International Journal of Engineering & Technology, 7 (4.36) (2018) 533-541

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Finding Efficient Positive and Negative Itemsets Using Interestingness Measures

P. Asha[1*], T. Prem Jacob[2], A. Pravin[3]

[1]*Asst.Prof, Department of Computer Science and Engineering, Sathyabama Insttitue of Science and Technology, Chennai.*
[2]*Asst.Prof, Department of Computer Science and Engineering, Sathyabama Insttitue of Science and Technology, Chennai.*
[3]*Asst.Prof, Department of Computer Science and Engineering, Sathyabama Insttitue of Science and Technology, Chennai.*
**Corresponding author E-mail:ashapandian225@gmail.com*

Abstract

Currently, data gathering techniques have increased through which unstructured data creeps in, along with well defined data formats. Mining these data and bringing out useful patterns seems difficult. Various data mining algorithms were put forth for this purpose. The associated patterns generated by the association rule mining algorithms are large in number. Every ARM focuses on positive rule mining and very few literature has focussed on rare_itemsets_mining. The work aims at retrieving the rare itemsets that are of most interest to the user by utilizing various interestingness measures. Both positive and negative itemset mining would be focused in this work.

Keywords: *Association rule, positive association rules, negative association ruls, Interestingness measures.*

1. Introduction

Data Mining has grasped people's interest due to the availability of a wide range of raw data where useful information is poor. Thus, there is a need for extracting meaningful information from the large amount of data that is of user understandable form. Data Mining is an analysis process where patterns are mined from a large database or repositories. To retrieve useful information, data preprocessing techniques such as Data cleaning and integrating, after selection of data followed by transformation using which the mining is done and such data representation is popularly known as Knowledge Discovery in KDD. Mining is performed on relational databases, Data Warehouse, Transaction Databases and Advanced databases such as OODB, ORDB, Multimedia, Spatial and Web Mining Databases. Mining can be applied in various fields such as Sales/Marketing, Healthcare/Insurance, Banking/Finance, Medicine, Biomedicine, Transportation, Telecommunication etc.It is advantageous as a powerful tool that find patterns and relationships among data which helps in discovering hidden information from the large and useless datasets available. But Data Mining cannot work without human effort and also it cannot tell the value of information mined for our need. Thus to ensure meaningful Data Mining results, the researcher must understand the data available.

Association Rule can otherwise be called as pattern. It has two constituents: antecedent and consequent which is similar to if - then respectively. The item P corresponds to the antecedent endowed in the data whereas the consequent Q is endowed in the combination of the antecedent P. Patterns that are mined must be meaningful. Such a meaningful information or pattern is discovered through some interesting measures. Most important measure is supported that indicates how frequently the itemset P or Q occur in the dataset and confidence attests the number of counts, the if/then (P->Q) statements have been committed to be true. Other measures are Laplace, Pearson coefficient, conviction, P-S measure, interest factor, chi-square test, lift, leverage. Some of these measures are discussed and efficient method is comparatively determined in the study. Also the relationship between these meaningful pattern can be identified through a traditional method called Association Rule Mining (ARM). ARM is becoming a research topic that mines associated rules. The strength of the associated rules is determined by interestingness measure. Mined rules have to satisfy some user specified minimum value of support and the confidence. Pretty good algorithms such as Apriori, Elcat, FP-growth are available to generate association rules. In mining, association rules favours us to analyze and predict the customer's behavior. Also, it plays an starring role in the market basket analysis and many other real time applications.

ARM finds relationship between two itemsets of the pattern, P→Q, in which P and Q are disjoint items. Likewise ARM can be used to mine k-itemsets. Most of the existing research is on mining positive_association_rules of the pattern P→Q, but we also focus on mining negative association rules of the pattern, P→~Q, ~P→Q, ~P→~Q. This absence of itemsets is also considered in our study. Different methods of ARM are discussed which mines frequent and infrequent itemsets from where both positive and negative rules are mined. Strong rules can be mined by applying interestingness measure. Major usage of mining negative association rules are fraud detection and to find genetic disorders in the field of Bioinformatics. Primary issues in mining negative rules are identification of appropriate patterns from the large database, thus making it a challenging attempt. Also, various methods that were incorporated to lessen the number of rules generated and number of scans to the database is the main objective discussed in the study.

