

# Prognosis of Disease that may Occur with Growing Age using Confabulation Based Algorithm

Neha Srivastava\* and Jyoti Gautam

Computer Science Department, JSSATE, Noida, India

\*E-mail: [nehasrivastava@jssaten.ac.in](mailto:nehasrivastava@jssaten.ac.in)

## ABSTRACT

The enduring diagnosis of patient's medical records might be useful to determine the causes that are responsible for a particular disease. So that, one can take early preventive measures to curtail the risk of diseases that may occur with the growing age. Consequently, this can enhance the life expectancy probability. Here, a new algorithm CMARM is proposed for analysis of symptoms in order to find out the disease that may occur frequently and rarely with growing age. It uses map reduce paradigm inspired by cognitive learning. It is concerned with acquisition of problem-solving skills, intelligence and conscious thought and uses prevailing knowledge to generate new rules. It has been evaluated over synthetic data sets collected from the health data repository. Since, CMARM requires one-time file access therefore, it is consistently faster and also consuming less memory space than the FP tree based algorithm.

**Keywords:** Associations rule mining; Cogency; Confabulation; Map reduce; Data mining; Disease; CMARM

## NOMENCLATURES

HDFS	Hadoop distributed file system,
$S_0$	Minimum support value,
LS	Least minimum support

## 1. INTRODUCTION

Presently, diseases are considered as one of the major factors for illness and death in human life span. According to WHO, ischemic heart disease, stroke, lower respiratory infections and chronic obstructive lung disease are the major cause of death from the past decade<sup>12</sup>. The early analysis of any such disease that frequently and rarely occur with the increasing age can be helpful in curing the disease entirely or to some point at early stage and increase the life expectancy probability. There are number of factors that are responsible for affecting the health of individuals and communities; and eventually give birth to the numerous kinds of diseases. Since, the medical domain is becoming an increasingly data intensive field as doctors and researchers generate gigabytes of medical data related to patients and their illnesses. Also, the rapid advancement in automation of the healthcare industry produces a vast amount of complex and heterogeneous, both structured and unstructured data which it is difficult to analyze in order to make any important decision regarding patient health<sup>14,16</sup>.

There are number of algorithms implemented in order

to classify, cluster and find hidden patterns in data sets. Basic approaches of association rule mining<sup>1,4,7</sup> may be good for frequent itemset mining but is not necessarily appropriate when rare item set are most desirable<sup>2</sup>, as the number of uninteresting rules increases with decrease in minimum support value. To overcome this issue, a new algorithm CARM is proposed based on cognitive learning<sup>18</sup>, where author introduces a new parameter called an interestingness measure to find out rare itemset among the given datasets. The negative aspect of CARM algorithm is that the parameter for selecting active items and interesting rules is predefined by user<sup>18</sup>. It would be desirable to set these parameters automatically using data statistics. Moreover, statistical analysis of CARM is difficult due to use of matrices in the implementation of CARM algorithm<sup>18</sup>. Another algorithm named FSM-H based on concept of map reduce paradigm<sup>18</sup> is very efficient in terms of time, accuracy and for handling big data. However, author shows the performance of FSM-H is complex to a certain extent compare to other algorithms and also complexity increases with increase in dataset size. Majali<sup>19</sup>, *et al.* proposed an algorithm which is an automatic procedure, called self-structured confabulation network for fast anomaly detection and reasoning. This model comprises a distinct node that contain both spatial and temporal relations among the features of the dataset.

Different mining algorithms are proposed to find out frequent and rare itemset and also association rules among them. But there are certain drawbacks associated with them like some required multiple database scanning, time consuming,

less efficient etc<sup>1,2,11,21</sup>. Also, in order to apply mining methods for clinical data, the researchers shall additionally resolve the problems related to patient privacy, semantic interoperability, heterogeneous data sources and unstructured data presented in text or media formats. To overcome all these issues, a new algorithm CMARM-confabulation based map reduce association rule mining algorithm is proposed. The algorithm is based on how human brain recognises the existing itemset and finds association rules among them<sup>15,18,19</sup>.

