European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 7, Issue 4, 2020

Development of Top K-Association Rule Mining for Discovering pattern in Medical Dataset

Aakriti Sharma[1], *Anjana Sangwan[2], Blessy Thankchan[3], Sachin Jain[4], Veenita Singh[5], Shantanu Saurabh[6]

[1,2]*Department of Computer Science & Engineering, Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur, India*
[3,4]*School of Computer & Systems Sciences, Jaipur National University*
[5]*Department of Business Studies in the discipline of Commerce, RGSC, Banaras Hindu University.*
[6]*Faculty of Commerce, Maharaja Sayajirao University of Baroda, Vadodara, Gujarat*
[1]*aakritivashishtha@gmail.com*
*[2]*sangwan.anjana@gmail.com*
[3]*blessyt218@gmail.com*
[4]*sachinjain4u@gmail.com*
[5]*veenitasingh1990@gmail.com*
[6]*shantanu.saurabh-cmr@msubaroda.ac.in*

Abstract:** Association rules consist of the discovery of association between mining transaction items. This is one of the most important information mining jobs. It has been integrated into many commercial data mining software and has a wide variety of applications on a number of domains. So, computing the prediction rules in top rank data set is very difficult task. Finding the pattern in large data set require memory computational power high rate of I/O. and it is possible only on high computational machine. In this paper, selection of parameter which is used to compute is chosen based on minimum support and minimum confidence value. In this paper proposed a new algorithm which generates the association rule for the input parameters to finding the pattern in large data set. The algorithm starts searching the rules. As soon as a rule is found, it is added to the list of order rules list by support. The list is used so far to maintain top N rules found. Once valid rules are found, the minimum support for the internal minsup variable list is raised to support the rule. When the Minsup value is raised, the search space is robbed while searching for more rules. Then, every time a valid rule is found, the list is inserted into the list, the lists that are not listed in the list are excluded from the list and the minsup is raised for the price of the least fun rules in the list. Result shows that new method is efficient technique to mine data set from standard data with minimum configuration system.

***Keywords:** Data Mining, Clustering, Classification, K-Means, Data Processing, Performance Optimization, Medical dataset*

1. INTRODUCTION:

Data and figures are real facts. Numbers sign and symbols may be given to the data. However, data are meaningless without explanations; it is just only signs or symbols. It is raw and can

be used another way. Intrinsically, data is of little value and the information is generated by a collection of valuable data. In other words, information is called meaningful data. It collects meaningful data and communicates its importance to people and allows awareness in a community of information [1]. If the information is processed, it produces knowledge. This information offers an effective path for everything, but it does not offer an application of itself that can be more derived by it. Wisdom is a non-predictable process that is extrapolated and non-definitive. The category of relevant knowledge is intelligence. Since we know how to

find the right thing. We know the right thing. Because information aggregation is so we can intertwine good or evil with good or evil [2]. Wisdom is part of the human condition only. Human beings fail in every field without using their knowledge. Since knowledge gathering is known as wisdom, we can say that knowledge plays a very important role in human activity. Figure 1 It represents the flow from data to knowledge and finally from knowledge to knowledge, and shows the flow of transfer between phases. Time oriented refers the

Fig. 1. Discovering pattern in Medical Dataset

data which is stored in the database are related to the time. These databases are logged activities it means they are related to stock market [3]. In this, sometimes there is a need of challenging real time anal- ysis because this database has a record of incoming data. Study of different-2 variables and relations between them shows the way of data mining in such type of database [4]. Data mining in such type of database use the concept of movements and prediction of variables.

2. LITERATURE SURVEY

Association rules consist of the discovery of association between mining trans- action items. This is one of the most important information mining jobs. It has been integrated into many commercial data mining software and has a wide variety of applications on a number of domains [5].As the association rules are called mining problems as follows. Let me be a limited set of items = {a1, a2 ... a} A transaction is a set of transaction transactions T = {t1, t2 ... tm} where each transaction $\subseteq$ (($1 \leq x \leq m$) represents a set of items bought by the client at a particular time. One of the items item X I. The support of an item X is identified as Super (X) and is defined as the number of transactions containing X. An association rule is the $X \rightarrow Y$ two items x, e.g. , $X \subseteq I$ $X \cap Y = \emptyset$.

