iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** IV **Month of publication:** April 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Optimized Association Rule Mining with Maximum Constraints using Genetic Algorithm

Rajdeep Kaur Aulakh
Department of Computer Science and Engineering

Abstract: Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database.. In this paper initially we applied Apriori algorithm in order to generate frequent item-sets and then frequent item-sets are used to generate association rules. After getting association rules from Apriori algorithm we applied Genetic Algorithm (GA) to obtain reduced number of association rules. For this we used Selection, Genetic Operators. We designed a new fitness function for the proposed algorithm. It is observed that this algorithm greatly reduces the problem of generation of huge association rules using Apriori algorithm. The implementation of the proposed algorithm is easier than other popular algorithm for association rule mining. The proposed algorithm performs much better when compared to Apriori algorithm and other previous technique used to optimize association rule mining.
Keywords- Association rule mining, Apriori algorithm, Led dataset.

I. INTRODUCTION

The size of data stored in database is growing rapidly and the development of new and efficient methods for extracting the useful information from this huge amount of data is one of the key research areas in which most of the researchers are working. Different users need different sort of knowledge depending on their respective usage. Therefore, efficient data techniques are required to find and analyze the required information. The various techniques of data mining help in building a better understanding of the data and in building characterization of the data that can be used as a basis for further analysis to extract valuable information from a large volume of data. In many cases, the algorithms generate an extremely large number of association rules, often in thousands or even millions. Further, the association rules are sometimes very large. It is nearly impossible for the end users to comprehend or validate such large number of complex association rules, thereby limiting the usefulness of the data mining results. Several strategies have been proposed to reduce the number of association rules, such as generating only "interesting" rules, generating only "non redundant" rules, or generating only those rules which satisfy certain other criteria such as coverage, leverage, lift or strength.

Data mining is the process of extracting interesting knowledge such as association rules, patterns, regularities or constraints from large amounts of data stored in databases, data warehouses, or other information repositories. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data analysis, data archaeology and data dredging. Many people treat data mining as a synonym for another popular used term Knowledge Discovery of Data (KDD) and others view data mining as an essential step in the process of knowledge discovery. The various steps of KDD are described below.

Data Integration: First of all, the data is collected and integrated from all the different data sources.

A. Data Cleaning

Data cleaning is the process of removing noise and inconsistent data. It is found that the data collected may contain missing values, errors, inconsistency or noise so it is required to clean the data to get rid of such anomalies.

B. Data Selection

After Data integration and data cleaning the process called data selection is performed. Data selection is done by using various sampling techniques of data. In data selection only the relevant or required data is retrieved from the database.

C. Data Transformation

Data transformation is the process of transforming the data from one form to another required form. Smoothing, Aggregation, Normalization etc, techniques are applied to perform data transformation.

D. Data Mining

Data mining is an essential process where different intelligent methods such as association rule analysis, clustering, classification, prediction etc. are applied in order to extract the interesting data patterns.

E. Pattern Evaluation

In this step, interesting patterns representing knowledge are identified based on given measures.

F. Knowledge Presentation

In this the discovered or mined knowledge is presented to the user by applying visualization and knowledge representation techniques

II. DIFFERENT TECHNIQUES FOR OPTIMIZATION OF ASSOCIATION RULE MINING

A. Optimization using Genetic Algorithm

Genetic algorithms are methods based on biological mechanisms, such as, Mendel's laws and Darwin's fundamental principle of natural selection. The most important biological terminology used in a genetic algorithm is [6]:

1) The chromosomes are elements on which the solutions are built (individuals).
2) Population is made of chromosomes.
3) Reproduction is the chromosome combination stage. Mutation and crossover are reproduction methods.
4) Quality factor (fitness) is also known as performance index, it is an abstract measure to classify chromosomes.
5) The evaluation function is the theoretical formula to calculate a chromosome's quality factor. Genetic algorithms simulate a population evolution process.

A problem represented by individuals since a population of solutions, operators which simulate interventions about the genome such as crossover or mutation in order to achieve a population of solutions increasingly adapted to the problem. This adaptation is evaluated by the quality factor (fitness).