CMARM is implemented using concept of map reduce paradigm of distributed computing because it is based on the data centric approach of distributed computing for analyzing huge data sets<sup>18</sup>. The key idea of the proposed work is to design an involuntary procedure that learns structure of a confabulation network from input datasets and mine association rules for frequent and rare itemset by strengthening knowledge link between itemset.

## 2. METHODS

The proposed algorithm is designed as an iterative map reduce procedure based on confabulation method<sup>20,22,23</sup>. It is processed in two steps:

1. Finding frequent and rare itemset using map reduce procedure<sup>5,8,17,24</sup>.
2. Mining association rules for generated frequent and rare itemset using confabulation method.

To implement first step, map reduce framework<sup>17</sup> is used to signify knowledge links with weak strength using threshold value of support count. This leads to the generation of frequent and rare itemset, whose support count is greater than the minimum specified support count value. In second step, strong association rules are generated based on confabulation theory, which states<sup>20</sup>;

$$Fr = \{x \in I \mid \text{support\_count}(x) \geq S_0\} \quad (1)$$

Here  $S_0$  is the minimum support value. It considers  $S_i = F_r$  where  $S_i$  is set of initial itemset. After finding n-itemset, algorithm generates all rules using their support, confidence and interestingness parameter for selecting active items and interesting rules that strengthen the knowledge link between the itemsets<sup>13,20</sup>.

### 2.1 Design Model

The CMARM approach is divided into three modules<sup>20,22,23</sup>,

- (i) Creating distributed cluster: Since, the medical data deals with voluminous and heterogeneous data structure. Therefore, to handle such big amount of data HDFS is used<sup>3,10,14</sup>. The data files are also stored across multiple machines in redundant fashion. Thus, makes the proposed system is also fault tolerant. It also makes applications available to parallel processing.
- (ii) Finding frequent and rare itemset: Here, minimum support value is used to delete irrelevant items by ignoring the items whose count is less than specified minimum support value. There are two functions defined; mapper and reducer to work on this module. Mapper divides input dataset into small datasets such that these small datasets

are passed to each iterations of mapper function and corresponding result of reducer function is observed. The mapper via map method processes single line at a time. It then splits the line into key-value pair of < <disease>, count> for each age-group. The Reducer, via reduce method sums up the values, which are the occurrence count for each item set.

- (iii) Finding association rules for frequent and rare itemset: This module deals with the finding of association rules based on confabulation theory. This means in every loop, mapper generates some rules such that only those rules are considered whose probability to win is maximum. Association rules are used to determine the category in which given disease is fall. This category would be used in the generalisation of patient record intensely and categories the patient into specific category indicating the disease with which he is suffering mainly. For example, if a patient suffers from 'abdominal hernia', then this disease is falls in category 'diseases and disorders of the digestive system' indicating he suffers from diseases and disorders of digestive system. Each category is associated with some confidence value which show that if rule  $X \rightarrow Y$  holds with confidence  $c$  then  $c\%$  of transaction in data set  $D$  that contain  $X$  also contain  $Y$ . Confidence value is calculated as;

$$\text{Confidence}(X \rightarrow Y) = \text{SUPP}(X \cup Y) / \text{SUPP}(X) \quad (2)$$

Where,  $\text{SUPP}(X \cup Y)$  and  $\text{SUPP}(X)$  is support value for  $(X \cup Y)$  and  $X$ . Support value determines how often a rule is applicable to a given data set. For rule  $X \rightarrow Y$ , support value of  $X$  with respect to transactions  $T$  is defined as percentage of transactions which contains item-set  $X$ . It is calculated as;

$$\text{Support}(X \rightarrow Y) = \text{probability}(X \cup Y) / N \quad (3)$$

Where,  $N$  is the total number of transactions. Basically, association rules are find out in order to conclude if a person is suffering from any particular disease then, there is also a chance that he may also suffer with other disease arises from that particular disease.