2. Review on existing work

Rakesh et al. compared Apriori and Apriori_tid algorithms for Association Rule Mining and combined the advantages of both the algorithms, which was termed asApriori_hybrid. Apriori algorithm uses Apriori_generation to generate the set of candidates and is stored in a hash tree. Apriori_tid also uses Apriori_generation along with a unique ID, to index the set of candidates stored in an array. Thus it generates lesser candidates and is much more effective in later passes. The combined Apriori_hybrid algorithm uses Apriori at the initial pass and then switches to Apriori_tid for further passes that involves switching cost which in turn affects the system performance. Gyorgyet al. proposed an algorithm on Survival Association Rule (SAR) to find survival outcomes where Association Rule Mining (ARM) lacks. ARM finds co-occurring frequent patterns, but fails to consider other risk factors like age, because depending on the age factor, the hypertensive and hyperlipidemic factors do vary.Thus SARM extends ARM by handling survival outcomes, making adjustment for confounders as well as incorporating other factors to mine the effective rules. Thulasi et al.has suggested four summarization techniques and evaluated the best and suitable technique among them. ARM is combined with survival analysis and summarization techniques based on greedy set coverage, which is used to summarize the whole database to a smaller set. APRX_collection finds superset of all rules where most subset will be valid. But the dilution of high risk rules and redundancy were its consequences. RPGlobal is also similar to APRX_collection, except that there is a slight difference in redundancy as each individual record is covered. Redundancy-Aware Top-K algorithm reduces redundancy by covering each record and thus it identifies rules with high risk records. Bottom Up Summarization (BUS) was concluded as the best technique, as it effectively controlsredundancy issues than Top-K with quality factor similar to BUS.

Ramakrishnan et al. made a study of different algorithmsthat searches for the subsets of the associated rules. Initially, a review is made to Apriori algorithm, which losses some frequent candidates during prune step. So new methods such as Multiple join and Reorder, produced a list of selected items from the database, and then applied a modifiedcandidate generation algorithm to find the satisfied frequent items. But during prune step, Reorder performed a little better than multiple join. The Direct method uses the subset of database to modify candidate generation and then it counts the items, but is more expensive. It performs better in low minimum support and with larger data sets. Xiaoxin et al. extended the classifying methodology based on Predictive Association Rules (CPAR) as traditional approaches like FOIL and PRM propogates a significant number of association_rules with high processing overhead as well as over fitting. First Order Inductive Learning (FOIL) used greedy algorithm as it searches for the current best rule and it lacks in accuracy due to the generation of very small set of rules. Predictive Rule Mining (PRM) modifies FOIL by multiplying a factor inorder to decrease weight instead of removing it. So it produced more rules, hence increasing efficiency with low accuracy.CPAR improves accuracy and efficiency using dynamic programming, as it prunes out redundant rules. More than one literal can be selected at a time and close-to-the-best literal is chosen based on the gain value, maintaining efficiency and accuracy.

Nancy et al. proposed a new method for positive and negative sequential pattern mining called PNSPM, where ahigher degree of interestingness is achieved even by considering the absence of itemsets. Initially positive and negative patterns are found by Apriori then interestingness measure is calculated by finding the difference between support of preceding and target subsequence and its degree is calculated to mine meaningful patterns.Thus redundancy can be avoided. Yang et al. concentrated on mining the negative information that solves the deficiency of positive association mining.Association rules are generated by Apriori to find candidates and frequent itemsets satisfying the minimum threshold. PNARC algorithm is used to mine positive and negative association rules based on correlation. After Apriori, the correlation coefficient is applied to detect and remove self contradictory rules so as to mine negative association rules. Asha et al. proposed an innovative algorithm called Incremental Positive and Negative Association Rule Mining (IPNAR) on web data. Incremental algorithm updates the database with respect to the transactions ofthe originaldatabase, making it more efficient by reducing the number of passes and the generation of meaningless patterns. IPNAR includes two phases,one to generate itemsets using Apriori and store it in a set.In the second phase, both positive and negative itemsets are mined through a PNARC algorithm and then update the database incrementally with some modifications using a support_conf function, so to reduce the number of passes.It is more efficient and effective as it mines association rules in a single scan.

