Novel Method for Association Rule Construction from Crime Pattern

D Usha[1], K Rameshkumar[2]
Department of Information Technology, Hindustan Institute of Technology & Science
Chennai, Tamil Nadu, India
{dusha@hindustanuniv.ac.in} {krkumar@hindustanuniv.ac.in}

ABSTRACT: *Association rule mining aims to extract interesting correlations, frequent patterns, associations among sets of items in the transaction database. In this paper, association rule is constructed from the proposed rule mining algorithm. Efficiency based association rule mining algorithm is used to generate patterns and Rule Construction algorithm is used to form association among the generated patterns. This paper aims at applying crime dataset, from which frequent items are generated and association made among the frequent item set. It also compares the performance with other existing rule mining algorithm. The algorithm proposed in this paper overcomes the drawbacks of the existing algorithm and proves the efficiency in minimizing the execution time. Synthetic and real datasets are applied with the rule mining algorithm to check the efficiency and it proves the results through experimental analysis.*

Keywords: ARM, IRM, Rule Construct, Crime Dataset, Information Gain

Received: 18 August 2019, Revised 29 November 2019, Accepted 8 December 2019

DOI: 10.6025/jdp/2020/10/1/1-10

1. Introduction

Association rule mining is one of the important techniques in data mining and was first introduced by Agrawal [1]. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. Association rules are created by analyzing data from frequent patterns and using the criteria support and confidence to identify the most important relationships. Support is an indication of how frequently the items appear in the database. Confidence indicates the number of times the statements have been found to be true [17]. The existing rule mining algorithms are inefficient due to so many scans of database and also if the database is large, it takes too much time to scan the database. The proposed rule mining algorithm proves efficiency by reducing the execution time. The aim of this paper is application of crime pattern [5] and generates association rules from the mined frequent pattern with the help of rule construct algorithm and find the suitable measures to validate the rule.

2. Problem Statement

Association rule mining [2] helps to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two sub problems. One is to find frequent itemset from the database. Those itemsets whose occurrences exceed a predefined support threshold are called as frequent itemsets. The second problem is to generate association rule among frequent item set and validate the rule. The main drawback of the existing Apriori algorithm is generation of large number of candidate itemsets [4], which requires more space and efforts. Because of the above facts the algorithm needs too many passes and multiple scans over the whole database, so that it becomes waste and useless. The existing frequent pattern based classification also has some drawbacks. Since the frequent pattern growth approach lies on tree structure, it constructs the tree to store the data but when the data are large it may not be fit in main memory. During the computation process, the results become infeasible and over-fit the classifier. Also the existing FPM algorithm is not scalable for all types of data.

The proposed algorithm is used to retrieve the frequently occurred patterns from large amount of database with the help of newly constructed data structure. Based on frequent patterns mined from FPM algorithm, association rules are generated and it has to find out the rules that satisfy the predefined minimum support and confidence from a given database. The existing rule mining algorithm generates wide number of association rules which contains non interesting rules also. While generating rule mining algorithm [14], it considers all the discovered rules and hence the performance becomes low. It is also impossible for the end users to understand or check the validity of the large number of complex association rules and thereby restricts the usefulness of the data mining results. The generation of large number of rules [14] also led to heavy computational cost and waste of time. Various methods have been formulated to reduce the number of association rules like generating only rules with no repetition, generating only interesting rules, generating rules that satisfy some higher level criteria's etc.

3. Related Work

3.1 Constraint Based Association Rule Mining algorithm [13] (CBARM)

In the above algorithm constraints were applied during the mining process to generate only those association rules that are interesting to users instead of all the association rules [13]. The cost which is spend for mining the non interesting rules can also be saved. Usually constraints are provided by users, it can be knowledge based constraints, data constraints, dimensional constraints, interestingness constraints or rule formation constraints [13]. CBARM is to find all rules from a given data-set meeting all the user-specified constraints. Apriori and its variants only employ two basic constraints: minimal support and minimal confidence. However there are two points, one is some of the generated rules may be usefulness or not informative to individual users; another point is that with the constraints of minimal support and confidence those algorithms may miss some interesting information that may not satisfy them.

3.2 Rule Based Association Rule Mining Algorithm [12] (RBARM)

This algorithm builds a rule-based classifier from association rules. This approach overcomes some limitations of greedy methods like decision-tree, sequential covering algorithms [12], which considers only one attribute at a time. Rule based association rule mining algorithm mines any number of attributes in the consequent. Class association rules set the consequent to be the class label. The drawback of this algorithm is since the numbers of attributes are more, efficiency cannot be achieved fully.