The Apariri Association's rule [6] was a major development in the history of mining, the non-

pragmatic algorithm was first proposed by author [7]. AIS is only for a direct ward method, which requires a lot of passage on the database, many candidates create items and preserve the counter of each candidate, but most of them are not frequencies. Most algorithms described above reduce the number of scans on the Apriori algorithm and database, try to prove the improof of some pass-through skills, reduce the size of the database to scan each pass; Using trick and sample techniques by different techniques. There are two disadvantages of the Apriori algorithm. First, the process of apriori algorithm generation generation is very complex. It uses most of the time and space. Secondly, multiple passes on the database. Apriori series algorithm breaks two bottlenecks, using the structure of the trees, the association rules are designed to work on some of the mining. FP-Tree [Han et al. 2000], frequent pattern mining[8], new re- search development of the association rule mining, which Apriori crosses two limitations. This algorithm has three main advantages. These discussions are as follows. Firstly, the FP-Tree database reduces the structure compared to the main database because only frequently items are considered and other varnish items are omitted. Some algorithms have been proposed to find top rules [9]. But they are ineffective. Each algorithm has its limitations, such as the FP-tree algorithm is difficult to use on an interactive mining system. During the interactive mining process, users can change the threshold support according to the rules. But if we change support in the FP-tray, then the whole process of the mining will be revived. Another more limitation of the FP-tree is that it is not suitable for growing mining.

3. METHOD AND TECHNOLOGY

This section describes the concept, technology; proposed method and snapshot of GUI of new proposed method Mining of Top Raked Data Set Form Standard Data.

3.1 Methodology/Algorithm

Mining association rules in dataset with user given minconf and minsup is big challenge, many researches also done in past and there is many method (algorithm) available for mining association rule. In which “Mining Top-k association rules” is one of them. Where K is a number of rules users wants to generate. This algorithm is suited for our problem solution but they are applied to mining streams or mining non-standard rules and not remove duplicate data set in output [10] .

3.2 Proposed Algorithm

As a transaction table the Top Rank algorithm enters a number of n rules the user needs to learn and uses a threshold minConf. The TopRank algorithm initials an intrinsic minsup value of 0, adopting the principal concept the algorithm then continues to search the rules. If a rule is observed, it will be applied by assistance to the list of order rules. In this point the list is used to hold top N rules. The minimum support for the internal minsup variable list is increased to follow the law until the rules are clear. The quest room is robbed when searching [11] for further laws while the Minsup value is elevated. Instead the list will be added into the list where a correct rule is identified, lists not included in the list will be omitted and minsup will be set at a premium for least fun rules in the list. The algorithm will continue to follow further rules before a rule is reached, meaning that the rules of the top ranking have been discovered.

Algorithm 1: Algorithm for Finding Top ranked data set values from standard Data set

1: Initialization of process
2: Input parameters are provided like database transition value, parameters value , minimum confidence value .
3: Initialize the minimum support value to zero.
4: do 1 to 3 until parameters value are remain

5: If supp (rule) => minsup and conf (rule) => minconf then it adds a list of rules (RULELIST). A list of top N rules ordered by RULELIST support.
6: Set minsup = support of rule with the lowest support in RULELIST
7: Exit

Output of Implemented Method

Figure 2 show the file storing transaction data where each row is group of items generated by transaction. Numbers in row represent item id.