Diana et al. proposed a new approach, such as,MultiObjective Positive and Negative Association Rule Mining(MOPNAR) with decomposition to measure quality along with maximum interestingness, comprehensibility and performance. Then Yule's Q method is finalized as the best Quantitative Association Rule mining method by comparing it with other measures such as support,confidence,conviction,lift and netconf. Initially Multi Objective Evolutionary Algorithm (MOEA) decomposes MO into subproblems and uses EA to optimize the subproblems. But MOPNAR introduces Extended Population(EP) and restarting processto improve dataset coverage along with nondominant rules storage.Lift provides interestingness and the number of attributes in a rule provides comprehensibility of that rule.Thus MOPNAR is concluded as the better,reduced and strong set of interesting PNQAR's generation method.

Kavitha et al. analysed Item based Bit Pattern (IBP) approach, that reduces the number of scans to the database with less memory access. Interestingness is provided by the correlation coefficient measure.

The existing methods require more scans to the database leading to the scalability issue. It is overcome by using IBP and also it provides better memory utilization along with effective rule generation. At first IBP constructs a table from the existing database, then finds frequent and infrequent itemsets, which is then used for generating positive and negative rules based on the minimum threshold and correlation coefficientthat evaluates the strength of the associated itemsets (Asha et al., 2016).When individual candidates are generated a new IBP overwrites the previous table, thus providing better memory access with increased processing speed.

Ramakrishnudu et al. suggested a tree based approach which reduces the input and output overhead while mining positive (Prem et al) and negative association rules. Moreover it requires only a single scan of the database. Frequent and Infrequent Itemset (FII) tree works as follows, i)DB is scanned once and the root node of the tree is assigned null, and the non root nodes are of the form <item set I, support(I)> stored in a vector. ii)Then insert items one after the other to the tree which is indexed by frequent and infrequent index. iii) Thus k-itemsets are generated that satisfies the support count and correlation coefficient(Azadeh et al) , which is then indexed based on its nature.All the nodes of the tree are analyzed to produce positive (Asha et al) and negative association rules based on its indexing.Thus, this approach appends new nodes to the tree without rebuilding from the scratch.

3. Interestingness measures

Various rule_ interestingness_measures are employed to retrieve the best rules out of all collective rules obtained from the resultant frequent patterns.

1. Primary Measures
   Confidence (cf)
   Support (spt)
   Completeness

Confidence

$$P(L_t - > R_t) = \frac{spt(L_t \cup R_t)}{spt(L_t)}$$

// Lt - items of LHS (left hand side) , Rt- items of RHS (right hand side), spt – support count of item_sets

Support

$$P\frac{(L_t \cup R_t)}{N_{AT}}$$ // NAT - total instances

Completeness

$$P(L_t - R_t) = \frac{spt\ (L_t \cup R_t)}{spt(R_t)}$$

Piatetsky-Shapiro (PS) and Rule Interestingness(RuI)

Gregory Piatetsky-Shapiro suggested laws which have to be satisfied by the RuI measure.

Law_1:

$$spt\ (L_t \cup R_t) = \frac{spt(L_t)*spt(R_t)}{N_{AT}}$$

RuI should be zero, if precedent & subsequent are independent statistically.

Law_2:

Must possess monotonic increase with P (Lt ∪ Rt) // fixed parameters.

Law_3:

Must possess monotonic increase with every P(Lt) & P(Rt). // fixed parameters.

