Results on Apriori algorithm for *0.3* min support

S.No	Parameter/ Dataset	Wi	F.F	B.C	Mush.	ComH
1	Execution Time	0.078	0.203	3.556	34.179	0.14
2	No.of Assoc.Rules Generated	2	477	10207	11678	441
3	Average Support	0.3483	0.416	0.480	0.3572	0.7475
4	Average Confidance	0.8612	0.973	0.981	0.9538	0.9428
5	Covered Records	34.831	100	100	100	99.521

Table 5 Results on EARMGA algorithm for *0.3* min support

S.No.	Parameter/ Dataset	Wi	F.F	B.C.	Mush.	ComH
1	Execution Time	1.732	4.01	4.426	54.687	5.655
2	No.of Assoc.Rules Generated	10	30	13	26	13
3	Average Support	0.623	0.8680	0.8261	0.9447	0.8803
4	Average Confidance	1	1	1	1	1
5	Covered Records	100	100	100	100	100

We have done the comparative study of two algorithms by plotting the results obtained in excel charts shown in Fig.1, Fig.2, Fig.3, Fig.4, and Fig.5.

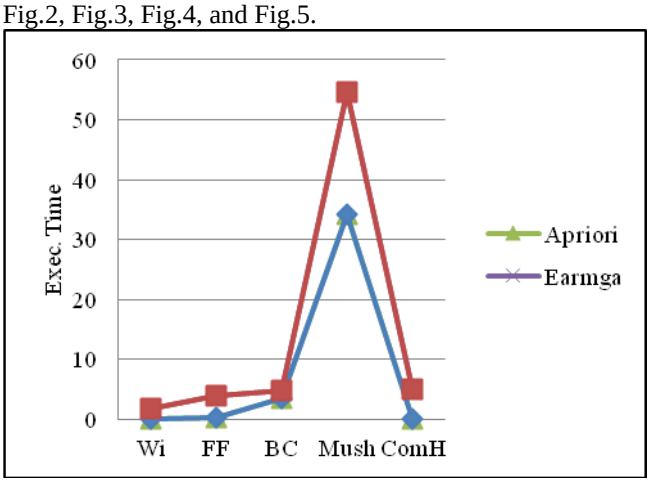

Fig. 1. Graph between execution time and different datasets comparing Apriori and EARMGA.

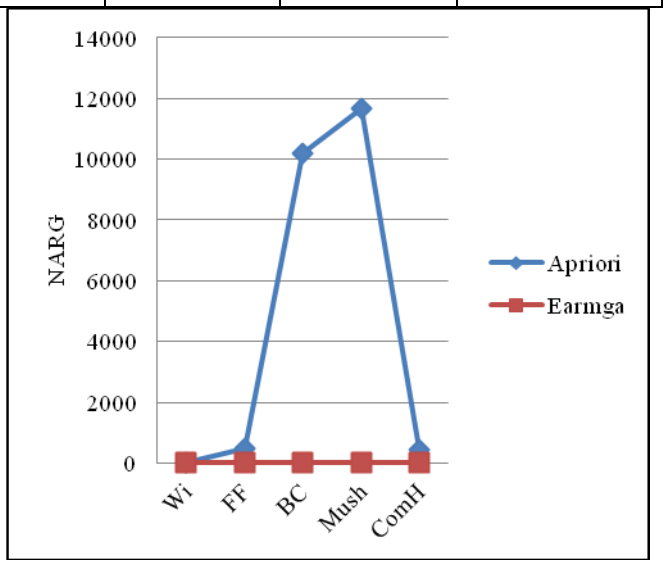

Fig. 2. Graph between No. of Association rules generated (NARG) and different datasets comparing Apriori and EARMGA

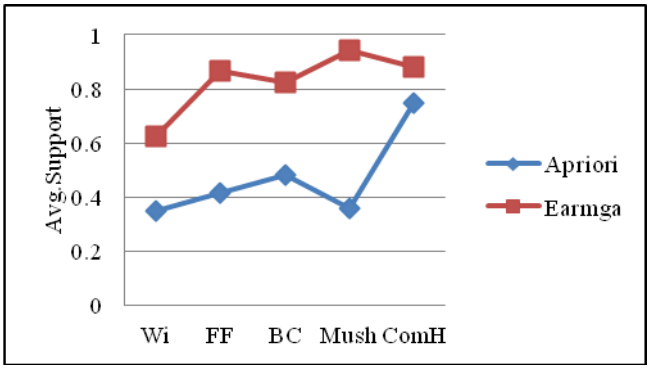

Fig. 3. Graph between Average Support and different datasets comparing Apriori and EARMGA