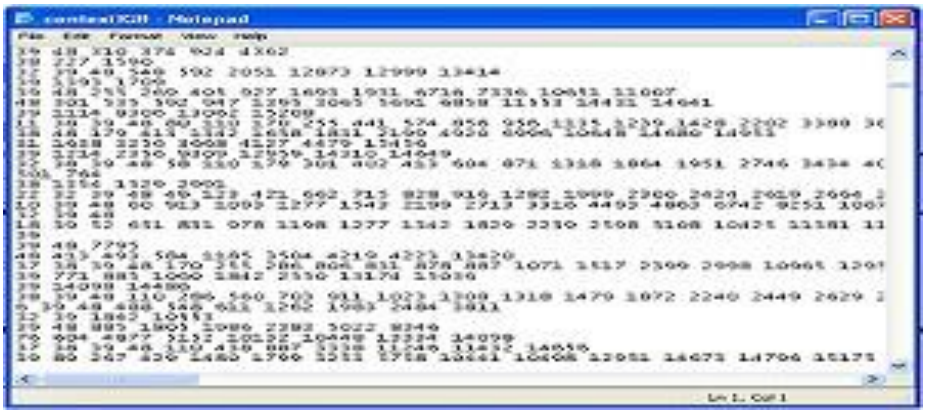

Fig. 3. Flow chart of Top Ranked rules from Standard data set

Graphical user interface built on JAVA Applet. In GUI window user can select method that could be Apriori, Top ranked data set, FP-Tree, with input data source file where transnational data is placed and select output file. Both output and input file in .txt format. It has option where user wants to display output file at same time or not. figure 3show the detail of how input values taking for selected mining algorithm. Here user can select transnational file that is store in secondary memory [12] and also select file where he want to save output result. The output of algorithm with option to show result files LIST.txt on same time. Output box show the process statistics for memory and time utilization and Minsup of transaction data for given input parameters.

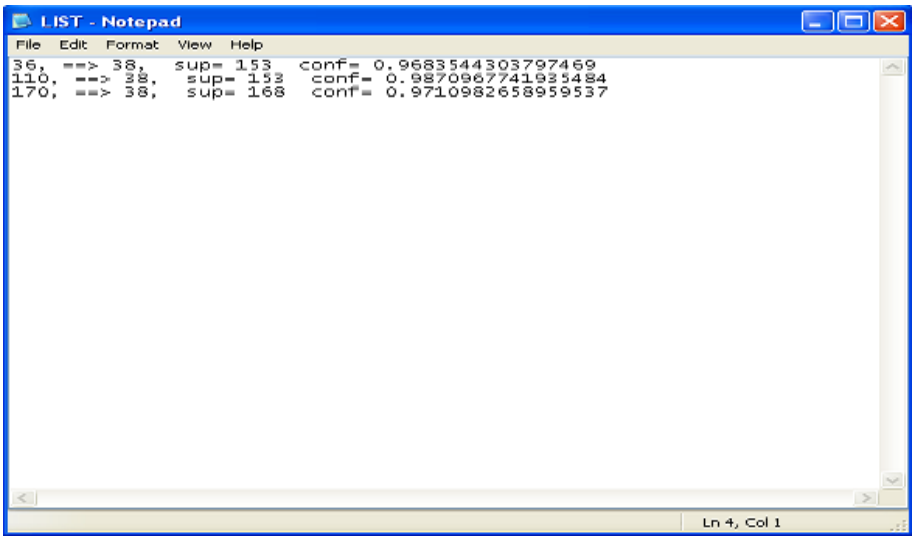

Fig. 4. Output of selected algorithm from GUI

4. Results

Result Comparison for 1000 Transaction Data:

For testing initial condition for new method taking 1000 transaction data. Result comparison table is shown below:

Table 1. Result comparison for 1000 Transaction Data taking as Input

Method use	Input Parameters			Rules count	Memory Use	Execution Time
	minsup	minconf	No of Rule			
Apriori	1	0.1	-	19	12.96	136
FP	1		-	19	14.9	132
Top Ranked Data Set	-	0.1	2	2	14.62	62

First Run of algorithms with input parameters minsup=1 and minconf=0.1 Apriori algorithm will mine 19 rule, MS priori mine19 rule, HMINE mine 15 rule, Relim mine 15 rules with taking average time 15 ms and other hand TOP Ranked Data Set will mine two rule that is given by user and also the execution time of algorithm is half of other mining algorithms[13]. Show in figure 5.2. For More Accuracy of Results Testing implement algorithm with 10,000 transaction data.

Result Comparison for 10,000 Transaction Data:

For Second level testing use 10,000 transaction data.

Table 2. Result comparison for 10000 Transaction Data taking as Input

Method use	Input Parameters			Rules count	Memory Use	Execution Time
	minsup	minconf	No of Rule			
Apriori	1	-	-	400	132.96	336
FP	1	01	-	346	130.9	342
Top Ranked Data Set	-	0.1	2	2	115.69	179

With run new method 10,000 data row find big difference between time ex- ecution, where Apriori, MS Apriori, HMINE and Relim will execute on average 260 ms Top Ranked Algorithm finish on time 179 and also clear that as the input transaction data increase mining[14] rules or rule count by algorithm is also increase here old mining algorithm will give average 350 rules that quantity is large when user want to display only those rule which is most frequent in data base. To test accuracy new approach on analysis factors required to increase input transaction data set row. So next run algorithms to 25,000 transaction data row and compare analysis factors for accuracy of result.

