



A NEGATIVE ASSOCIATION RULES FOR WEB USAGE MINING USING NEGATIVE SELECTION ALGORITHM

¹R. GOBINATH, ²M. HEMALATHA

¹Research Scholar, Department of Computer Science, Karpagam University, Karpagam University, Coimbatore, Tamilnadu, India

²Assoc. Prof., Department of Computer Science, Karpagam University, Coimbatore, Tamilnadu, India
E-mail: iamgobinathmca@gmail.com, csresearchhema@gmail.com

ABSTRACT

The immense capacity of web usage data which survives on web servers contains potentially precious information about the performance of website visitors. Pattern Mining involves applying data mining methods to large web data repositories to extract usage patterns. Due to the emerging reputation of the World Wide Web, many websites classically experience thousands of visitors every day. Examination of who browsed what can give necessary approach into the buying pattern of obtainable customers. Right and timely decisions made based on this acquaintance have helped organizations in accomplishing new heights in the market. The aim of discovering rules in Web log data is to obtain information about the navigational behavior of the users. This can be used for marketing purposes, dynamic creation of user profiles, finding user pattern navigation, etc. This suggested study, particularly focuses on examining an efficient algorithm for mining negative association rules from the clustered weblog navigational patterns. The algorithm indicates traditional association rules to include negative association rules to be an alternative for using positive association rule itself. When mining negative association rules take place, the same minimum support vestibule is used to mine frequent negative item sets. The rules gathered in such a manner are bridged by applying a negative selection algorithm, which helps in discovering all convincing negative association rules quickly and solves some of the limitations in traditional association rule mining. The experimental result reveal the proposed work to be highly effective.

Keywords: *Data Mining, Web mining, Web Usage Mining, Association Rule Mining, Negative Selection Algorithm.*

1. INTRODUCTION

Web usage plays an important role for making out the behavior of website visitors and their browsing patterns. This knowledge enhances the effectiveness of website personalization and for web marketing. The extraction of knowledge from World Wide Web has become an essential and complicated task owing to the enormous amount of data which is semi structural in nature. The scrutiny of web log files is used for applications like customer shopping sequence, web clicks, biological sequence, and creation of dynamic users profile. In this paper, discovery of relevant rules from the identified frequent patterns is considered from web log files. Traditionally the web log data are in raw format. It needs to be cleaned, condensed and transformed in order to scrutinize the apparent and useful information. After performing cleaning process, the relevant features which are useful for pattern recognition as to be performed for grouping similar patterns. Finally pattern mining can be

performed for extracting rules by means of association in navigation behavior[1-3].

Mining of association rules is an important area of research in web usage mining, correspondingly the hotspot which uses large amounts of data in the web server log and other relevant data sets for mining, analysis and gains valuable knowledge model about usage of relevant web sites[4].

Many application domains use association rules for their usefulness in decision support, diagnosing capacity and much more. Finding such rules is a very important component for improving the version of the association rule mining algorithm.

Scrutinizing of association rules is a matter of identifying association rules which appear spasmodically in a given transaction data set [5]. Every traditional association rule mining algorithm was ameliorated for positive associations between items and also deals with problem of generating all association rules that consist of frequent item set

and confidence greater than the user specified minimum confidence.

In order to manage web space efficiency the association rule discovery helps in personalizing set of pages with a minimum support value [6].

Eg: confidence is 80% of the user who accessed "News/onair/business/sports.asp" also accessed "living/sport/travel/frontpage.asp", but only 20% of those accessed "News/onair/business" accessed "living/sport/travel/frontpage.asp", then it shows that some information in "sports.asp" is making the clients to access "frontpage.asp".

With the mounting exploiting and growth of data mining patterns in recent times numerous studies are paying attention on discovering navigational patterns which can offer precious information. However mining negative association rule is a difficult task due to the reality that there are essential differences between positive and negative association rule mining. In this proposed work we focus on three major key issues in negative association rule mining [7].