So, RuI should meet all these three criterions.

$$RuI = P(L_t \cup R_t) - \frac{(P(L_t)*P(R_t))}{N_{AT}}$$

Normally, RuI value would be positive.

$$PI = \begin{cases} RuI > 0 & Normal \\ RuI = 0 & Rule\_better\_than\_chance \\ RuI < 0 & Rule\_lesss\_successful\_than\_chance \end{cases}$$

2. Filter the deserving rules

Various other Rule Interestingness measures are also used to rank the rules, in the order of priority and are truly useful.

Lift (Lif)

$$L_{if}(L_t - > R_t) = \frac{P(L_t \cup R_t)}{count(L_t)*count(R_t)}$$

For the rule to be a better one, its lift should be > = 1 (for 10000 transactions).

Leverage (Lge)

Lge (Lt - > Rt) = spt (Lt∪ Rt) - spt(Lt) * spt(Rt)

For a rule to be a better one, its leverage should be > = 0.0001 (for 10000 transactions).

Interest Factor

$$Inf(L_t, R_t) = \frac{spt(L_t \cup R_t)}{spt(L_t)*spt(R_t)}$$

To be a better one, it should possess a positive.

Pearson's Correlation (PC) Coefficient

$$\Phi(L_t - > R_t) =$$

$$\frac{spt(L_t)*spt(R_t)}{\sqrt{\{spt(L_t)*spt(R_t)*(1-spt(L_t))*(1-spt(R_t))\}}}$$

Should range from -1 to +1. Φ(Lt->Rt) =0, if Lt and Rt are independent.

4. Propopsed system

The architecture of the proposed system is explained in Figure 1.The various modules involved in the work are explained below.

1. Data Preprocessing

Multiple input multiple output engine give initial set of values from dataset by finding time series vector format. Initial values are predicated using explanatory values present in the data set. Forecasting values for each item is compared with initial values and then finds the deviations, which is considered as data values in data set.

2. Hash Table Generation

While generating hash table, it consists of infrequent item and infrequent transaction removal. All Item sets which are less than the minimum support threshold are removed. If an item is greater than minimum support threshold and if no combination of that item satisfies the minimum support threshold, that those transactions are removed. Create a table that contains item name and its support for next level reference (Figure 2).

3. Frequent Itemset Mining

In hash table a large number of item set combinations are generated. Final combination of large item sets which are satisfying minimum support threshold are taken as frequent item set.

4. Rule Generation

It consist of two sub operations rule body creation and corresponding rule head generation. Rule body and corresponding rule head were created from frequent item set. Finally combine both the rule body and rule head and they form the set mined rules.

5. Dissimilar rule Mining

From the large number of general rules, fruitful rules are gathered by adding filtering measures.