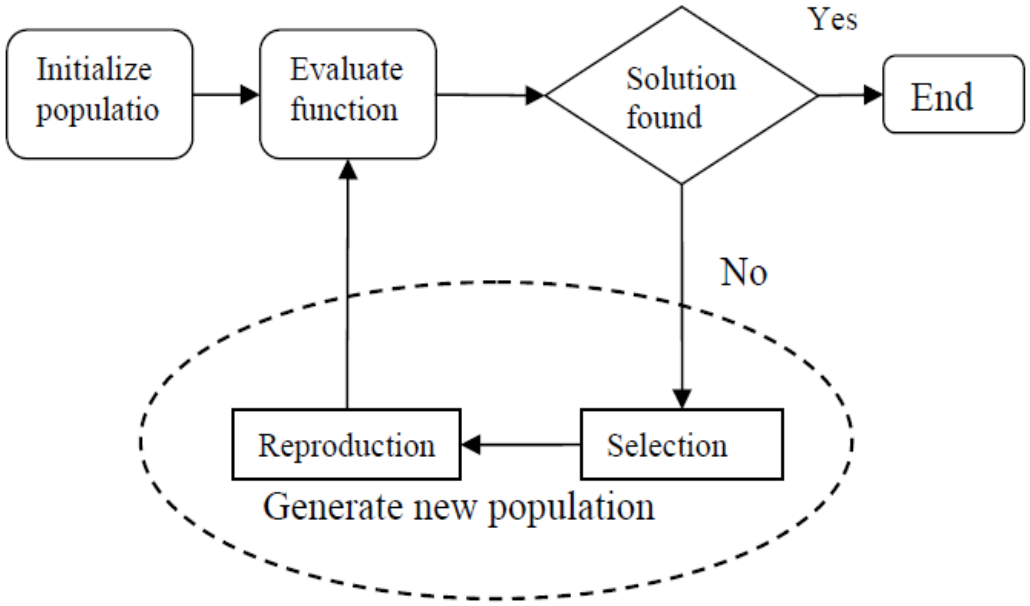


Fig.1 flow chart of genetic algorithm

III. RELATED WORKS

Many surveys have been conducted on previously developed optimization techniques. Logical organization of this literature survey proved to be a vital task for filling the gap between researches recently to improve performance of the association rule mining.

In 2013 Kannika Nirai Vaani M, E Ramaraj [1] presented an integrated method to derive effective rules from Association Rule

Mining using Genetic Algorithm. Currently Apriori algorithm uses the conjunctive nature of association rules, and the single minimum support factor to generate the effective rules. However the above two factors are alone not adequate to derive useful rules effectively hence authors have taken Apriori Algorithm as a reference and included disjunctive rules [2][3] and multiple minimum supports also to capture all possible useful rules. The concept is to integrate all into one that lead to a robust algorithm. And the salient feature of their work is introducing Genetic Algorithm (GA) in deriving possible Association Rules from the frequent item set in an optimized manner. Besides authors have taken one more add-on factor 'Lift Ratio' which is to validate the generated Association rules are strong enough to infer useful information.

In 2013 Basheer Mohamad Al-Maqaleh [4] proposed a multi-objective genetic algorithm approach for the discovery of interesting association rules with multiple criteria such that support, confidence and simplicity (comprehensibility). With Genetic Algorithm (GA), a global search can be achieved and system automation is developed. The proposed algorithm could identify interesting association rules from a dataset without having the user-specified thresholds of minimum support and minimum confidence. The most important difference between the proposed algorithm and the existing mining strategies is that this algorithm does not require the minimum support and minimum confidence thresholds.

In 2013, Shanta Rangaswamy and G.Shobha [5] presented a method in which genetic algorithm [6] is applied over the rules fetched from Apriori association rule mining. By using Genetic Algorithm the proposed system can predict the rules which contain negative attributes in the generated rules along with more than one attribute in consequent part. The goal of generated system was to implement association rule mining of data using genetic algorithm to improve the performance of accessing information from databases (Log file) maintained at server machine and to improve the performance by minimizing the time required for scanning huge databases maintained at server machines.