1. What methodology should be used to effectively improvise search for a negative frequent item set

2. Determining efficiencies to identify negative association rule

3. The technique to be implemented for optimized negative association rules are selected to improve classification accuracy.

2. RELATED WORK

The largest number of rules created by association rule algorithms are partially used for information analyst. Research proposal in [8] for pruning the group of rules generated may have irrelevant rules during the rule generation stage. The mining process for removing irrelevant rules from the set of rules generated by [9] is a formal approach. The mining techniques to snip off the rule set size remains abandoned and many researchers focus on reducing rule cluster size. The most common dispute in mining association rules is hand picking the perfect motivating measures as mentioned in [10]. The rule clusters over abandoned number of motivating measures is a analysing study of [11] for mining relevant data and conversation about association rule mining limitations are prescribed in [12]. Over a generation of rules, false findings and rule discovery, which are hard to understand and minimize constant support for some of the disputes which are solved

with a proposed tactic of filtering top k- association discoveries. Association rule mining comes under the web usage mining environment and web usage deals with the website link structure data and closely correlated items. The web pages which are closely linked may fall under same session and has very high confidence and makes user apathy. The method introduced by [13], introduces additional association rules for web usage data during the classification of association rule mining. The Generalized association rule is a basic technique used in web usage mining for analysing patterns. The hierarchical conceptual model is used to demonstrate the relationship between attributes in different levels. The Generalized association rule permits rules in different levels of attributes [14]. The Generalized association rules are based on the URL query and cannot be used for the regular updating websites.

3. MATERIALS AND METHOD

To find all co-occurrence relationships called associations, Aggarwal [15] introduced the concept of finding the main objective of association rules and it has gained a great deal of attention. An association rule is an emersion $x \rightarrow y$ where x and y are set of items. In detail given a Data base D of transaction where each transaction $T \in D$ is a set of items. $X \rightarrow Y$ express that whenever transaction T contains X then T probably contains Y also. The probability is rule confidence which is defined as the percentages of transactions containing Y in addition to X with regard to occurrence of number of transaction containing X . In case of web log mining

Eg: /index/content \rightarrow /dunia

This example represents the pages index and content are accessed along with "dunia" \neg page. It shows association that the users who are navigating index and content will also probably access "dunia" webpage. The truth of the rule can be identified by finding the support and confidence of the above mentioned rules with the whole transaction data base. If the value gained is greater than minimum support and minimum confidence, then the rule is valid.

3.1 Counting Occurrence Algorithm

Before support and confidence for each rule is determined, the number of occurrences for each rule must be calculated. An algorithm for counting the number of occurrence is shown below [16]:

Pseudo code: Counting Occurrences Algorithm
Input: Rules generated from web log data
Procedure: Read record in database For each record in database If Filter ItemLevel1 \cap ItemLevel2 \diamond 0 then Counter = Counter + 1 End if Next record
Output: number of occurrence of each rule

Procedure: Confidence = $\frac{\text{Total occurrence for item X and Y}}{\text{Total occurrence for item X}}$
Output: Computing confidence value of each rule

3.2 Algorithm for support

Support measures how often the rules occur in database. To determine the support for each rule produced, several arguments have been identified in calculating the support such as Total Transaction in database and number of occurrences for each rules. The formula for support is shown below [16].

Pseudo code: Support calculation formula
Input: Total Transaction in DB No. of occurrences each item {x,y}
Procedure: Support = $\frac{\text{Number of occurrences } \{x, y\}}{\text{Total Transaction in DB}}$
Output: Computing support value of each rule

3.3 Algorithm for Confidence

The Formula for calculating the confidence value is shown below [16].