### 2.2 Algorithm

2.2.1 Pseudocode of Iterative map reduce algorithm:

1. While (itemset! = null)
2. Execute Mapreduce Job
3. Write result to DFS
4. Update condition

Here, line 2 and 3, performs the filtering and partitioning of input dataset horizontally into  $P$  node subsets and distribute it to  $Q$  nodes supersets and writes each partition result in the HDFS<sup>20</sup>.

2.2.2 Pseudocode of CMARM algorithm:

1.  $S1 = Fr$  // set of all frequent and rare itemsets-1
2. Find  $S2$ , the set of all frequent 2 itemsets
3. Find all rules from 2-itemsets (according to their support and confidence)

4. While  $Sk \neq \emptyset$ 
  - 4.1.  $Sk = \text{Frequent-Rare-itemset\_Generation}(Sk, \text{count})$
  - 4.2. for all  $c \in D_{k+1}$  //for all item  $c$  belongs to given datasets
  - 4.3. while  $c\_count > \text{min\_Support}$
  - 4.4. //Support\\_counting( $c, D$ )
  - 4.5. if  $\text{cogency } X \rightarrow Y > (\text{interesting\_measure and min\_confidence})$
  - 4.6. add  $Sk+1 = Sk+1 + \{c\}$  //find association rules for frequent and rare itemsets
  - 4.7. end for
5.  $K = K+1$   
end while

Here, line 4.1 signify that mapper generates static data structure and emits key-value pair of all single length patterns to reducer. Line 4.3, discards the itemset whose support count is less than minimum support, which is a user defined variable. After finding frequent and rare itemset, next step is to extract strong association rules and it is done in line 4.4 and 4.5. Here, mapper computes variations of user-feature and item-feature pair. These variables are used as an interestingness measure for selecting active items and interesting rules for frequent and rare itemset. Then, reducer computes the sum of changes and applies this to each item rating by updating appropriate feature and directs this to mapper function again for next iteration.

### 2.3 Dataset

In order to validate CMARM, an empirical evaluation is performed on dataset retrieved from a health data repository provided by SPARCS<sup>9</sup>. The patient record is grouped into 5 different age groups such as 0-17, 18-29, 30-49, 50-69, 70 or older and for a particular age group frequent and rare disease is evaluated for both female and male category.

### 2.4 Results

Tables 1 and 2 show the summary of frequent and rare disease in each age group for year 2012 and 2013, respectively at  $S_0 = 5$ . Table 4 display summary of association rules with confidence value for both frequent and rare disease for year 2013 at  $S_0 = 5$ . It shows that at the age group of 30-49, Most of the male person suffers from back problem that comes under the category of diseases and disorders of musculoskeletal system and conn tissue and rarely suffers from other nutrition disorder which may also suffers from endocrine, nutritional and metabolic diseases and disorders.

### 2.5 Performance Metrics

Comparison of CMARM with other existing algorithms is done by considering three parameters as accuracy (in terms of Recall and Precision), computational performance (in terms of time taken to process the dataset), and memory consumption in terms of number of nodes maintained<sup>18</sup>.

**Precision and Recall:** Let  $Qp$  be the set of frequent itemsets generated by CMARM and  $Qc$  be the set of correct frequent itemsets. Then,

$$\text{Precision} = |Qp \cap Qc| / |Qp| \tag{4}$$

$$\text{Recall} = |Qp \cap Qc| / |Qc| \tag{5}$$

Similarly, precision and recall is calculated for rare itemset. Table 3 shows precision and recall value of CMARM and CFPgrowth at different minimum support value with noise level 0.05. A high value of precision indicates obtained items from the algorithm have been predicted correctly, but there might be some items have not been identified yet<sup>6</sup>. The good accuracy will be achieved by getting the highest precision and recall simultaneously. Conversely, algorithm should predict the maximum number of features correctly while generating less irrelevant results. Hence, for this F-score is calculated to trade off precision versus recall.

$$\text{F-score} = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall}) \tag{6}$$

Figure 1 and Fig 2 shows that CMARM is more efficient and scalable in terms of execution time and memory usage than CFP growth and FIN algorithm, since:

By increasing number of dataset, memory usage of CFPgrowth increases distinctly due to boost in the size of tree, whereas the memory usage of CMARM remains nearly constant with increase in number of items as shown in Fig. 4.