In 2012, K.Poornamala, R.Lawrance[7] presented an approach based on genetic algorithm. Genetic algorithm is used to optimize the large dataset. After that advanced frequent pattern tree is used to mine the frequent item set without generating conditional FP-tree. The proposed algorithm uses the GA to optimize the database to get high quality chromosomes and to find the frequent item sets using the Advanced FP algorithm from those high quality chromosomes. This algorithm mines the entire possible frequent item set with the compressed tree structure and without generating the conditional FP-tree.

In 2012, Sanat Jain, Swati Kabra[8] presented an Apriori-based algorithm that is able to find all valid positive and negative association rules in a support confidence framework. The algorithm can find all valid association rules quickly and overcome some limitations of the previous mining methods. Authors have designed pruning strategies for reducing the search space and improving the usability of mining rules, and have used the correlation coefficient to judge which form association rule should be mined.

In 2011, Peter P.Wakabi–Waiswa, Venansius Baryamureeba et al. [9] have proposed a multi–objective approach to generating optimal association rules using two new rule quality metrics: syntactic superiority and transactional superiority. These two metrics ensure that dominated but interesting rules are returned and not eliminated from the resulting set of rules. The introduction of the superiority measure causes more rules to be discovered which requires better presentation of the results to the user and it takes the algorithms longer to generate optimal rules. The weighted sum method [4] is the most popular approach used for multi–objective ARM where the fitness value of a candidate rule is derived using a linear transformation formula.

In 2011, Rupali Haldulakar and Jitendra Agrawal [10] designed a novel method for generation of strong rule. For which a general Apriori algorithm is used to generate the rules after that optimization techniques are used for optimized rules. Genetic algorithm is one of the best ways to optimize the rules .In this direction authors designed a new fitness function that uses the concept of supervised learning then the GA will be able to generate the stronger rule set. The new fitness function divides into two classes' c1 and c2 one class for discrete rule and another class for continuous rule.

In 2011, Wilson Soto and Amparo Olaya Benavides [11] proposed a genetic algorithm in his paper for discovery of association rules. The main characteristics of the proposed algorithm are: (1) The individual is represented as a set of rules (2) The fitness function is a criteria combination to evaluate the rule's quality – high precision prediction, comprehensibility and interestingness (3) Subset Size–Oriented Common feature. Crossover Operator (SSOCF) is used in the crossover stage (4) mutation is calculated through non–symmetric probability and selection strategy through tournament. The proposed algorithm is an alternative to find a set of association rules with high precision prediction, comprehensibility and interestingness. The use of kind of crossover (SSOCF) on the proposed algorithm allows the sets of useful information continuance in order to be inherited, regardless the number of generations individuals have.

In 2011, Huang Qiu-yong, Tang Ai-long et al. [12] presented an Apriori's optimization algorithm. Because there are some problems about some optimization algorithms of Apriori such as they consume large memory space although they reduce the numbers of database scanning, or the problem about the difficulties to realize programming. The algorithm first uses the order character of itemsets to reduce the times of comparison and connection when it connects and generates the candidate itemsets, then compresses the candidate itemsets according to the following condition: whether the number of element "a" in the frequent

K-itemsets is less than K or not. It is proved that the algorithm can not only realize programming easily but also improve the efficiency of mining association rules. The authors presented an Apriori's optimization algorithm based on reducing transaction. The algorithm is prone to programming and implementation and no additional storage space.

IV. TOOLS USED

A. *Matlab*

One most attractive aspect of MATLAB is that it is relatively easy to learn. It is written on an intuitive basis and it does not require in-depth knowledge of operational principles of computer programming like compiling and linking in most other programming languages. This could be regarded as a disadvantage since it prevents users from understanding the basic principles in computer programming. The interactive mode of MATLAB may reduce computational speed in most applications. The power of MATLAB is represented by the length and simplicity of the code. For example, one page of MATLAB source codes. Numerical calculation in MATLAB uses collections of well-written scientific/mathematical subroutines such as LINPACK and EISPACK. MATLAB provides Graphical User interface (GUI) as well.