3.3 Classification Based on Multiple Association Rule Mining Algorithms [12] (CMAR)

CMAR uses the FP-growth algorithm to mine the class based association rules [12]. It employs the CR tree structure to efficiently store and retrieve rules. It applies rule pruning whenever a rule is inserted in the tree: If R_1 is more general than R_2 and conf $(R_1) > conf(R_2)$: R_2 is pruned. The main drawback of CMAR algorithm is that the rules where the classes are not positively correlated are also pruned. CMAR considers multiple rules when classifying an instance and use a weighted measure to find the strongest class.

4. Improved Rule Mining Algorithm (IRM)

The improved rule mining algorithm increases the efficiency through the process of reducing the computational time as well as cost. It can be succeeded by reducing the number of passes over the database, by adding additional constraints on the pattern.

In legal applications, some rules will have less weightage and inefficient and some rules will have more weightage. Generation of the entire constructed rule will lead to waste of time. So, researcher likes to validate the rule to find the most efficient rule. The two important basic measures for association rules are support (Supp) and confidence (Conf) [14]. Support [166] [14] is defined as the percentage or fraction of records that contain XUY to the total number of records in the database. For example, if the support of an item is 0.5% it means that only 0.5% of the transaction contain purchasing of that item. Confidence of an association rule is defined as the percentage of the number of transactions that contain XUY to the total number of records that contain *X*. Confidence is the measure of strength of the association rule. For example, if the confidence of the association rule $X => Y$ is 75%, it means that 75% of the transactions that contain *X* also contain *Y* together.

The additional measure Information Gain [18] is a statistical property that measures the validity of the generated rules. The Information Gain calculates lift value and with the input of the lift value it calculates information gain by taking all the possible combination of rules that is generated from the proposed pattern mining algorithm. The proposed Improved Rule Mining algorithm consists of association rule generation phase.

The drawback of the above mentioned CBARM, RBARM and CMAR algorithm is that the number of rules can be extremely large. So, it takes more time to execute the process. The proposed IRM algorithm minimize the number of rules mined but proves the efficiency because of its limited and exactly predictable space overhead and is faster than other existing methods. It also increases the efficiency through the process of reducing the computational time as well as cost. It can be succeed by reducing the number of passes over the database, by adding extra constraints on the pattern. Association rule construction is a straight forward method. The rule construct algorithm uses the following procedure to construct association rule [1].

```
Procedure Gen_AR (k-itemsets)
begin
Step 1 : for all large k-itemsets l_k, k ≥ 2 do begin
Step 2 : H_1= { consequents of rules derived from l_k with one item in the consequent };
Step 3 : call rue_construct (l_k, H_1);
end procedure
Procedure rule_construct (l_k: large k-itemset, H_m: set of m-item consequents)
begin
Step 1 : if (k > m + 1) then begin
Step 2 : H_m+1 = Apriori_gen(H_m)
Step 3 : for all h_m+1 ∈ H_m+1 do begin
Step 4 : conf = support(l_k)/support(l_k- h_m+1)
Step 5 : if (conf ≥ MinConf) and (gain >- MinGain) then
Step 6 : output the rule (l_k-h_m+1) → h_m+1with confidence=conf
Step 7 : information gain = gain and
Step 8 : support = support(l_k)
         else
Step 9 :  delete h_m+1from H_m+1
end if
Step 10: call rule_construct (l_k, H_m+1)
end for
end if
end procedure
```

Figure 1. Rule construct procedure to construct association rules

4.1 Rule Evaluation Metrics

Support (s) : Fraction of transactions that contain both *P* and *Q*

Confidence (c) : Measures how often items in *Q* appear in transactions that contain *P*

{lonely house, bureau pulling} => chain hooks

$s = \sigma$ (lonely house, bureau pulling, chain hooks) / $|T| = 2 / 5 = 0.4$

$c = \sigma$ (lonely house, bureau pulling, chain hooks) / σ (lonely house, bureau pulling) = $2 / 3 = 0.67$

Information Gain

Information Gain is a statistical property that measures how well a given attribute separates the training samples according to their target classification. The measure of purity is called the information. Information Gain increases with the average purity of the subsets that an attribute produces. It is used to determine which attribute in a given set of training feature vectors is most useful for discriminating between the classes to be learned [16]. Information gain tells us how important a given attribute of the feature vectors is. Information Gain Feature Ranking is derived from Shanon entropy as a measure of correlation between the feature and the label to rank the features. Information Gain is a famous method and used in many papers [11]. The other method is CFS which uses the standardized conditional Information Gain to find the most correlated features to the label and eliminates those whose correlation is less than a user defined threshold [10]. It will use it to decide the ordering of attributes in the nodes of a decision tree.

Attribute Selection:

• Select the attribute with the highest information gain

• Let p_i be the probability that an arbitrary tuple in *D* belongs to class C_i, estimated by $|C_{i,D}| / |D|$

• Expected information (entropy) needed to classify a tuple in *D*:

$$Info\ (D) = -\sum_{i=1}^{m} p_i \log_2 (p_i)$$

• Information needed (after using *A* to split *D* into *v* partitions) to classify *D*:

$$Info_A\ (D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I\ (D_j)$$

• Information gained by branching on attribute *A*

$$Gain\ (A) = Info\ (D) - Info_A\ (D)$$

Association rule mining falls under the descriptive category [15]. Association rules aims in extracting important correlation among the data items in the databases [18]. Association rule [3] [15], basically extracts the patterns from the database based on the two measures such as minimum support and minimum confidence. To select the best measures for mining rules based on constraints such as multiple criteria is discussed in [11]. The support and confidence measures are described as stated in [8] for mining frequent itemset mining and association rule generation.

5. Dataset Description

The real and synthetic dataset are taken from UC Irvine Machine Learning Database Repository [6] and FIMI [7] repository. The below mentioned table 1 shows the number of transactions and attributes of real and synthetic datasets used in this evaluation. Typically, these real datasets are very dense, i.e., they produce many long frequent itemsets even for very high values of support threshold. Usually the synthetic datasets are sparse when compared to the real sets.

6. Experimental Study and Analysis

The Constraint Based Association Rule Mining algorithm is compared with the proposed Improved Rule Mining algorithm. The

Dataset	No. of Transactions	No. of Attributes
T40I10D100K	100000	942
Mushroom	8124	119
Gazella	59601	497
Crime Dataset	5000	83

Table 1. Synthetic and Real Datasets

below figure nos. 2, 3, 4 and 5 gives the comparative study of execution time between IRM and CBARM varying minimum support using Gazella [9], T40I10D100K and Mushroom dataset resp. The MinSupp is set as 70%, the MinConf is set as 75% and Information Gain is set as 60%. Most of the transactions are replaced by the earlier TID which are effectively used for support calculations and easy to construct frequent patterns.

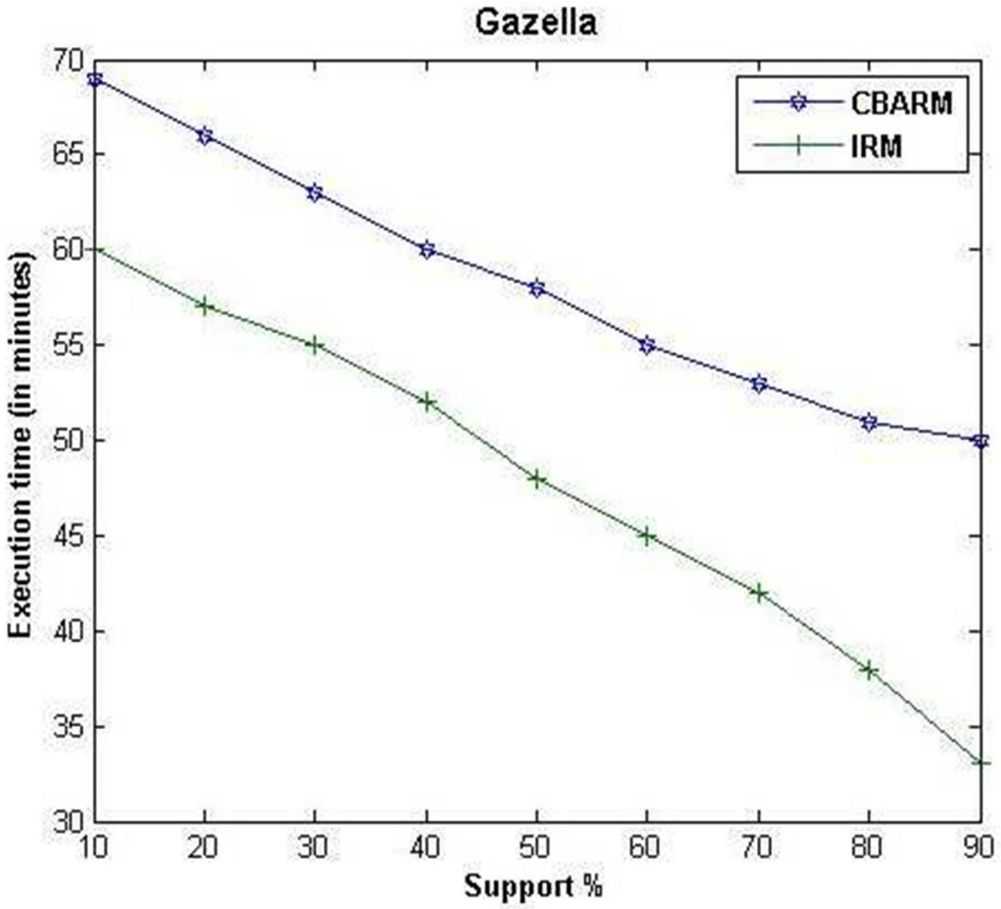

Figure 2. Comparison - execution time between IRM and CBARM algorithm with varying support in Gazella dataset

In the figure nos. 2, 3, 4 and 5, the performance of IRM algorithm is better than CBARM algorithm in case of small dataset and also in high density dataset. In TTC dataset, in the comparison of execution time, IRM is better than CBARM varying levels 10% to 20%. The execution time of IRM is reduced in case of Gazella and Mushroom dataset varying ratio between 10% to 14% respectively. In T40I10D100K, the IRM algorithm performs better than CBARM. When compared with CBARM algorithm, the IRM algorithm has reduced for about 30% to 35% of execution time. Hence it is proved that the proposed IRM algorithm performs better in high density dataset when compared with small dataset.

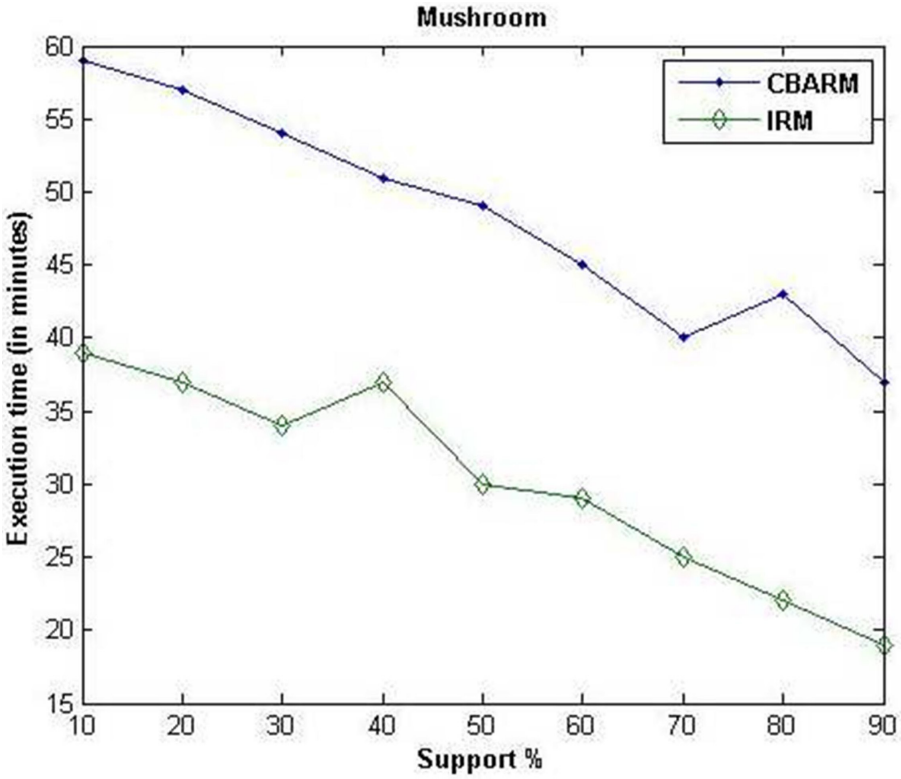

Figure 3. Comparison –Execution time between IBM and CBARM algorithm with varying support in Mushroom dataset