Execution time has a steady trend as shown in Fig 1, due to the enlarging tree size with growing number of transactions. Execution time includes both time spent for constructing data structure and time spent for mining<sup>18</sup>.

Though CFPgrowth has a flat memory usage trend, it still needs more memory space than CMARM<sup>18</sup>. Consequently, height of the tree and mining time decreases in CFPgrowth as shown in Fig. 2.

When CMARM compare with FIN algorithm, memory usage of FIN is increased markedly due to enhancement in size of tree again. Execution time also increases tremendously with increase in number of transactions as size of tree also increases. So, CMARM is better than FIN algorithm as memory usage and execution time are only depend on the number of transactions. Hence, it requires only one scanning of the database and show a stable execution time and memory usage with increase in number of transactions as shown in Fig 1 and Fig 4 respectively. To shows that CMARM is more reliable than existing algorithm like CFPgrowth and FIN, let's find out the time and space complexity of CMARM. Since, CMARM is based on the concept of map reduce paradigm. Therefore, each node in the cluster of map reduce is likely to report intermittently with completed work and status updates. If a node waits for longer than specified time, the mapper function archives the node as dead and sends out the node's allocated work to other nodes. Similarly, reducer works in the same way.

**Time and space complexity<sup>18</sup>:** A language  $L$  is accepted by CMARM  $[f(n), g(n)]$ , if there is a constant  $0 < c < 1$  and a sequence of mappers and reducers  $m_1, r_1, m_2, r_2, m_3, r_3, \dots$  such that for all  $x \in \{0,1\}^n$  the following is satisfied:

- a. Let  $S = O(F(n))$  and  $M = (m_1, r_1, m_2, r_2, m_3, r_3, \dots, m_n, r_n)$ ,  $M$  accepts  $x$  if and only if  $x \in L$ .
- b. For all  $1 \leq r \leq n$ ,  $m_r, r_n$  use  $O(n^c)$  space and  $O(g(n))$  time.
- c. Each  $m_n$  outputs  $O(n^c)$  keys in round  $n$ .

Therefore, space and time complexity of CMARM is polynomial bounded.

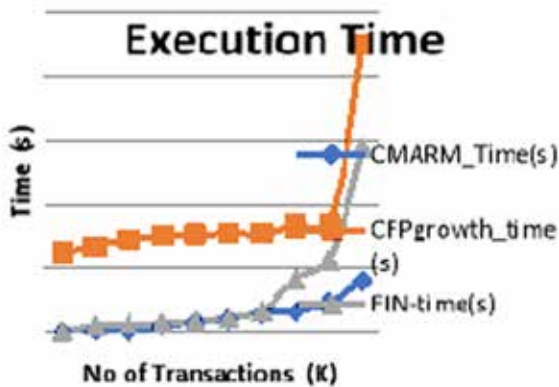


Figure 1. Run Time on Data sets with 1K item.

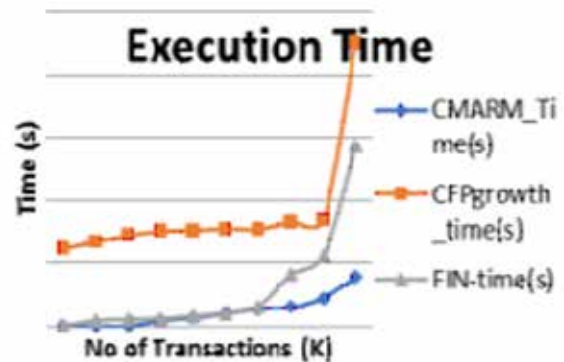


Figure 2. Memory Usage on data sets with 1K items.