1) *MATLAB Genetic Algorithm Toolbox:* The Genetic Algorithm and Direct Search Toolbox includes routines for solving optimization problems using:

a) Genetic Algorithm
b) Direct search

All the toolbox functions are MATLAB M-files, made up of MATLAB statements that implement specialized optimization algorithms.
To use the Genetic Algorithm and Direct Search Toolbox, you must first write an M-file that computes the function you want to optimize. The M-file should accept a row vector, whose length is the number of independent variables for the objective function and return a scalar. Calling the function ga at the Command line to use the genetic algorithm at the command line, call the genetic algorithm function ga with the syntax.
[x.fval]=ga@(fitnessfun,nvars,options)

a) Using the Genetic Algorithm GUI Tool

The Genetic Algorithm Tool is graphical user interface that enables you to use the genetic algorithm without working at the command line. To open the Genetic Algorithm Tool, enter gaoptim. To use the Genetic Algorithm tool, you must first enter the values of parameters:

(i) *Fitness Function*- The objective function you want to minimize.
(ii) *Number of variables*- the length of the input vector to the fitness function.

To run the GA, click the start button. The tool displays the results of the optimization in the status and result pane. You can change the options for the GA in the options pane.

B. *MATLAB Language*

This is a high-level matrix/array language with control flow statements, functions, data structures, input/output and object-oriented programming features. It allows both "programming in small" to rapidly create quick and dirty throw-away programs, and "programming in large" to create large and complex application programs.
Key features of MATLAB are:

a) High-Level language for technical computing.
b) Development, environment for managing code, files and data.
c) Interactive tools or iterative exploration, design and problem solving.
d) Mathematical functions for linear algebra, statistics, Fourier analysis, filtering, optimization and numerical integration.
e) 2-D and 3-D graphics functions for visualizing data.
f) Functions for integrating MATLAB based algorithms with external applications and languages such as C, C++, FORTRAN, Java, COM and Microsoft Excel.

C. *Pseudo Code of Genetic Algorithm*

The standard GA can be coded as in Figure 3.11.

```
While ( Generation_count =  max_generation)
{
Initialize populationP(0)
Evaluate populationP(0)
Generation_count = Genration_count+1
Select P(Generation_count) from
P(generation_count-1)
Crossover P(Generation_count)
Mutate P(Generation_count)
Evaluate P(Generation_count)
If (optimizationcriteriamet)
Break:
Outputbestsolution
}
```

Figure 3.11: Pseudo Code of Genetic Algorithm

D. Applied Algorithm

Algorithmic Structure: The proposed method for generating association rule by GA is as follows:

Step 1: Start

Step 2: Load a sample of records from the database that fits in the memory.

Step 3: Apply Apriori algorithm to find the frequent item sets with the minimum support. Suppose A is set of the frequent item set generated by Apriori algorithm.

Step 4: Set Z= 0 where Z is the output set, which contains the association rule.

Step 5: Input the termination condition of GA.

Step 6: Represent each frequent item set of A as a binary string using the combination of representation.

Step 7: Select the two members from the frequent item set using Roulette Wheel sampling method.

Step 8: Apply the crossover and mutation on the selected members to generate the association rules.

Step 9: Find the fitness function for each rule X□Y and check the following condition.

Step 10: If (fitness function > min confidence)

Step 11: Set Z = Z U {X □Y}

Step 12: If the desired number of generations is not completed, then go to Step 3.

Step 13: Stop.

V. RESULTS AND DISCUSSION

To evaluate the performance of the proposed algorithm, Extensive simulation experiments have been performed. The goal is to optimize the number of generated association rules. In this research work, the performance of proposed algorithm is compared with that of Apriori algorithm and previous work in terms of number of association rules generated.

A. Results on LED 7 Dataset

We applied our proposed algorithm on LED 7 dataset. We obtained 7 association rules and we calculated support, confidence, simplicity and fitness for each generated rule using eq. (3.1-3.4) defined in section 3.2. These values are shown in Table 4.4. The lowest value in the sup column is 0.197 and the lowest value in the conf column is 0.900 in Table 4.4. So, these two values are used as the threshold values of the minimum support and minimum confidence in the Apriori algorithm respectively. Based on these constraints, the Apriori algorithm would generate 14 association rules from the Led 7 dataset.