Result Comparison for 25,000 Transaction Data:

For third level testing use 25,000 transaction data.

Table 3. Result comparison for 25,000 Transaction Data taking as Input

Method use	Input Parameters			Rules count	Memory Use	Execution Time
	minsup	minconf	No of Rule			
Apriori	1	-	-	1270	332.96	1336
FP	1	01	-	1246	330.9	1342
Top Ranked Data Set	-	0.1	2	2	215.69	979

Testing Top Ranked Data Set method with Data set row 25,000 will shows that algorithm will

take less time and memory with mine 2 data set as given by user same. Results show that new algorithm take 979 ms process execution time[15] and 215 MB of RAM (primary memory) With compare to average rule mine on old algorithms. Old algorithms mine average 1250 rule on given data set with average process time 1250 ms and 328 MB of RAM that ration is too high. To find more accuracy in result for new method required large data set as application of data mining area like super market where owner want to mine frequent pattern[16] have large data set which can be generated in Two, Three or more years of transaction so next result analysis implement algorithm on 50,000 transactional data set row.

Result Comparison for 50,000 Transaction Data:

For fourth level testing use 50,000 transaction data.

Table 4. Result comparison for 50,000 Transaction Data taking as Input

Method use	Input Parameters			Rules count	Memory Use	Execution Time
	minsup	minconf	No of Rule			
Apriori	1	-	-	2770	532.96	2336
FP	1	01	-	2546	530.9	2442
Top Ranked Data Set	-	0.1	2	2	315.69	1023

Testing algorithm with 50,000 transaction data set row show that mine two top ranked data set with 0.1 minconf takes 315.69 MB memory and 1023 ms execution time that values is less than old mining approached. For final test we take super market transaction data that is available in General Public Licenses and contain more than 1,00,000 data set row . Each row contain group of item set generated by each sell. Data file contain transaction row of each row and row contain item ids sold on transaction. If algorithm take less time and memory for mine two data set rows in 1,00,0000 transactions than it is best approach for mine data in compare to previous mine methods.

Result Comparison for 1, 00,000 Transaction Data:

For fifth level testing 1, 00,000 transaction data use.

Table 5. Result comparison for 100000 Transaction Data taking as Input

Method use	Input Parameters			Rules count	Memory Use	Execution Time
	minsup	minconf	No of Rule			
Apriori	1	-	-	11435	756	5300
FP	1	01	-	11343	798	5065
Top Ranked Data Set	-	0.1	2	2	315.69	1023

Mine Two Top Ranked Data from Super Market Data set that contain 1,00,000 data set show that process take 479 MB of ram and 2789 ms execution time . That values show that new approach is best suited for large data set also. Form all previous tests also show that the time and memory increment with large data is less than other mining approach. Old method will mine all rules on other hand top ranked Data Set mine algorithm will stop after mine user give

top data set value. Result also arrange in descending order according to support of item set in transaction data.

Result Analysis

Considering the result comparison of 1000, 10000, 25000, 50000 and 100000 shows that new method will take less memory and time. The most advantage of new method is user choose how many rows, which want to display. This is important because old mining methods will display all possible data set row according to minsup and minconf, thus user can't predict Top frequent items acceding to support in data set. Figure 7 and 8 shows the time and memory comparison with all previous methods and new approach "Top Ranked Data Set".

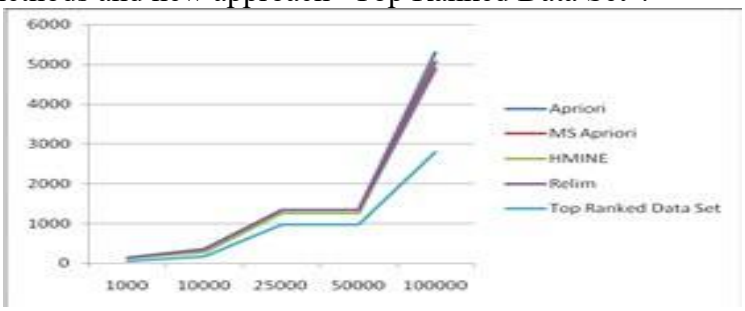

Fig. 7. Comparison of Execution Time with Different Data Row Value in Data Set

Second advantage of the method is it can reduce duplicate row or redundant data set and produce output contain unique rows. In old mining approach it is biggest issue, output contain more than many common rows that will increase when data is too large.

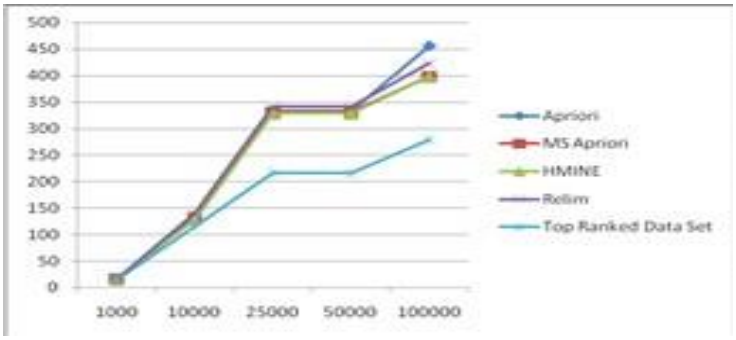

Fig. 8. Comparison of Memory Utilization with Different Data Row Value in Data Set

The third advantage is output store in .txt file and user also able to select output file thus he can compare and store result permanently in secondary memory. Thus, all testing spastics show that new method is efficient technique to mine data set from standard data with minimum configuration system.

5. CONCLUSION

Every algorithm defines for mining data or information has some limitation. As in testing condition new algorithm required minimum 1GB ram and required data set in text file with .txt format or Standard Data Set, where each row represent group of items and each column value represent item identity num- ber. Testing show that mine information with large data set required minimum 3GB ram and duel core or above processor. High speed process and RAM re- duce the size of memory and time. The third advantage is output store in .txt file and user

also able to select output file thus he can compare and store result permanently in secondary memory. Thus, all testing spastics show that new method is efficient technique to mine data set from standard data with minimum configuration system. Mining of Top Ranked dataset from standard dataset has immense the scope in future. It can be helpful for decision support System and future prediction. It can be also helpful if implementation of relational database using database management software like SQL, ORACLE. It gives more efficient output and support large dataset.

6. REFERENCES

[1] R. C. Agarwal, C. C. Aggarwal, and V. Prasad, "A tree projection algorithm for generation of frequent item sets," Journal of parallel and Distributed Computing, vol. 61, no. 3, pp. 350–371, 2001.
[2] P. Jadwal and M. R. Dave, "An improved and customized ik means for avoiding similar distance problem."
[3] [V. S. Tseng, C.-W. Wu, P. Fournier-Viger, and S. Y. Philip, "Efficient algorithms for mining top-k high utility itemsets," IEEE Transactions on Knowledge and data engineering, vol. 28, no. 1, pp. 54–67, 2016.
[4] P. K. Jadwal, S. Jain, U. Gupta, and P. Khanna, "Clustered support vector machine for atm cash repository prediction," in Progress in Advanced Computing and Intelligent Engineering. Springer, Singapore, 2019, pp. 189–201.
[5] M. Deshpande and G. Karypis, "Item-based top-n recommendation algorithms," ACM Transactions on Information Systems (TOIS), vol. 22, no. 1, pp. 143–177, 2004.
[6] M. J. Zaki, "Scalable algorithms for association mining," IEEE transactions on knowledge and data engineering, vol. 12, no. 3, pp. 372–390, 2000.
[7] T. M. Quang, S. Oyanagi, and K. Yamazaki, "Exminer: an efficient algorithm for mining top-k frequent patterns," in International Conference on Advanced Data Mining and Applications. Springer, 2006, pp. 436–447.
[8] Kumar, A., & Sinha, M. (2019). Design and analysis of an improved AODV protocol for black hole and flooding attack in vehicular ad-hoc network (VANET). Journal of Discrete Mathematical Sciences and Cryptography, 22(4), 453-463.
[9] Kumar, A., Dadheech, P., Singh, V., Poonia, R., & Raja, L. (2019). An improved quantum key distribution protocol for verification. Journal of Discrete Mathematical Sciences and Cryptography, 22(4), 491-498.
[10] Dadheech, P., Goyal, D., Srivastava, S., & Kumar, A. (2018). A scalable data processing using Hadoop & MapReduce for big data. J. Adv. Res. Dyn. Control. Syst, 10, 2099-2109.
[11] Kumar, A., Goyal, D., & Dadheech, P. (2018). A novel framework for performance optimization of routing protocol in VANET network. J. Adv. Res. Dyn. Control. Syst, 10, 2110-2121.
[12] Kumar, A., & Sinha, M. (2019). Design and development of new framework for detection and mitigation of wormhole and black hole attacks in VANET. Journal of Statistics and Management Systems, 22(4), 753-761.
[13] Kumar, A., Dadheech, P., Kumari, R., & Singh, V. (2019). An enhanced energy efficient routing protocol for VANET using special cross over in genetic algorithm. Journal of Statistics and Management Systems, 22(7), 1349-1364.
[14] Dadheech, P., Kumar, A., Choudhary, C., Beniwal, M., Dogiwal, S., & Agarwal, B. (2019). An enhanced 4-way technique using cookies for robust authentication process in wireless network. Journal of Statistics and Management Systems, 22(4), 773-782.
[15] Kumar, A., Dadheech, P., Beniwal, M. K., Agarwal, B., & Patidar, P. K. (2020). A Fuzzy Logic-Based Control System for Detection and Mitigation of Blackhole Attack in

Vehicular Ad Hoc Network. In Microservices in Big Data Analytics (pp. 163-178). Springer, Singapore.

[16] A., Dadheech, P., & Chaudhary, U. (2020, February). Energy Conservation in WSN: A Review of Current Techniques. In 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE) pp. 1-8

[17] Abhishek Kumar, et al. "The state of the art of deep learning models in medical science and their challenges". Multimedia Systems. (2020).

[18] Abhishek Kumar, et al. "Efficient data transfer in edge envisioned environment using artificial intelligence-based edge node algorithm". Transactions on Emerging Telecommunications Technologies. (2020).

[19] Ambeth Kumar, V.D. et al. "Active volume control in smart phones based on user activity and ambient noise". Sensors (Switzerland) 20. 15(2020): 1-17.

[20] Vengatesan, K. et al. "Analysis of Mirai Botnet Malware Issues and Its Prediction Methods in Internet of Things". Lecture Notes on Data Engineering and Communications Technologies 31. (2020): 120-126.

[21] Vimal, V. et al. "Artificial intelligence-based novel scheme for location area planning in cellular networks". Computational Intelligence. (2020).

[22] Kumar, A. et al. "Comparative Analysis of Data Mining Techniques to Predict Heart Disease for Diabetic Patients". Communications in Computer and Information Science 1244 CCIS. (2020): 507-518.

[23] Sayyad, S. et al. "Digital Marketing Framework Strategies Through Big Data". Lecture Notes on Data Engineering and Communications Technologies 31. (2020): 1065-1073.

[24] Kumar, V.D.A. et al. "Exploration of an innovative geometric parameter based on performance enhancement for foot print recognition". Journal of Intelligent and Fuzzy Systems 38. 2(2020): 2181-2196.

[25] Vengatesan, K. et al. "Secure Data Transmission Through Steganography with Blowfish Algorithm". Lecture Notes on Data Engineering and Communications Technologies 35. (2020): 568-575.

[26] Lone, T.A. et al. "Securing communication by attribute-based authentication in HetNet used for medical applications". Eurasip Journal on Wireless Communications and Networking 2020. 1(2020).

[27] Vengatesan, K. et al. "Simple Task Implementation of Swarm Robotics in Underwater". Lecture Notes on Data Engineering and Communications Technologies 35. (2020): 1138-1145.

[28] Kesavan, S. et al. "An investigation on adaptive HTTP media streaming Quality-of-Experience (QoE) and agility using cloud media services". International Journal of Computers and Applications. (2019)

[29] Ankit Kumar, Vijayakumar Varadarajan, Abhishek Kumar, Pankaj Dadheech, Surendra Singh Choudhary, V.D. Ambeth Kumar, B.K. Panigrahi, Kalyana C. Veluvolu, Black Hole Attack Detection in Vehicular Ad-Hoc Network Using Secure AODV Routing Algorithm, Microprocessors and Microsystems, 2020, 103352, https://doi.org/10.1016/j.micpro.2020.103352