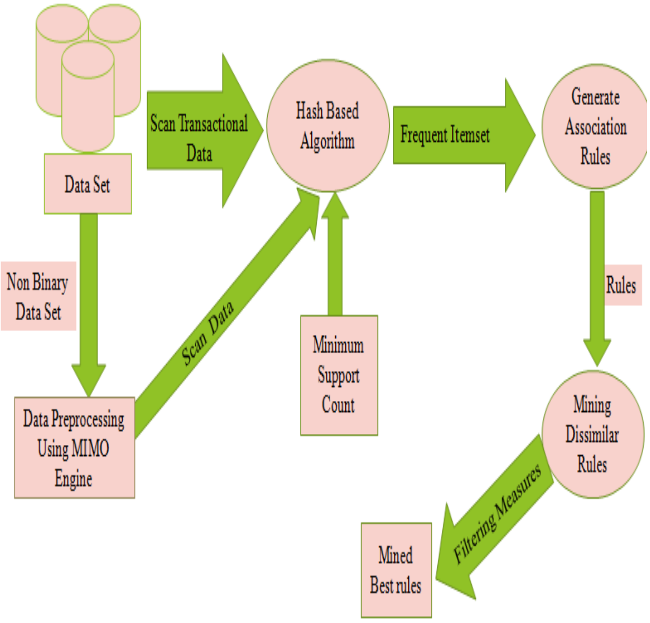


Fig. 1: Architecture of hash algorithm

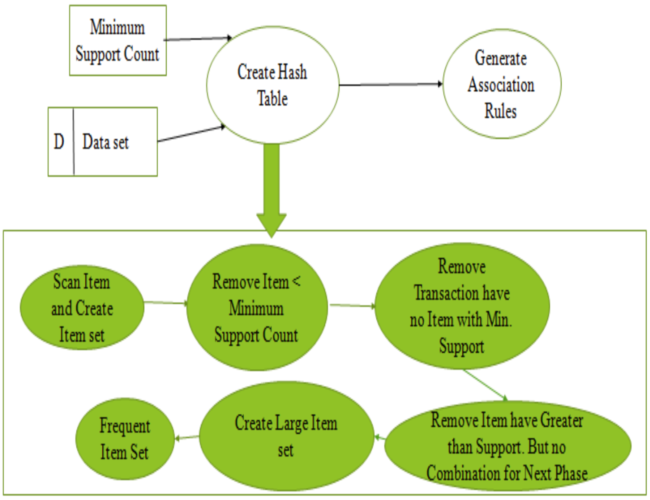
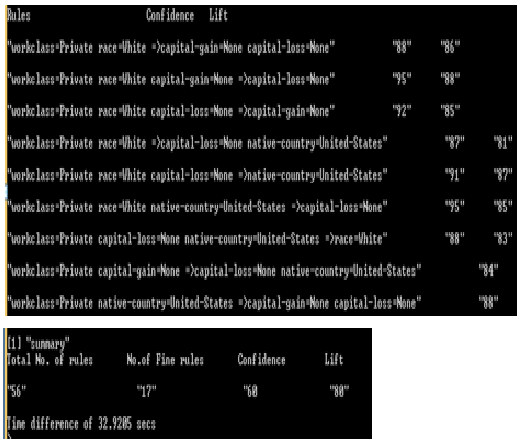

Minimum Support Count
D
Data set
Create Hash Table
Generate Association Rules
Scan Item and Create Item set
Remove Item < Minimum Support Count
Remove Transaction have no Item with Min. Support
Frequent Item Set
Create Large Item set
Remove Item have Greater than Support. But no Combination for Next Phase

Fig. 2: Detailed design

5. Results and discussion

The software used for implementation is R-3.0.1, R -studio. The datasets used are Groceries, SER Prediction, Genome and Adult Datasets. Figure 3 displays the promising rules that were generated by the Hash algorithm for the Adult dataset along with its computation time. Initially the 1- recurrent itemsets are culled, after which the 2-recurrent, 3-recurrents etc are culled from the datasets using the proposed algorithm.



```
Rules                          Confidence   Lift
"workclass=Private race=White =>capital-gain=None capital-loss=None"          "88"     "86"
"workclass=Private race=White capital-gain=None =>capital-loss=None"          "95"     "88"
"workclass=Private race=White capital-loss=None =>capital-gain=None"          "92"     "85"
"workclass=Private race=White =>capital-loss=None native-country=United-States"          "87"     "81"
"workclass=Private race=White capital-loss=None =>native-country=United-States"          "91"     "87"
"workclass=Private race=White native-country=United-States =>capital-loss=None"          "95"     "85"
"workclass=Private capital-loss=None native-country=United-States =>race=White"          "88"     "83"
"workclass=Private capital-gain=None =>capital-loss=None native-country=United-States"          "84"
"workclass=Private native-country=United-States =>capital-gain=None capital-loss=None"          "88"
```

```
[1] "summary"
Total No. of rules      No.of Fine rules      Confidence      Lift

"56"                      "17"                  "68            "80"

Time difference of 32.9205 secs
```

Fig. 3: Rule generation with interestingness measures

Table 1: Comparison of Existing and Proposed System

DATASETS	HASH Vs. APRIORI ALGORITHMS			
	TIME		RULES	
	APRIORI (Minutes)	HASH (Minutes)	APRIORI	HASH
GROCERIES	29	7	53010	27031
SER PREDICTION	19	0.6	8410	2000
GENOME	10	.9	1200	629
ADULT	23	3	800	530

Table 1 displays the number of rules that were generated and the time taken for computation. From the table it is evident that the proposed Hash algorithm consumes less time and ahs produced promising rules. Figure 4 displays the computation time taken by the Apriori and Hash based algorithms for retrieving the recurrent items and promising association rules. Table 2 explains the advantages and disadvantages of prevailing mining algorithms.

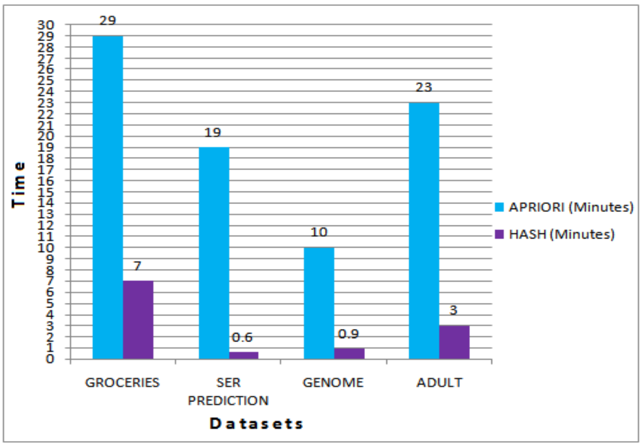


Fig. 4: Computation time of apriori and hash algorithm

Table 2: Comparitive Study of Existing Algorithms

S.No	Method Used	Advantages	Drawbacks
1	Apriori_Hybrid	Scales Linearly With The Number Of Transactions.	Quantities Of Items Are Not Found, Execution Time Depends On The Item Count Present In The Db.
2	Survival Association Rule_Mining (Sarm)	Rules Are More Interpretable, More Suitable To Assess Risk.	Prediction Is Not Appropriate For Individual Patients.
3	Bottom Up Summarization (Bus)	Higher Patient Coverage, Higher Ability To Reconstruct The Information Base.	Slightly Additional Redundancy Than Other Techniques.
4	Integrated Algorithm	Reduce The Execution Time.	More Number Of Redundancies.
5	Classification Based On Predictive Association Rules (Cpar)	Highly Efficient, Accurate And High Quality Classification.	Lacks In Performance.
6	Positive And Negative Sequential Pattern Mining (Pnspm)	High Degree Interesting Rules Are Selected, Redundancywas Reduced To A Greater Extent.	Memory Access Is High.
7	Positive And Negative Association Rule On Correlation (Pnarc)	Much Beneficial To Web Administrator – Amend The Page Structure Easily And Supports Efficient Decision Making.	Negative_Association Mining Is Not Supported.
8	Incremental Positive And Negative Association Rule Mining (Ipnar)	Efficient As Itemsets Mined Are Updated Incrementally,Avoids Multiple Scan By Means Of Improved Search.	Time Consuming.
9	Multiobjective Positive And Negative Association Rule Mining (Mopnar) With Decomposition	Maximum Comprehensibility And Performance, Improves The Diversity Of Rules Obtained, High Coverage, Good Computational Cost And Scalability.	Redundancy Is More.

10	Item Based Bit Pattern (Ibp)	Better Memory Utilization, Reduce No Of Scans To The Db, Scalable, Increase Processing Speed.	Relies On Correlation Coefficient For Rule Interestingness.
11	Fii Tree Approach	Reduced I/O Overhead With Single Scan To Db, Time Efficient, Appends The New Items Directly Into A Tree.	Difficult To Add New Data Or To Delete Old Data Into The Tree While Mining Frequent Patterns.
12	Frequent Pattern Growth Tree (Fp Tree) Approach	Rules Are Categorized For Establishing Them In A Better Way.	Context And Semantics Of Textual Data Is Not Considered, Lacks To Enhance Eminence And Practicality Of Engendered Rules.
13	Approach For Mining Confined Rules	Correlation Coefficient Threshold Value Is Automatically Progressed And Even Be Adjusted, Avoids Redundancy.	Lacks In Accuracy.
14	Improved Fp Tree Called Positive And Negative Association Rules (Pnar)	Achieves Efficiency And Accuracy.	Lacks Better Classification.
15	Confabulation Inspired Association Rule Mining (Carm)	Better Runtime And Memory Access.	Concentrates More On New Rules.

6. Conclusion

The work presented above is all about positive and negative itemset mining. The concepts used in various papers for mining these positive and negative association rules has been discussed along with few interestingness measures, the one that helps for better and meaningful mining. Advantages and drawbacks of each reviewed method are tabulated, which helps to differentiate each method. The proposed work attempts to provide better and effective rule mining with less computation time using various concepts like classification that cluster's data from the database to provide better base for association rule mining, stemming normalizes the whole data for quick and easy mining, after which the ranking is applied to prioritize the mined rules, as it helps the user to get their accurate result. Even the negative associations are given much importance, as they do carry essential and important combinations.

References

[1] Agrawal R & Srikant R, "Fast algorithms for mining association rules", 20th int. conf. very large data bases, VLDB, Vol.1215, (1994), pp.487-499.

[2] Simon GJ, Schrom J, Castro MR, Li PW & Caraballo PJ, "Survival association rule mining towards type 2 diabetes risk assessment", AMIA annual symposium proceedings, American Medical Informatics Association, (2013).

[3] Sinduja K & Saravanan N, "Predicting Relative Risk for Diabetes Mellitus Using Association Rule Summarization Techniques", Imperial Journal of Interdisciplinary Research, Vol.2, No.6, (2016).

[4] Srikant R, Vu Q & Agrawal R, "Mining association rules with item constraints", Kdd, Vol.97, (1997), pp.67-73.

[5] Yin X & Han J, "CPAR: Classification based on predictive association rules", SIAM International Conference on Data Mining, 2003, pp.331-335.

[6] Lin NP, Chen HJ, Hao WH, Chueh HE & Chang CI, "Mining strong positive and negative sequential patterns", WSEAS Transactions on Computers, Vol.7, No.3, (2008), pp.119-124.

[7] Bin Y, Xiangjun D & Fufu S, "Research of web usage mining based on negative association rules", IEEE International Forum on Computer Science-Technology and Applications, Vol.1, (2009), pp.196-199.

[8] Pandian A & Thaveethu J, "SOTARM: Size of transaction-based association rule mining algorithm", Turkish Journal of Electrical Engineering & Computer Sciences, Vol.25, No.1, (2017), pp.278-291.

[9] Martin D, Rosete A, Alcala-Fdez J & Herrera F, "A new multiobjective evolutionary algorithm for mining a reduced set of interesting positive and negative quantitative association rules", IEEE Transactions on Evolutionary Computation, Vol.18, No.1, (2014), pp.54-69.

[10] Asha P & Srinivasan S, "Analyzing the associations between infected genes using data mining techniques", International Journal of Data Mining and Bioinformatics, Inderscience Publishers, Vol.15, No.3, (2016), pp.250–271.

[11] Ramakrishnudu T & Sbramanyam RBV, "Mining positive and negative association rules using fii-tree", Editorial Preface, Vol.4, No.9, (2013).

[12] Prem Jacob T & Ravi T, "An Optimal Technique for Reducing the Effort of Regression Test", Indian Journal of Science and Technology, Vol.6, No.8, (2013), pp.5065-5069.

[13] Soltani A & Akbarzadeh TMR, "Confabulation-inspired association rule mining for rare and frequent itemsets", IEEE Transactions on neural networks and learning systems, Vol.25, No.11, (2014), pp.2053-2064.

[14] Asha P & Srinivasan S, "Distributed association rule mining with load balancing in grid environment", Journal of Computational and Theoretical Nanoscience, Vol.13, No.1, (2016), pp.33-42.