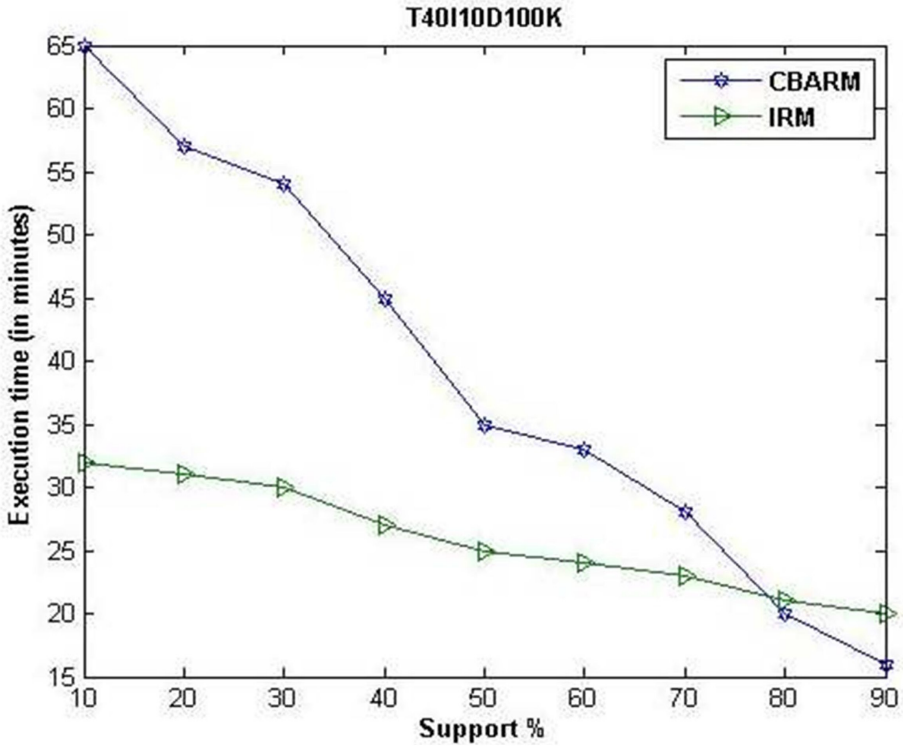

Figure 4. Comparison –Execution time between IRM and CBARM algorithm with varying support in T40I10D100K dataset

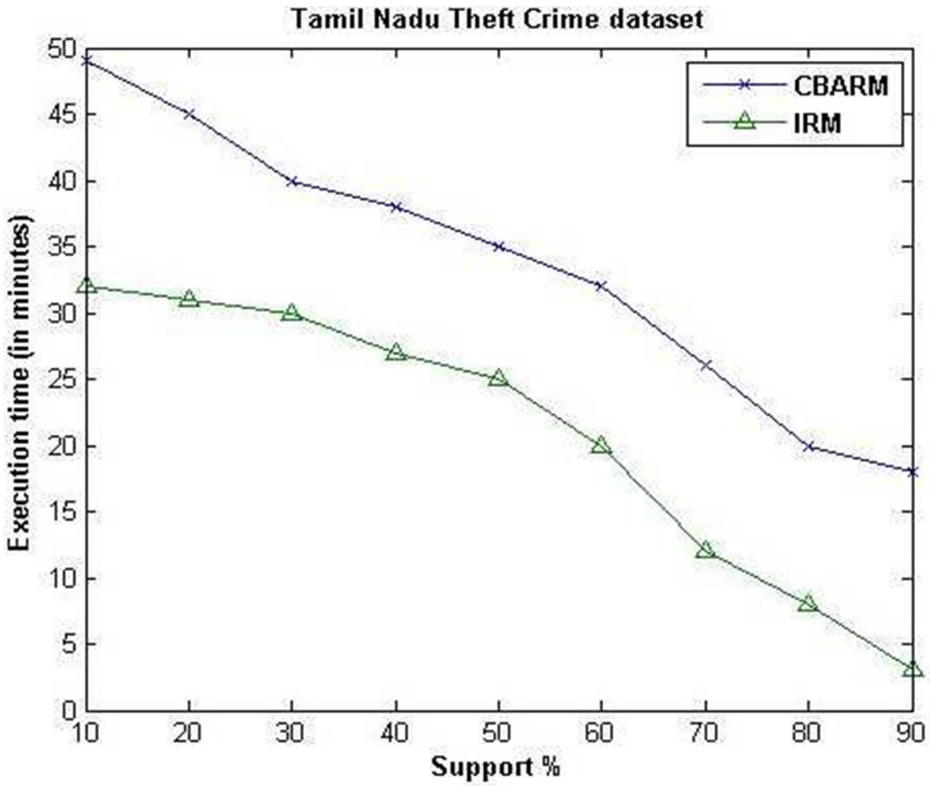

Figure 5. Comparison – Execution time between IRM and CBARM algorithm with varying support in Theft dataset

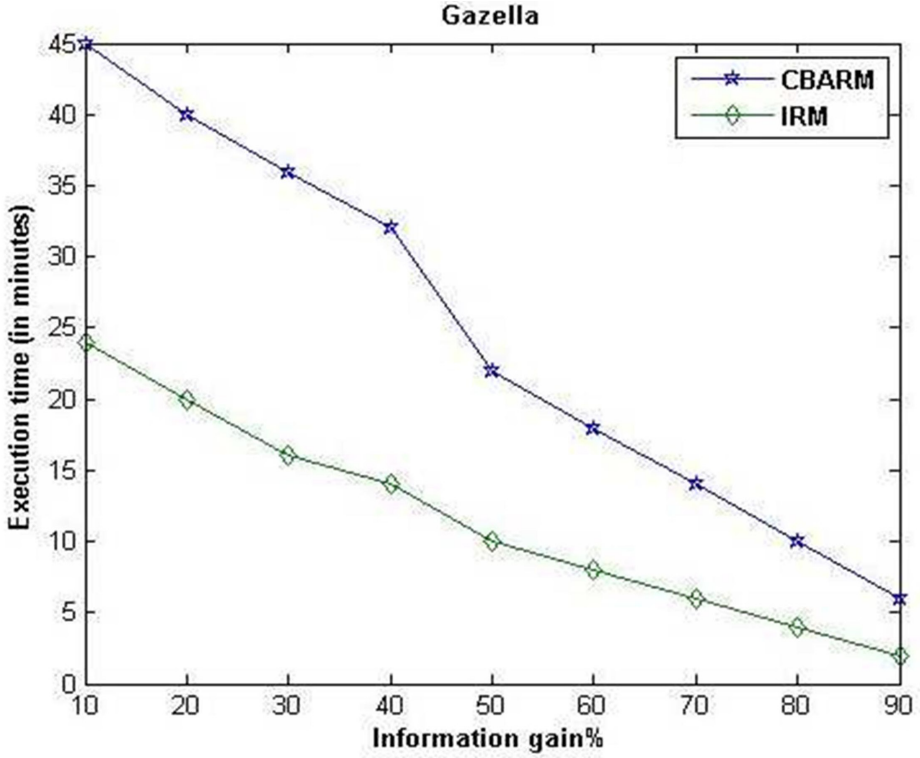

Figure 6. Comparison – Execution time between IRM and CBARM algorithm with varying information gain in Gazelle dataset

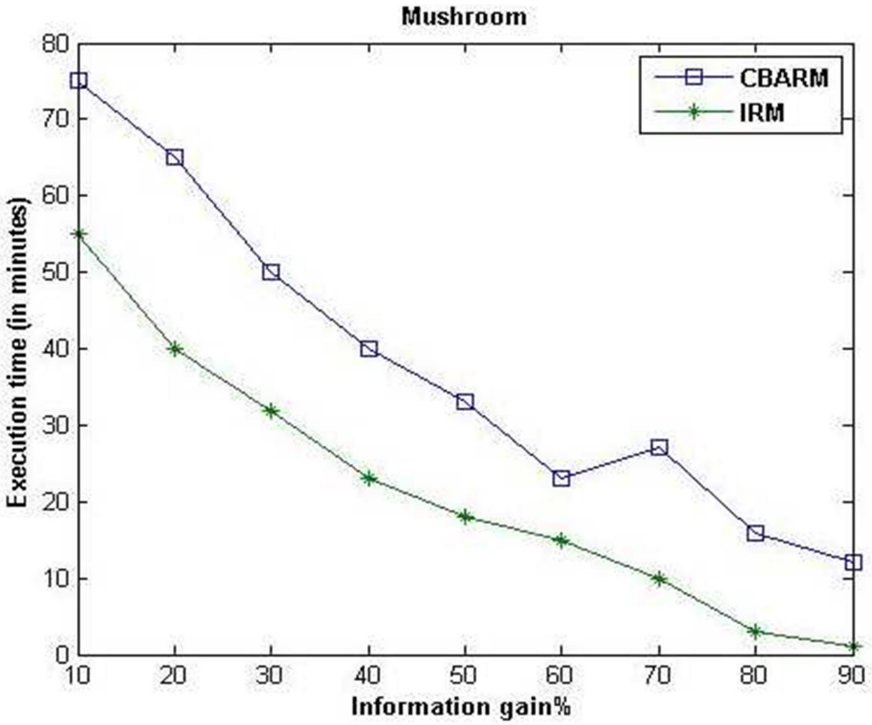

Figure 7. Comparison – Execution time between IRM and CBARM algorithm with varying information gain in Mushroom dataset

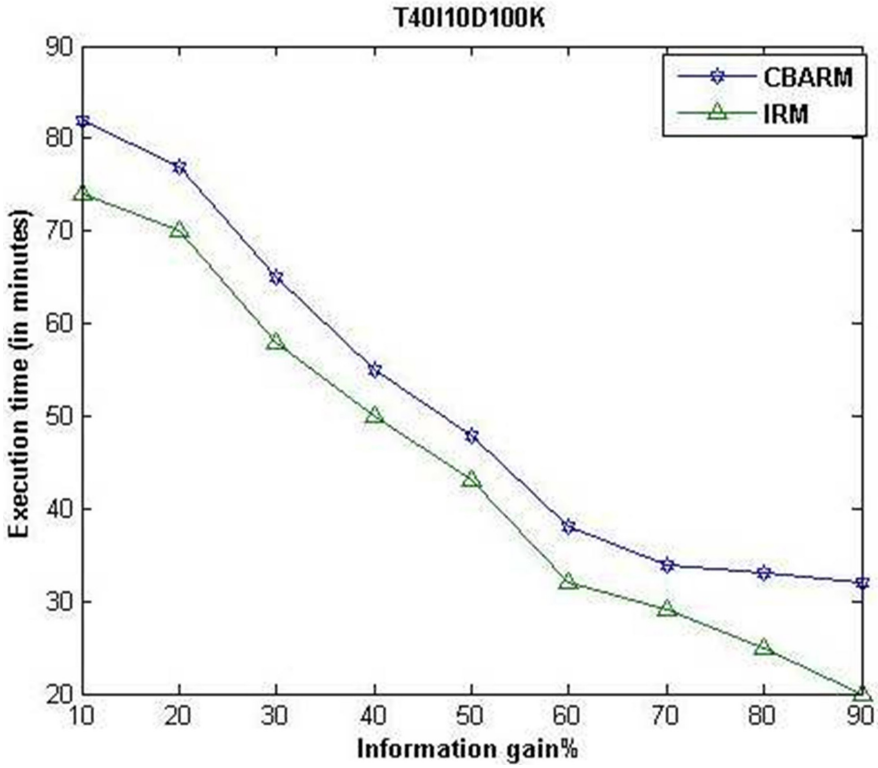

Figure 8. Comparison – Execution time between IRM and CBARM algorithm with varying Information Gain in T40I10D100K dataset