4. CONCLUSIONS

CMARM is proposed to diagnose patient medical records in order to find out frequent and rare diseases that may occur in different age groups. Also, this leads to the determination of association rules for the diagnosed disease set. It is processed in two steps i.e. knowledge acquisition and rule extraction. Result show that CMARM is much faster than Apriori and FP tree based algorithm due to one-time file access. It is also efficient and scalable in terms of memory usage, Fig. 2 and execution time, Fig. 1. in comparison to other FP tree based algorithms like

CFPgrowth and FIN algorithm. It is also concluded that CMARM can produce higher performance for mining association rules from rare items, particularly when rare items are important.

5. FUTURE SCOPE

There are a number of interesting directions for future work of CMARM. First, CMARM uses the interesting measure as a parameter which set automatically using data set statistics. In future, this measure can be modified to enhance the accuracy rate in result by understanding the

Table 1. Frequent and rare disease for dataset of year 2012 at  $S_0 = 5$

Age Group	Frequent disease		Rare disease	
	Male	Female	Male	Female
0 -17	Liveborn	Liveborn	GU congenital anomaly	Fluid disorder
18-29	Schiz/Other Psych Disorder	OB perineal trauma	Back problem	Other benign neoplasm
30-49	Back problem	comp birth /puerpm	Bronchial/Lung cancer	Endometriosis
50-69	Osteoarthritis	Osteoarthritis	Fluid/Electrolyte disorder	Enteritis/ulcer colitis
70 or older	Aneurysm	Osteoarthritis	Biliary tract disease	Epilepsy/ convulsions

Table 2. Frequent and rare disease for dataset of year 2013 at  $S_0 = 5$

Age Group	Frequent disease		Rare disease	
	Male	Female	Male	Female
0 -17	Liveborn	Liveborn	Other conn tissue disease	Headache/ migraine
18-29	Crush/ internal injury	Other pregnancy complication	Chest pain	Headache /migraine
30-49	Back problem	nutrition disorder	Psychotropic poisoning	genital prolapse
50-69	Osteoarthritis	Osteoarthritis	Other endocrine disorder	Abdominal pain
70 or older	Device/ implant	Acute CVD	Other benign neoplasm	genital prolapse

Table 3. Precision and recall of mining result by CMARM and CFPgrowth

Min_sup(%)	Precision		Recall	
	CMARM	CFPgrowth	CMARM	CFPgrowth
0.8	91.04	99.99	75.56	75.76
1.0	90.52	100	77.56	77.56
1.5	91.69	100	78.78	78.78
2.0	97.52	99.98	82.03	82.03
3.0	99.40	99.99	83.04	83.04

**Table 4. Association rules for frequent & rare disease for dataset of year 2013 at  $S_0 = 5$ . Here, c is the confidence percentage**

Age group	Frequent disease						Rare disease					
	Male	Rules	C	Female	Rules	C	Male	Rules	C	Female	Rules	C
0-17				Liveborn								
		Liveborn			Newborns and other neonates with conditions originating in the perinatal period	50						
18-29				Other pregnancy compl								
	Crush/internal injury	Diseases and disorders Of respiratory system	12		Pregnancy, childbirth And the puerperium	50	Chest pain	Diseases and disorders Of the circulatory system	50	Headache / migraine	Migraine and other headaches	50
30-49				Other nutrition disorder								
	Back problem	Diseases and disorders of the musculoskeletal system and conn tissue	48		Endocrine, nutritional and metabolic diseases and disorders	50	Psychotropic poisoning	Poisoning of medicinal agents	50	Female genital prolapsed	Diseases and disorders of the female reproductive system	50
50-69				osteoarthritis								
		Diseases and disorders of the musculoskeletal system and conn tissue	50		Diseases and disorders of the musculoskeletal system and conn tissue	50	other endocrine disorder	Endocrine, nutritional and metabolic diseases and disorders	50	abdominal pain	abdominal pain	50
70 or older				Acute CVD								
	device/implant/ graft comp	Diseases and disorders of the circulatory system	24		Diseases and disorders of the nervous system	50	Other benign neoplasm	Diseases and disorders of the digestive system	50	Female genitl prolapse	Diseases and disorders of the female reproductive system	50

data statistics. Second, CMARM could be mainly useful in case of data sets that are dynamic due to its one scanning of file access and hence frequently updated. Third, due to distributed nature of CMARM, it may present a logical advantage predominantly on the more modern processing units such as cluster-based and graphical processor units. Hence, one can hope to tackle efficiency of CMARM

implementation further from a software/hardware aspect.

**REFERENCES**

- Gupta, A.; Arora, R.; Sikarwar, R. & Saxena, N. Web usage mining using improved frequent pattern tree algorithm. *In* IEEE conference on communication system and network technologies, 2014, 573-578.

doi: 10.1109/ICICICT.2014.6781344

2. Ghosh, A. & Nath, B. Multi-objective rule mining using genetic algorithm. *Information Science*, 2004, **163**, 123-133.  
doi: 10.1016/j.ins.2003.03.021
3. Bhandarkar. MapReduce programming with Apache Hadoop. *In IEEE International Conference on Parallel & Distributed Processing Symposium (IPDPS)*, 2010. doi: 10.1109/IPDPS.2010.5470377
4. Agnihotri, D.; Verma K. & Tripathi P. Pattern and cluster mining on text data. *In IEEE conference on communication system and network technologies*, 2014, 428-432. doi: 10.1109/CSNT.2014.92
5. Afrati, F.; Fotakis D. & Ullman J. Enumerating subgraph instances using map-reduce. *In International Proceedings of IEEE 29<sup>th</sup> International Conference Data Engineering*, 2013, 62-73. doi: 0.1109/ICDE.2013.6544814
6. Ghorashi, S.H.; Ibrahim. R.; Noekhah, S. & Dastjerdi, N.S. A Frequent pattern mining algorithm for feature frequent pattern mining algorithm for feature extraction of customer reviews. *Int. J. Comp. Sci.*, 2012, **9(4)**, 29-35.
7. Liu, G.; Zhang, H. & Wong, L. A Flexible approach to finding representative pattern sets. *IEEE Trans. Knowledge Data Eng.*, 2014, **26(7)**, 1562-1574. doi: 10.1145/2339530.2339543
8. Chen, G. P.; Yang, Y. B. & Zhang, Y. Mapreduce-based balanced mining for closed frequent itemset. *In International Proceedings of IEEE 19<sup>th</sup> International Conference Web Service*, 2012, 652-653. doi: 10.1109/ICWS.2012.19
9. Health data repository [Online] Available: [http://www.healthdata.gov/dataset/resource/c77ebf0e-4189-4bd8-b887-be269520ebf?f\[0\]=field\\_tags%3A996](http://www.healthdata.gov/dataset/resource/c77ebf0e-4189-4bd8-b887-be269520ebf?f[0]=field_tags%3A996) [Accessed on 10 August 2015].
10. Zhu, Jie.; Li, Juanjuan.; E, Hardesty.; Jiang, Hai. & Li, Kuan Ching. GPU-in-Hadoop: Enabling mapreduce across distributed heterogenous platforms. *In IEEE 13<sup>th</sup> International Conference on Computer and Information Science (ICIS)*, 2014, 321-326. doi: 10.1109/ICIS.2014.6912154
11. Mutakabbir, K.; Mahin, S. & Hasan, M. Mining frequent pattern within a genetic sequence using unique pattern indexing and mapping techniques. *In IEEE Conference on Informatics Electronics & Vision*, 2014, 1-5. doi: 10.1109/ICIEV.2014.6850729
12. Srinivas, K.; Rani, B. Kavihta. & Govrdhan, A. Applications of data mining techniques in healthcare and prediction of heart attacks. *Int. J. Comp. Sci. Eng.*, 2010, **2(2)**, 250-255. doi: 10.17485/ijst/2016/v9i39/102078
13. Cagliero, L. & Garza, P. Infrequent weighted itemset mining using frequent pattern growth. *In IEEE transactions on knowledge and data engineering*, 2014, **26(4)**, 903-915. doi: 10.1109/TKDE.2013.69
14. Wang, L.; David.; Lok, Wai. & Cheng, R. Efficient mining of frequent item sets on large uncertain databases. *IEEE Trans. Knowledge Data Eng.*, 2012, **24(24)**, 2170-2183. doi: 10.1109/TKDE.2011.165