Pseudo code: Confidence calculation formula
Input: Total occurrence for item X Total occurrence for item X and Y

4. PROPOSED FRAMEWORK

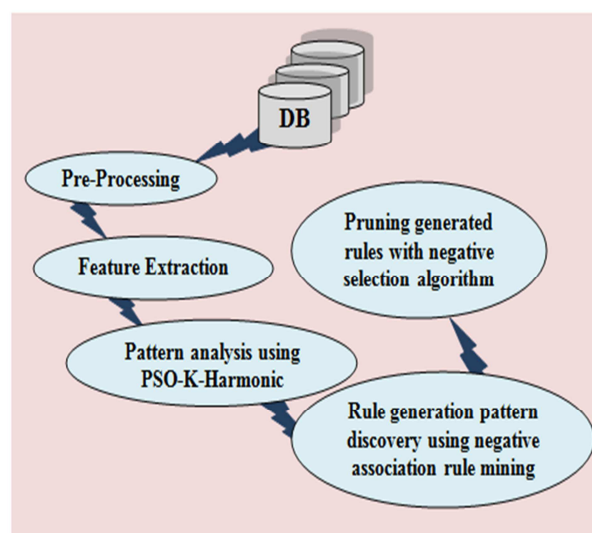


Figure 1. Architecture for Rule Mining.

Step 1: Collect the raw data set from web log file

Step 2: In pre- processing stage performs cleaning process by eliminating multimedia files, duplicates etc. And as a data extraction process user and session are determined

Step 3: Future extraction is done on the web log file based on the user navigation behavior.

Step 4: In order to identify similar patterns PSO-K-Harmonic mean was implied to cluster the similar user navigation patterns.

Step 5: From the clustered data set association among each item set was identified applying negative association rule mining.

Step 6: The resultant rules are justified by negative selection algorithm which prunes the resultant rules for better optimization.

The system is proposed to discover interesting patterns in weblogs navigational sequence. Information about accesses to various Web pages within the Web space associated with a particular server is in the Weblog.

Association rules capture relationships among page views based on navigation patterns of users in case of Web transactions [7], [17].

Steps involved in the proposed system

Proposed system involves the following steps:

- 1) The input is a set of Weblogs for which Positive and Negative association rules have to be found. We have chosen PSO-K Harmonic based clustered navigational patterns of three different data sets as an input in this paper.
- 2) To obtain Negative association rules with minimum support and confidence, both positive and Negative association rules are applied on those navigational sequence clusters.
- 3) Interesting patterns of infrequent path traversal can be discovered and client's web usage can be evaluated as a result of negative association rule mining.

4.1 Negative Association Rule Mining

The elements other than the item set A in a cluster can be denoted with $\neg A$. The positive rule associated with positive rule can be considered as $A \Rightarrow B$. Likewise negative association rules of the various forms $A \Rightarrow \neg B$, $\neg A \Rightarrow B$ and $\neg A \Rightarrow \neg B$ are considered [18],[19],[20],[21].

The minimum support of negative association rules can gather information from positive association rules [22]. The same method is followed in acquiring minimum confidence of negative association rules. Support of negative association rules is given by the following formulae.

$$\sup p(\neg A) = 1 - \sup p(A) \quad (3)$$

$$\sup p(A \Rightarrow \neg B) = \sup p(A) - \sup p(A \cup B) \quad (4)$$

$$\sup p(\neg A \Rightarrow B) = \sup p(B) - \sup p(A \cup B) \quad (5)$$

$$\sup p(\neg A \Rightarrow \neg B) = 1 - \sup p(A) - \sup p(B) + \sup p(A \cup B) \quad (6)$$

The traditional way of many association framework technique follows "support-confidence". However these two parameters allow in pruning of many interesting rules from the discovered data. This proposed framework is not only pruning interesting rules from discovered data, but it also added a measurement known as conviction for the process. The formulae for confidence is given below.

$$\text{conf}(A \Rightarrow \neg B) = \frac{\sup p(A) - \sup p(A \cup B)}{\sup p(A)}$$

$$\text{conf}(\neg A \Rightarrow B) = \frac{\sup p(A) - \sup p(A \cup B)}{1 - \sup p(A)}$$

$$\begin{aligned} \text{conf}(\neg A \Rightarrow \neg B) \\ = \frac{1 - \sup p(A) - \sup p(A) + \sup p(A \cup B)}{1 - \sup p(A)} \end{aligned}$$

The three equations have its capacity of discovering necessary rules with their support and confidence greater than, or equal to, by considering minimum support and minimum confidence thresholds respectively, which is specified by the user. The rules collected in such a manner can be considered as interesting negative association rules.

The process of mining negative association rules can be deployed for mining positive rules. The steps involved in extraction of interesting positive rule list and negative rule list are given below[7].

- (1) A positive list of frequent itemset is generated and negative list of infrequent itemset is generated.
- (2) Extract positive rules in the form $A \Rightarrow B$ in positive list.
- (3) Extract negative rules in the forms $A \Rightarrow \neg B$ and $\neg A \Rightarrow B$ in the negative list.

The correlation coefficient measurement for extraction negative association rules and pruning strategies are designed. The extraction of negative association rules can be possible with the rules extracted using associated technologies. The usability of the frequent negative item sets uses same threshold as in association rule mining. The strategy of Yule's correlation measurement is deployed in a quick pruning process [23]. The negative itemset mining is explained in following steps.

Algorithm: Negative Association Rules

Input: TDB-Transactional Database

MS-Minimum Support

MC-Minimum Confidence

Output: Negative Association Rules

Method:

1. $N \leftarrow \Phi$
2. Find $F1 \leftarrow$ Set of frequent 1- itemsets
3. for ($k=2; Fk-1 \neq \Phi; k++$)
4. {
5. $Ck = Fk-1$ join $Fk-1$
6. // Prune using Apriori Property
7. for each $i \in Ck$, any subset of i is not in $Fk-1$ then $Ck = Ck - \{i\}$
8. for each $i \in Ck$ find Support(i)
9. for each A, B ($A \cup B = i$)

```

10. {
11. QA,B= Association(A,B);
15. if Q<0
16. {
17. if(supp(A←¬ B)≥MS&&conf(A←¬ B)≥MC)
then
N← N U { A← ¬ B}
19.
if(supp(¬ A→B)≥MS&&conf(¬ A→B)≥MC)
then
20. N←N U { ¬ A← B}
21. }
23. }
24. AR← P U N
End
End
Return (Repertoire)
Input: InputSamples, Repertoire
For ( InputSamples)
"non-self"
For ( Repertoire)
If (Matches(, ))
"self"
Break
End
End
End

```

4.2 Optimization of Rule generation using Negative Selection Algorithm

The acquired immune system of mammalian is a major inspiration for developing the negative selection algorithm. The self-discrimination character plays a key role in the negative selection algorithm. The immunity acquired by clonal selection theory, which means adaptive behaviour of the immune system including ongoing selection and proliferation of cells that select for potentially harmful (and typically foreign) material in the body [24]. An interesting aspect of this process is that it is responsible for managing a population of immune cells that do not select-for the tissues of the body, specifically it does not create self-reactive immune cells known as auto immunity. This problem is known as 'self-non self-discrimination' and it involves the preparation and ongoing maintenance of a repertoire of immune cells such that none is auto immune. This is achieved by a negative selection process that selects for and removes those cells that are self-reactive during cell creation and cell proliferation. This process has been observed in the preparation of T-lymphocytes, naive versions of which are matured using both a positive and negative selection process in the thymus.

```

Input: SelfData
Output: Repertoire
Repertoire
While (StopCondition())
Detectors GenerateRandomDetectors()
For ( Repertoire)
If (Matches(, SelfData))
Repertoire
End

```

- The Negative Selection Algorithm was intended to change detection, novelty detection, intrusion detection and similar pattern recognition and two-class classification problem domains.
- Traditional negative selection algorithms used binary representations and binary matching rules such as Hamming distance, and -contiguous bits.
- A data representation should be selected that is most appropriate for a known problem domain, and a matching rule is in turn chosen or tailored to the data representation [24].
- Detectors can be equipped with no prior acquaintance of the problem domain other than the known (normal or self) dataset.
- The algorithm can be configured to balance between detector convergence and the space complexity .
- The lack of dependence between detectors means that detector preparation and application is inherently parallel and suited for a distributed and parallel implementation, respectively.
- In our proposed method the rules generated using negative association rule mining are pruned using the negative selection algorithm [24].

This negative selection algorithm optimizes the performance of the negative association rule for improving the accuracy of classification.

5. EXPERIMENTAL RESULTS

For conducting experimental result, this proposed work has used the three different log server files for feature extraction Kdlog, Online shopping server log file and msnbc.com datasets.



The below table shows the sample rule generation based on positive association rule with the transaction support and confidence

Table 1. POSITIVE ASSOCIATION RULE

Positive Association rule	Support	Confidence
Opinion and misc and travel -> on air	90.26	92.37
Living and sports and business and bbs ->front page	90.00	91.72
News and tech and living and business and sports ->front page	89.00	90.81
Front page and tech and opinion and living and news	87.60	88.59
Front page and tech and living and business ->sports	87.87	88.01
News and living and business and sports ->front page	87.18	88.14
Misc and living and travel ->on air	86.55	87.12
News and tech and misc and bbs ->front page	85.99	84.52
Misc and business and travel ->on air	85.65	85.39
Misc and business and bbs ->front page	85.57	85.41
Tech and living and sports and bbs ->news	85.49	86.09
Local and misc and business and sports ->frontpage	85.32	87.35
News and on air and business ->sports	85.01	85.69
Front page and opinion and living and sports ->news	87.81	87.94
News and misc and business and sports ->front page	86.22	86.84

Table 1 shows the support and confidence obtained using positive association rules. In this the “onair” page was mostly associated when accessing “Opinion\misc\travel” with 90.26% support and 92.37% confidence. It was leastly associated with “misc\business\travel” with 85.65% as support and 85.39% as confidence.

Table 2 shows the sample rule generation based on Negative association rule with the transaction support and confidence.

Negative Association rule	Support	Confidence
not Opinion and misc and Not business -> on air	89.21	91.53
Living and not sports and not travel ->front page	91.27	91.45
not News and not tech and living and not business and sports -> front page	93.56	93.94
Not Front page and not tech and not opinion and living -> news	94.28	94.63
Front page and not tech and not living and business ->sports	91.20	94.51
Not News and not living and not business and sports ->front page	89.49	89.53
Not Misc and not living and not travel ->on air	88.53	88.09

News and tech and misc and not bbs ->front page	87.28	87.05
Misc and not business and travel ->on air	86.14	86.13
Not Misc and business and not bbs ->front page	85.94	85.37
Not Tech and living and not sports and bbs ->news	86.31	86.25
Not Local and misc and not business and not sports ->frontpage	85.58	85.63
News and not onair and business ->sports	85.32	85.29
Front page and not opinion and not living and not sports ->news	85.36	85.41
News and not misc and business and not sports ->front page	85.39	85.04

The table above shows the support and confidence obtained using Negative association rules. In this the “news” page was mostly associated when accessing the pages which are not navigated through front page, tech, opinion but through living, which was represented as “Not Front page and not tech and not opinion and living” with 94.28% support and 94.63% confidence. It was leastly associated while accessing the pages through front page but not with opinion, living, sports. The negative association rule such as “Front page and not opinion and not living and not sports” contains 85.36% as support and 85.41% as confidence.

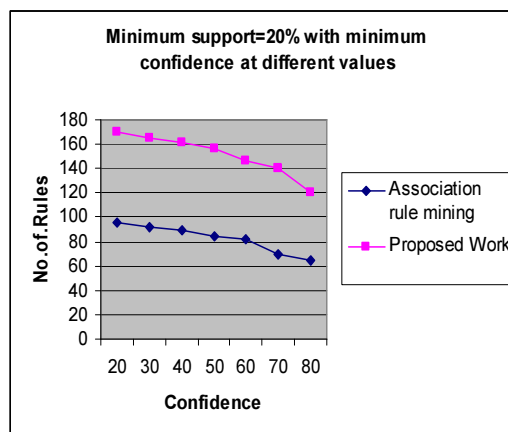


Fig Minimum Support =20% And Different Minimum Confidences

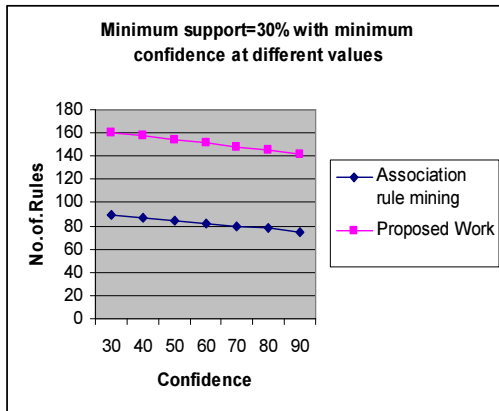


Fig Minimum Support =30% And Different Minimum Confidences

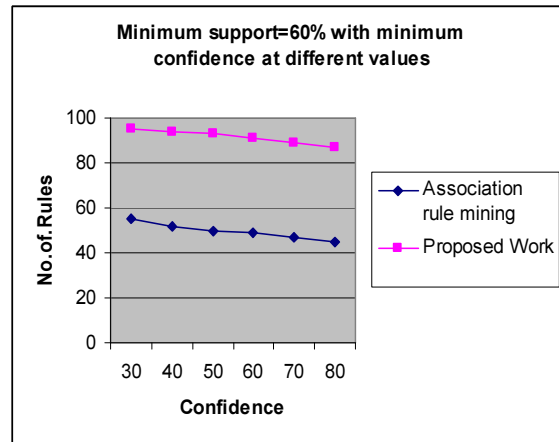


Fig Minimum Support =60% And Different Minimum Confidences

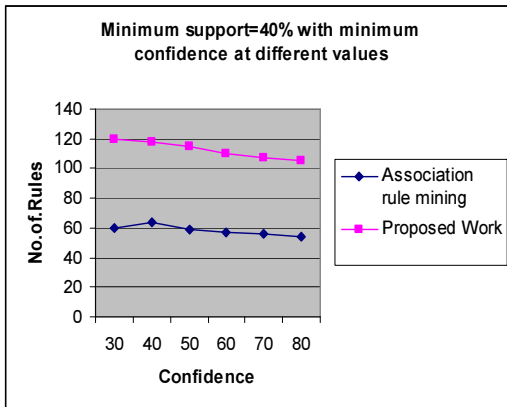


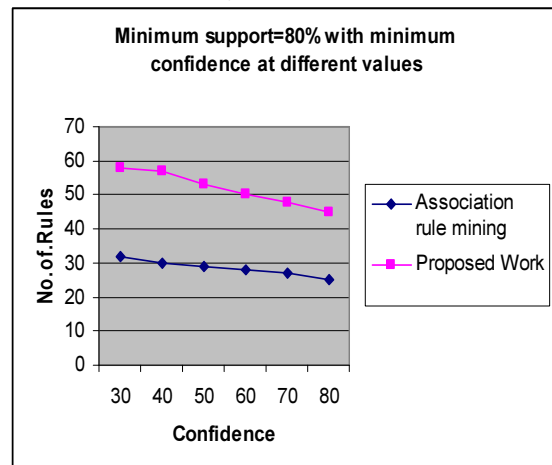
Fig Minimum Support =40% And Different Minimum Confidences



Fig Minimum Support =70% And Different Minimum Confidences

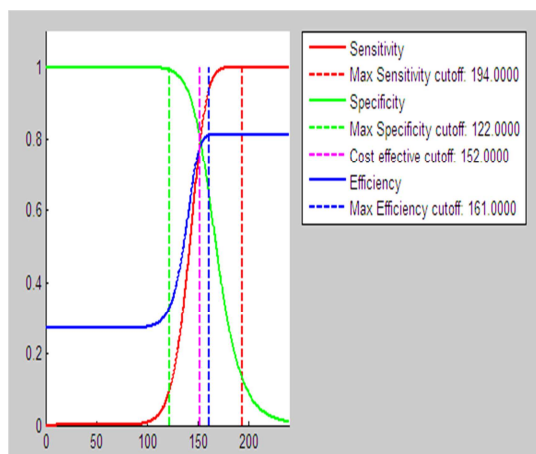
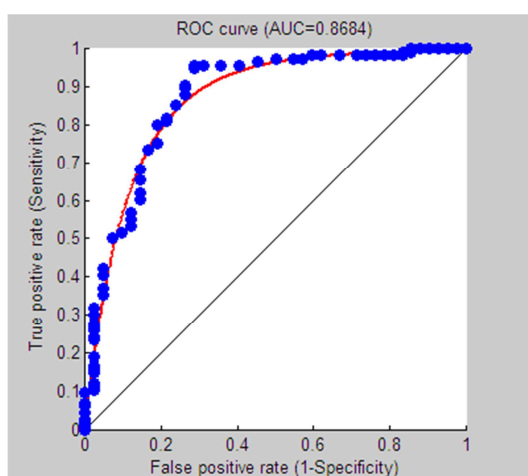


Fig Minimum Support =50% And Different Minimum Confidences



The Figure Above Shows, The Performance Of Association Rule Mining With The Proposed Approach Using Different Values Of Support And Confidence. When The Value Of Confidence Started Increasing The Number Of Rules Get Decreased To Sustain The Minimum Support And Confidence.

The proposed algorithm of Negative association rule mining was tested after pruning the generated rules with negative selection algorithm and the comparison was performed with existing traditional association rule mining which considers only the positive rules. A web log file transactions for three different datasets is considered this algorithm was tested with different minimum supports and minimum confidences. The suggested algorithm is performing better than the existing one. It focuses on the important fact that when the support and confidence level increases its minimum values the number of rules dropped considerably.



The result of ROC curve with performance of proposed work is displayed with the AUC 0.8684. The maximum specificity and efficiency is shown. The Max Sensitivity Cut-off point= 194.00, Max Specificity Cut-off point= 122.00, Cost effective Cut-off point= 152.00 and Max Efficiency Cut-off point= 161.00

6. CONCLUSION

In this study, a new architecture has been designed for efficiently mining negative association rules in clustered weblog dataset. To cluster the similar navigation pattern PSO-K-Harmonic mean from an earlier work was adopted. This approach is novel and unique from existing traditional association rule mining research work on weblog dataset. This algorithm is intended for pruning strategies for reducing the search space and improving the usability of mining rules, and have used the Negative selection algorithm which is an inspiration of artificial immune system to judge which selects the promising negative association rule should be used for better optimization of weblog data classification. The result shows that instead of considering the frequent item set in transaction, the infrequent item set should also be taken into consideration for better managerial decision support in designing the websites based on the category of user's behavior.

REFERENCES:

- [1] Goethals, B and Zaki, M, FIMI'03: Workshop on Frequent Itemset Mining Implementations. Volume 90 of CEUR Workshop Proceedings series. <http://CEUR-WS.org/Vol-90/>. 2003
- [2] Teng, W, Hsieh, M and Chen, M, 2002. On the mining of substitution rules for statistically dependent items. In: Proc. of ICDM. pp.442–449.
- [3] Tan, P and Kumar, V, 2000. Interestingness measures for association patterns: A perspective. In: Proc. of Workshop on Postprocessing in Machine Learning and Data Mining.
- [4] Gourab Kundu, Md. Monirul Islam, Sirajum Munir, Md. Faizul Bari ACN: An Associative Classifier with Negative Rules 11th IEEE International Conference on Computational Science and Engineering, 2008.
- [5] Mohd Helmy, A., W, Mohd, N., H., M and Mohamad Farhan, M., M, 2010. Discovering Web Server Logs Patterns Using Generalized Association Rules Algorithm, New Advanced Technologies, PP: 177-193.
- [6] Kiruthika, M, Rahul, J et., al., 2011, Pattern Discovery using Association Rules, International Journal of Advanced Computer Science and Application, 2(12): 69-74.
- [7] Honglei., Z and Zhigang., X, 2008. An Effective Algorithm for Mining Positive and Negative Association Rules, Computer Science



- and Software Engineering, 2008 International Conference on , 4(1): pp.455,458
- [8] Sahaaya, A. M. and Malarvizhi, M, 2010. Improving web navigation technique using weighted order representation. International Journal of Research and Reviews in Computer Science, 1(2): 55-60.
- [9] Balcázar, J., L., Redundancy, reduction schemes, and minimum-size bases for association rules, 2010, logical methods. Computer Science, 6(2:3): 1-33.
- [10] Huang, X. 2007. Comparison of interestingness measures for web usage mining: An empirical study, International Journal of Information Technology and Decision Making (IJITDM), 6(1): 15-41.
- [11] Webb, G., I. 2011. Filtered-top- k association discovery. WIREs Data Mining Knowledge Discovery 2011, 1: 183-192.
- [12] Dimitrijevic. M and Bosnjak. Z. 2011. Association rule mining system. Interdisciplinary Journal of Information, Knowledge, and Management, 6(1): 137-150.
- [13] Kosala, R and Blockeel, H. 2000. Web mining research: A survey, SIGKDD Explorations, 2(1): 1-15.
- [14] Tan, P., Kumar, V and Srivastava, J, 2004, Selecting the right interestingness measure for association patterns. Information Systems, 29(4): 293 – 313.
- [15] Agrawal., R, Imielinski., T and Swami.,A, 1993. Mining association rules between sets of items in massive databases, In Proc. of the 1993 ACM SIGMOD International Conference on Management of Data, ACM, Washington D.C., pp. 207-216.
- [16] Aditi Shrivastava and Nitin Shukla, 2012, Implementation of Navigation Patterns Mining in Dot Net Framework, International Journal of Computer Science and Information Technologies, 3(5): 5272-5275.
- [17] Ruchi Bhargava and Shrikant Lade, 2013, “Effective Positive Negative Association Rule Mining Using Improved Frequent Pattern, 3(2): 1256-1262.
- [18] Wu, X, Zhang, C and Zhang, S, 2004. Efficient mining both positive and negative association rules. ACM Transactions on Information Systems, 22(3), pp. 381-405.
- [19] Wu, X, Zhang, C and Zhang, S, 2002. Mining both positive and negative association rules. In: Proc. of ICML. pp. 658-66.
- [20] Yuan, X, Buckles, B, Yuan, Z and Zhang, J, 2002. Mining Negative Association Rules. In: Proc. of ISCC. pp. 623-629.
- [21] Honglei, Z and Zhigang, X, 2008. An Effective Algorithm for Mining Positive and Negative Association Rules. International Conference on Computer Science and Software Engineering.
- [22] Antonie, M., L and Zarane., O, R, 2004. Mining Positive and Negative Association Rules: an Approach for Confined Rules, Proc. Intl. Conf. on Principles and Practice of Knowledge Discovery in Databases, pp 27-38.
- [23] Honglei., Z, Zhigang., X, 2009. A Novel Incremental Updating Algorithm for Maintaining Discovered Negative Association Rules, icrccs, International Conference on Research Challenges in Computer Science, pp.164-167.
- [24] Jason Brownlee, 2012, Clever Algorithms: Nature-Inspired Programming Recipes, PP. 277-283.