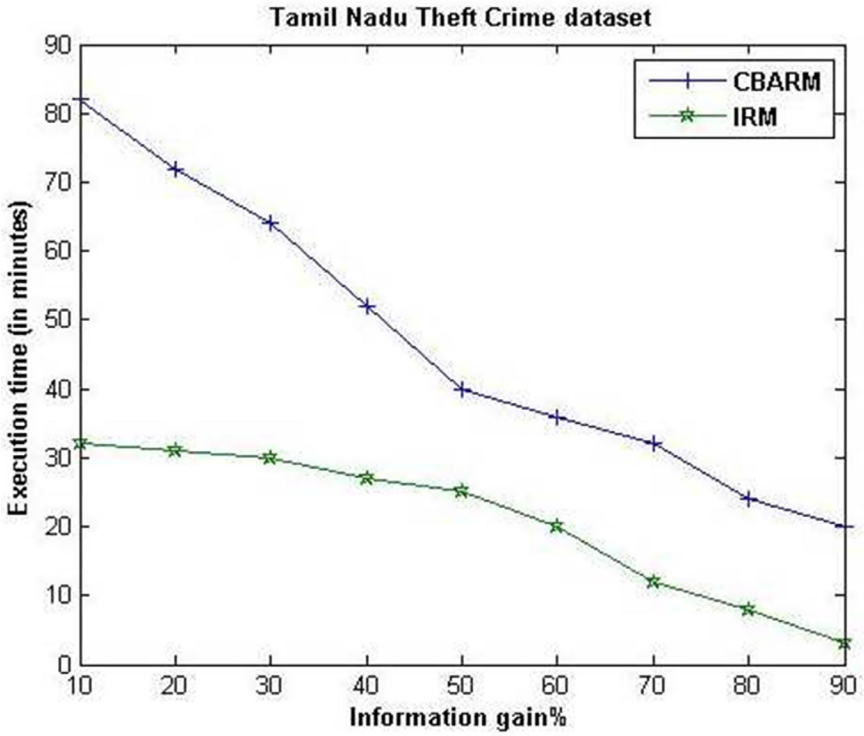

Figure 9. Comparison – Execution time between IRM and CBARM algorithm with varying Information gain in TTC dataset

The figures 6, 7, 8 and 9 shows the comparison between IRM and CBARM algorithm in terms of execution time with varying Information Gain ranges from 20% to 60%. The performance of IRM increases as the information gain increases. In TTC, Gazella and Mushroom dataset, the performance decreased by 8, 11 and 12% respectively when compared with the performance of CBARM. The proposed IRM algorithm performed in ratio varying from 30% to 40% with T40I1D100K dataset when compared with CBARM algorithm. The above mentioned figures shows the comparative study of execution time between IRM algorithm and CBARM algorithm with values using T40I10D100K, Mushroom, Gazella and TTC dataset respectively.

7. Conclusion

This paper detailed about association rule mining algorithm which is used to mine the rule from frequently occurred patterns from large amount of database. The projected frequent itemset acts as a input to generate association rules by using rule construct algorithm and validated the results using interesting and effective measures. The experimental study is conducted through T40I10D100K, Mushroom, Gazella and Crime dataset. The proposed IRM algorithm is compared with above mentioned four datasets. The experimental results prove that IRM algorithm is performed better than existing CBARM algorithm with varies values of measures.

References

[1] Agrawal, R., Srikant, R. (1994). Fast Algorithms for Mining Association Rules, *In*: Proceedings of the 20th International Conference on very large data bases, p. 487-499.

[2] Agrawal, R., Imielinski, T., Swami, A. N. (1993). Mining association rules between sets of items in large databases, *In*: Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, 207-216.

[3] Agrawal, R., Imielinski, T., Mining, Swami. A. (1993). Association Rules between Sets of Items in Large Databases, *In*: Proceedings of the 1993 ACM SIGMOD Conference, 1-10.

[4] Bhandari, Pranay. (2013). Improved Apriori Algorithms – A Survey, *International Journal of Advanced Computational Engineering and Networking*, 1 (2).

[5] Krishnamurthy, Revathy., Kumar, Satheesh. J. (2012). Survey of Data Mining Techniques on Crime Data Analysis, *International Journal of Data Mining Techniques and Applications*, 1 (2) 117- 120.

[6] Gazella dataset, 2000. [online] Available at www.gbif.org/species/5220149/datasets.

[7] Dataset [online] Available at www.fimirepository.com.

[8] Steinbach, Michael., Tan, Pang-Ning., Xiong, Hui., Kumar, Vipin. (2007). Objective Measures for Association Pattern Analysis, *International Journal of Contemporary Mathematics*, 205-226.

[9] Biesiada, Jacek., Duch, Wlodzislaw., Duch, Google. (2005). Feature Selection for High-Dimensional Data: AKolmogorov-Smirnov Correlation-Based Filter. *In*: Proceedings of the International Conference on Computer Recognition Systems, 2005.

[10] Mark A. Hall. (2000). Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. *In*: Proceedings of the Seventeenth International Conference on Machine Learning, 359–366.

[11] Azhagusundari, B., Selvadoss Thanamani, Antony. (2013). Feature Selection based on Information Gain, *International Journal of Innovative Technology and Exploring Engineering (IJITEE)* ISSN: 2278-3075, 2 (2) 18-21.

[12] Rule Based Association Rule Mining algorithm [online] Available at www.docplayer.net

[13] Constraint Based Association Rule Mining algorithm [online] Available at www.lsi.upc.edu

[14] Rule Mining algorithms [online] Available at www.math.upatras.gr

[15] Information Gain[online] Available at www.ijcse.com

[16] Information Gain [online] Available at www.hwsamuel.com

[17] Association Rule [online] Available at www.ijirset.com

[18] Attribute Selection [online] Available at www.research.ijcaonline.org