Table 5.1: Generated Rules from Led 7 Dataset

Sr. No.	Discovered Rules	Support	Confidence	Comp.	Fitness
1	Attr#4=1∧Attr#6=1∧ Attr#7=1→Attr#1=1	.352	.920	.431	.555
2	Attr#5=1 ∧ Attr#7= 1 → Attr#1=1	.331	.910	.500	.552
3	Attr#2=1 ∧ Attr#5= 0 → Attr#6=1	.267	.900	.500	.517
4	Attr#5=0 ∧ Attr#7= 0 → Attr#6=1	.252	.900	.500	.509
5	Attr#4=1 ∧ Attr#5= 1 ∧ Attr#7= 1→ Attr#1 =1	.239	.910	.431	.495
6	Attr#3=0 ∧ Attr#7= 1 → Attr#1 =1	.197	.920	.500	.489
7	Attr#1=1 ∧ Attr#2= 1 ∧ Attr#5= 1→ Attr#7 =1	.200	.910	.431	.492

Figure 5.1 shows the performance of proposed algorithm with compared to Apriori algorithm and previous work.

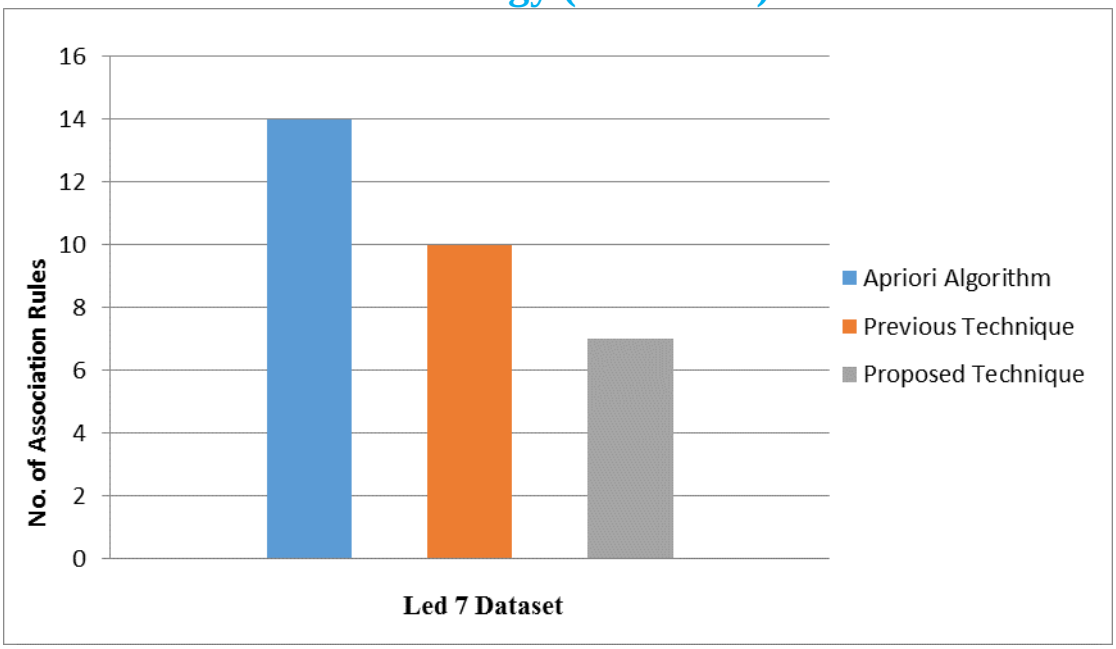


Figure 5.1: Comparison of performance for Led 7 Dataset

VI. CONCLUSION

Various association rule mining algorithms such as Apriori suffers from limitation of large number of association rules generation. An efficient method is developed in this research work to find the minimized number of association rules. The results of the experiments performed show the proposed model can attain considerable performance improvement in terms of the interesting association rules discovery and the number of discovered rules comparing to the Apriori algorithm. Comparison of the results obtained from proposed technique and other techniques shows that our approach is more efficient in terms of performance as compared to other previous algorithms. It is found that this algorithm finds the association rules in much easier and efficient. After applied Apriori algorithm on datasets, the process of optimization of association rules obtained from Apriori is performed using proposed GA