15. Karun, A. & Chitharanja, A. A Review on Hadoop HDFS infrastructure extensions. *In IEEE Conference of Information & Communication Technologies (ICT)*, 2013, 132-137. doi: 10.1109/CICT.2013.6558077
16. Srinivas, K.; Kavihta, B. & Govrdhan, A. Applications of data mining techniques in healthcare and prediction of heart attacks. *Int. J. Comp. Sci. Eng.*, 2010, **2(2)**, 250-255. doi: 10.5120/ijca2016911011
17. Bhuiyan, M. A. & Hasan, M. A. An iterative mapreduce based frequent subgraph mining algorithm. *IEEE Trans. Knowledge Data Eng.*, 2015, **27(3)**, 608-620. doi: 10.1109/TKDE.2014.2345408
18. Akbarzadeh-T, M. R. & Soltani, A. Confabulation inspired association rule mining for rare and frequent itemsets. *IEEE Trans. Neural Networks Learning Systems*, 2014, **25(11)**, 2053-2064. doi: 10.1109/TNNLS.2014.2303137
19. Majali, Jaimini.; Niranjana, Rishikesh.; Phatak, V. & Tadakhe, Omkar. Data mining technique for diagnosis and prognosis of cancer. *Int. J. Adv. Res. Comp. Comm. Eng.*, 2015, **4(3)**, 613-616. doi: 10.1007/978-81-322-2553-9\_38
20. Srivastava, Neha & Gautam, Jyoti. Analysis of Medical domain using CMARM: Confabulation mapreduce association rule mining algorithm for frequent and rare itemset. *In International journal of advanced computer science and applications (IJACSA)*, 2015, **6(11)**, 224-228. doi: 10.14569/IJACSA.2015.061129
21. Chen, Q.; Qiu, Q.; Li, H. & Wu, Q. A neuromorphic architecture for anomaly detection in autonomous large-area traffic monitoring. *In IEEE/ACM International Conference on in Computer-Aided Design (ICCAD)*, 2013, 202-205. doi: 10.1109/ICCAD.2013.6691119
22. Chen, Q.; Qiu, Q.; Wu, Q.; Bishop, M. & Barnell, M. A confabulation model for abnormal vehicle events detection in wide-area traffic monitoring. *In IEEE International Inter-Disciplinary Conference on cognitive methods in situation awareness and decision support (CogSIMA.)*, 2014, 216-222. doi: 10.1109/CogSIMA.2014.6816565
23. Chen, Qiuwen.; Wu†, Qing.; Bishop†, Morgan.; Linderman†, R. & Qiu, Qinru. Self-structured confabulation network for fast anomaly detection and reasoning. *In Air force research laboratory, Information Directorate*, 2015. doi: 10.1109/IJCNN.2015.7280371
24. Bai, S. & Bai, S. The maximal frequent pattern mining of DNA sequence. *In IEEE conference on informatics electronics and vision*, 2009, 2326. doi: 10.1109/CIT.2005.106

## CONTRIBUTORS

**Ms Neha Srivastava** received her B.Tech (Computer Science and Engineering) from UP Technical University, Lucknow, 2012 and M.Tech from JSSATE, Noida, in 2015. Currently, she is working with JSSATE, Noida as assistant professor. Data mining is her core research interest, within data mining she works on problems related to medical data mining, big data, and cognitive learning.

**Dr Jyoti Gautam** received her BE (Instrumentation and control engineering) from Delhi Institute of Technology, Delhi, and ME (Computer Technology and Application) from Delhi College of Engineering, Delhi and PhD in semantic web mining from Gautam Buddh University, Noida. She is a Head of department of computer science and engineering and coordinator in JSSATE, Noida. She has published more than 15 research articles on semantic web mining in conferences and journals. Her area of interest is semantic web mining.