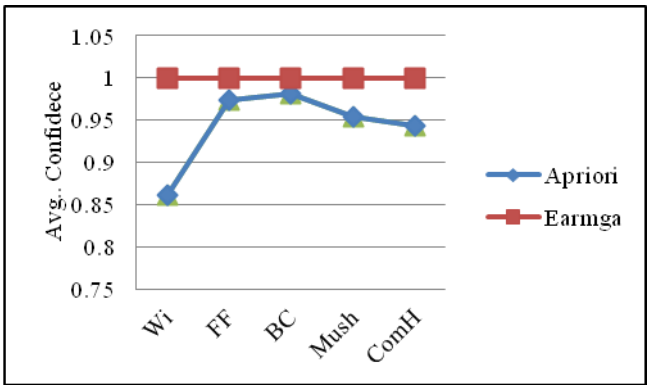

Fig. 4. Graph between Average Confidence and different datasets comparing Apriori and EARMGA

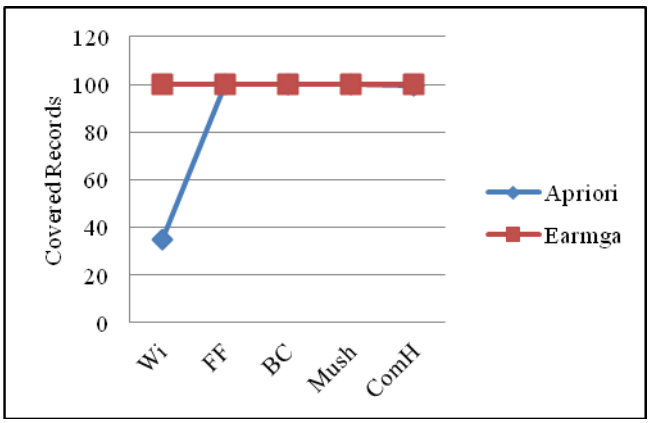

Fig. 5. Graph between Covered Records and different datasets comparing Apriori and EARMGA

Here we have plotted the graphs for 0.3 minimum support because it will give better results on considered datasets. If min support is taken as greater value, then Wine dataset may not generate sufficient rules and if min support lesser value is taken then, Mushroom dataset will generate excessive rules. EARMGA generally does not require minimum support, here it is considered as additional measure. Fig.1. shows that Apriori algorithm performs better than EARMGA algorithm on parameter of Execution Time with every dataset taken. Fig.2. shows that number of Association rules generated in Apriori algorithm is very large especially in Mushroom and Breast cancer dataset except Wine dataset. The value of this parameter is substantially decreased in EARMGA algorithm. So EARMGA algorithm performs much better than Apriori algorithm on parameter of number of Association rule generated. Fig.3. shows that Average Support in case of EARMGA is always higher than Apriori algorithm. So EARMGA gives better performance on the parameter of Average Support. Fig.4. shows 100% confidence obtained on every dataset by EARMGA algorithm. Due to this reason, EARMGA is very effective algorithm in terms of confidence. Fig.5. shows Covered Records in dataset is 100% in case of EARMGA, therefore it is again better than Apriori. Overall EARMGA scores over the Apriori algorithm on majority of the parameters.

VI. CONCLUSION

Association rule plays a prominent role in many data mining applications like market basket analysis, inventory management etc. It helps in generating interesting and useful patterns in large database. For this many association rule mining techniques are used to generate frequent itemset from which required interesting rules are deduced. Apriori algorithm is most common for this purpose, which takes less execution time but may generate too many rules which may not be appropriate. On the other hand EARMGA algorithm will take comparatively more time but will generate lesser no. of association rules and with high interestingness based on genetic algorithm. Therefore, to conclude Genetic algorithm gives optimum search results.

ACKNOWLEDGMENT

The authors would like to thank and acknowledge anonymous reviewers for their valuable consecutive and constructive comments to improve the quality of this paper.

REFERENCES

1. Agrawal, R., Imielinski, T., and Swami, A. N. 1993. Mining association rules between set of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Washington D.C.,207-216..
2. D. Ai, H. Pan, X. Li, Y. Gao, and D. He, “Association rule mining algorithms on high-dimensional datasets,” Artif. Life Robot., vol. 23, no. 3, pp. 420–427, 2018.
3. Agrawal, R. and Srikant, R. 1994. Fast algorithms for mining association rules. In Proc.20th Int. Conf. Very Large Data Bases, 487-499..
4. J. Kaur, R. Singh, and R. K. Gurm, “Performance evaluation of Apriori algorithm using association rule mining technique,” vol. 2, no. 5, 2016.
5. Park, J. S., Chen, M.-S., and Yu, P. S. 1995. An effective hash based algorithm for mining association rules. In Proceedings of the 1995 ACM SIGMOD..
6. Srikant, R., & Agrawal, R. (1995). Mining generalized association rules In Proceedings of the 21st international conference on very large databases (VLDB’95) (pp. 407–419).
7. Q.Zhao,“Association Rule Mining : A Survey Association Rule Mining : A Survey,” vol. 5, no. March, pp. 2320–2324, 2016.
8. S. O. Fageeri, R. Ahmad, and H. Alhussian, “A performance analysis of association rule mining algorithms,” 2016 3rd Int. Conf. Comput. Inf. Sci. ICCOINS 2016 - Proc., pp. 328–333, 2016.
9. D. Ai, H. Pan, X. Li, Y. Gao, and D. He, “Association rule mining algorithms on high-dimensional datasets,” Artif. Life Robot., vol. 23, no. 3, pp. 420–427, 2018.

10. Xiaowei Yan,Chengqi Zhang,Shicho Zhang.Genetic algorithm based strategy for identifying association rules without specifying actual minimum support.In Expert Systems with Application 36(2009)3066-3076.
11. Q. Zhao, "Association Rule Mining : A Survey Association Rule Mining : A Survey," vol. 5, no. March, pp. 2320–2324, 2016.
12. Srikant, R., & Agrawal, R. (1996). Mining quantitative association rules in large relational tables. In Proceedings of the ACM SIGMOD international conference on management of data (SIGMOD'96) (pp. 1–12).
13. J. Alcalá-Fdez, L. Sánchez, S. García, M.J. del Jesus, S. Ventura, J.M. Garrell, J. Otero, C.Romero, J. Bacardit, V.M. Rivas, J.C. Fernández, F. Herrera KEEL: A Software Tool to Assess Evolutionary Algorithms to Data Mining Problems.Soft Computing 13:3(2009) 307-318
14. J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework. Journal of Multiple-Valued Logic and Soft Computing 17:2-3 (2011) 255-287.
15. J. Deogun, W. Spaulding, B. Shuart and D. Li. Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method. 4th International Conference of Rough Sets and Current Trends in Computing (RSCTC'04). Lecture Notes in Computer Science 3066, Springer 2004, Uppsala (Sweden, 2004) 573-579.
16. F. Khan and D. Singh, "Knowledge Discovery on Agricultural Dataset Using Association Rule Mining," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 4, no. 5, pp. 925–930, 2014.
17. D. Barua, P. Jain, J. Gupta, and D. V Gadre, "Road Accident Prevention Unit (R.A.P.U) (A Prototyping Approach to Mitigate an Omnipresent Threat)," 2013 Texas Instruments India Educators' Conference, pp. 56–60, 2013
18. J. Alcalá-Fdez *et al.*, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *J. Mult. Log. Soft Comput.*, vol. 17, no. 2–3, pp. 255–287, 2011.
19. J. Kaur, R. Singh, and R. K. Gurm, "Performance evaluation of Apriori algorithm using association rule mining technique," vol. 2, no. 5, 2016.
20. T. Tasneem, T. Tasneem, and M. M. Jahangir Kabir, "Performance Analysis of Classical and Evolutionary Algorithms for Mining Association Rules," 2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE 2019, pp. 1–6, 2019

AUTHORS PROFILE

Sandeep Pratap Singh received his M.Tech degree in 2012 from Jaypee University of Information Technology, Solan (H.P.) and BE degree in 2009 from RGPV Bhopal . Presently he is working as Assistant Professor (SS) in Department of Virtualization, School of Computer Science, University of Petroleum and Energy Studies, Dehradun, India.. His research interest lies in data mining, fraud detection, biomedical image processing He has published 7 research papers in Referred Journals and International Conferences. He has also served as TPC member of International Conferences.

Dharani Kumar Talapula received his M.S degree in 2004 from Middlesex University, North London, United Kingdom and B.Tech degree in 1997 from Sri Krishnadevaraya University, A.P, India. Presently he is working as Industry Fellow in Department of Virtualization, School of Computer Science, University of Petroleum and Energy Studies, Dehradun. India. His research interests lies in Cloud, Machine data/logs, Bigdata, Machine Learning.