REFERENCES

[1] M. Kannika Nirai Vaani, E. Ramaraj."An Integrated Approach To Derive Effective Rules From Association Rule Mining Using Genetic Algorithm" Pattern Recognition, Informatics and Medical Engineering (PRIME), International Conference , pp 90 – 95, 2013.
[2] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques", Morgan- Kaufmann Publishers, 2000.
[3] R. Agrawal, R. Srikant, "Fast Algorithm for Mining Association Rules", Proc. of the Int. Conf on Very Large Database, pp. 487- 499, 1994.
[4] J. Han, J. Pei, Y. Yin, "Mining Frequent Patterns without Candidate Generation". Proc. of the ACM SIGMOD Int. Conf. on Management of Data, pp.1-12, 2000.
[5] J. Han, J. Pei, Y. Yin and R. Mao, "Mining Frequent Patterns without Candidate Generation: A Frequent Pattern Tree Approach", In Data mining and Knowledge Discovery, Vol. 8, pp.53-87, 2004.
[6] K. Poornamala, B. Lawrence, "A Frequent Pattern Tree Algorithm For Mining Association Rule Mining Using Genetic Algorithm," International Conference of Computing and Control Engineering, 2012.
[7] S. Jain, S. Kabra. "Mining & Optimization of Association Rules Using Effective Algorithm," International journal of Emerging Technology and Advanced Engineering (IJETAE), Vol.2, Issue 4, 2012.
[8] S. Jain, S. Kabra. "Mining & Optimization of Association Rules Using Effective Algorithm," International journal of Emerging Technology and Advanced Engineering (IJETAE), Vol.2, Issue 4, 2012.
[9] P. Wakabi-Waiswa, V. Baryamureeba, K. Sarukesi. "Optimized Association Rule Mining with Genetic Algorithms", In Natural Computation (ICNC), Seventh International Conference on, Vol. 2, pp. 1116-1120, 2011.
[10] M. Wang, Z. Qin, L. Caihui, "Multi-Dimension Association Rule Mining Based on Adaptive Genetic Algorithm.", IEEE International Conference , Vol. 2, pp. 150-153, 2011.
[11] Jun Gao, "A New Association Rule Mining algorithm and Its Applications", IEEE 3rdInt. Conf. on Advanced Computer Theory and Engineering (ICACTE), vol 5, pp. 122-125,2010.

[12] Li Juan and Ming De-ting, "Research of an association rule mining algorithm based on FP tree", IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS), Vol. 1, pp. 559-563,2010.

[13] Zhi Liu, Mingyu Lu, Weiguo Yi, and Hao Xu, "An Efficient Association Rules Mining Algorithm Based on Coding and Constraints", Proceedings of the 2nd International Conference on Biomedical Engineering and Informatics, pp. 1-5, 2009.

[14] Wanjun Yu , XiaochunWang, and Fangyi Wang, "The Research of Improved Apriori Algorithm for Mining Association Rules", 11th IEEE International Conference on Communication Technology Proceedings, pp. 513-516, 2008.

[15] Dongme Sun, Shaohua Teng, Wei Zhang and Haibin Zhu, "An Algorithm to Improve the Effectiveness of Apriori Algorithm", Proc. of 6th IEEE Int. Conf. on Cognitive Informatics", pp. 385-390, 2007.

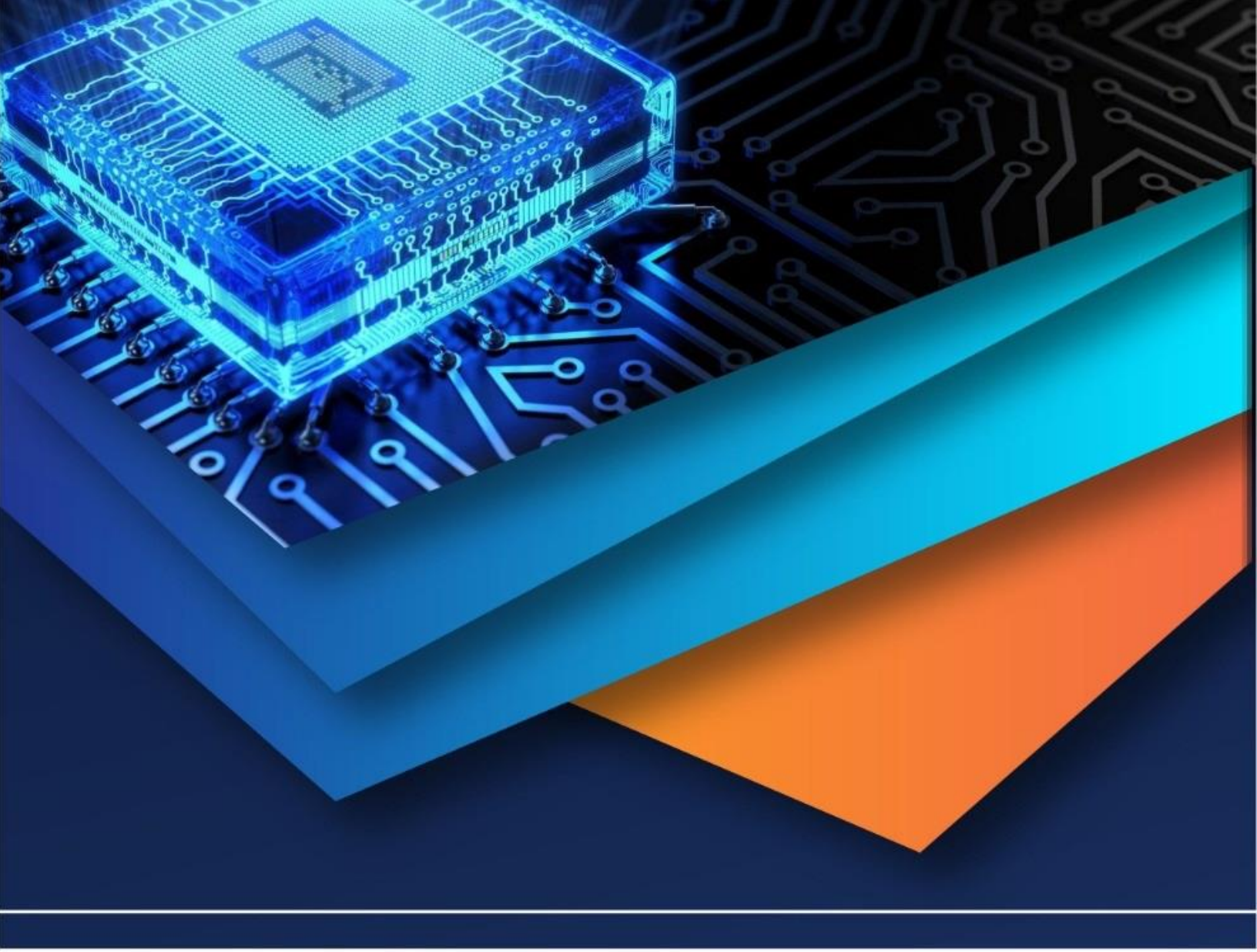
[16] Chin-Feng Lee and Tsung-Hsien Shen, "An FP-Split Method for Fast Association Rule Mining", Proc. of IEEE 3rd International Conference on Information Technology: Research and Education, June 27-30, pp. 459-463, 2005.

[17] Deepa, M. Kalimuthu. "An Optimization of Association Rule Mining Algorithm using Weighted Quantum behaved PSO", International Journal of Power Control Signal and Computation (IJPCSC), Vol.3, 2012.

[18] S. Dehuri, R. Mall, "Mining Predictive and Comprehensible Rules Using A Multi-Objective Genetic Algorithm", Advance Computing and Communication (ADCOM), India, 2004.

[19] J. Arunadevi, V. Rajamani. "Optimization of Spatial Association Rule Mining using Hybrid Evolutionary algorithm." International Journal of Computer Applications Vol. 1, Issue 19, 2010.

[20] Y. Cheung, A. Fu, "Mining Frequent Item sets without Support Threshold: With and Without Item Constraints", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, pp. 1052-1069, 1999.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY