COMPARATIVE ANALYSISOF ASSOCIATION RULE GENERATION ALGORITHMSIN DATA STREAMS

Dr. S. Vijayarani[1]and Ms. R. Prasannalakshmi[2]

,

[1]Department of Computer Science,School of Computer Science & Engineering, BharathiarUniversity,Coimbatore, Tamilnadu, India.
[2]Department of Computer Science, School of Computer Science & Engineering, Bharathiar University, Coimbatore, Tamilnadu, India.

ABSTRACT

Data mining technology is engaged in establishing helpful and unfamiliar data from the huge databases. Generally, data mining methods are useful for static databases for knowledge extraction wherever currently available data mining techniques are not appropriate and it also has a number of limitations for managing dynamic databases. A data stream manages dynamic data sets and it has become one of the essential research domains in data mining. The fundamental definition of the data stream is an arrival of continuous and unlimited data which may not be stored fully because it needs more storage capacity. In order to perform data analysis with this, many new data mining techniques are to be required. Data analysis is carried out by using clustering, classification, frequent item set mining and association rule generation. Association rule mining is one of the significant research problems in the data stream which helps to find out the relationship between the data items in the transactional databases. This research work concentrated on how the traditional algorithms are used for generating association rules in data streams. The algorithms used in this work are Assoc Outliers, Frequent Item sets and Supervised Association Rule. A number of rules generated by an algorithm and execution time are considered as the performance factors. Experimental results give that Frequent Item set algorithm efficiency is better than Assoc Outliers and Supervised Association Rule Algorithms. This implementation work is executed in the Tanagra data mining tool.

KEY WORDS

Data Stream, Association Rules, Assoc Outliers, Frequent Item sets and Supervised Association Rule, Tanagra.

1.INTRODUCTION

A data stream is an unbroken arrival of data which is boundless in nature. The foremost individuality of the data stream is it handles primary size of unremitting data and most perhaps infinite [1] [8]. The application locale of data streams is market-basket data analysis, cross-marketing, catalogue manner, loss-leader analysis, industry organizations (process credit card transactions), economic markets (stock alternates), engineering and industrial development (power supply and manufacturing), security (traffic engineering observing) and web (web logs and webpage click streams). Essential data mining tasks performed in data streams are clustering, classification, association rule generation, query optimization and frequent item set mining [2].

Association rules are described by finding the frequent pattern, links, relationship and the related structures among the data objects in the databases and in order repositories. There are two important steps in association rule mining; initial one is to find the frequent data items and the next step is to generate association rules via these frequent data items [4] [7]. The association rule mining problem is defined as, assume a given set of items $I= \{I_1,I_2,\ldots I_m\}$ and a database of transactions $D=\{t_1,t_2,\ldots t_n\}$ where $t_i=\{I_{i1},I_{i2},\ldots.I_{ik}\}$ and $I_{ij}\epsilon$ I, an association rule is an inference of the form $X \Rightarrow Y$ where $X,Y \subset I$ are sets of items called item sets and $X \cap Y=\theta$ [5].

Two important events support and confidence are used for association rule generation. The support of an item (or set of items) is the % of transactions in which that item (or items) happens. The support (s) for an association rule $X \Rightarrow Y$ is the percentage of transactions in the database that contain $X \cup Y$. The confidence or strength (α) for an association rule $X \Rightarrow Y$ is the ratio of the number of transactions that contain $X \cup Y$ to the number of transactions that include X. Usually, confidence measures the strength of the rule, while the support measures how frequently it should occur in the database [6]. Some of the important association rule mining algorithms are, a priori, fp-tree, fp-growth, dynamic item set counting, ECLAT, DCLAT and RARM.

This research work mainly focuses on generating association rules from data streams. The nonstop arrival of data is divided into many partitions as windows and it is stored in the form databases. For each and every partition, association rule generation algorithms are applied to generate the association rules. In this work, the traditional association rule algorithms specifically Assoc Outliers, Frequent Items and Supervised Association Rule are used for generating association rules in each partition. From this, we come to know that the advantages, drawbacks and limitations of these conventional association rule mining algorithms for generating association rules in data streams [8].

The remaining portion of this paper is prepared as follows. Proposed methodology and the traditional association rule algorithms are explained in Section 2. Section 3 talks about experimental results and conclusion is given in Section 4 [16].

2. PROPOSED METHODOLOGY

The system architecture of the proposed work is represented in Figure 1.

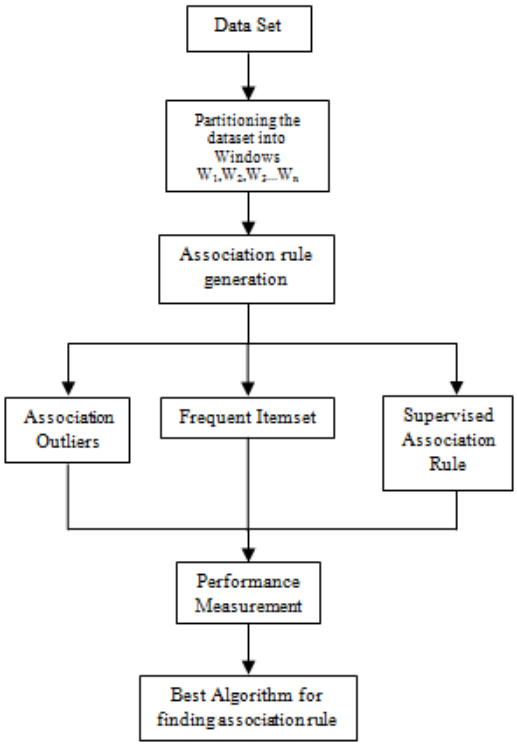

Figure 1. System Architecture

2.1 Dataset

The connect data set is used in this work. It is extorted from http://fimi.ua.ac.be/data/connect.dat. It consists of 67,558 instances and 48 attributes. In this work, 1K, 2K and 5K instances are used. In data streams, we imagine that the nonstop arrival of data is partitioned into five windows with a fixed size, i.e. $W_1, W_2, W_3 \ldots\ldots W_n$. [17].

Association Rule Generation

In order to generate association rules, three types of algorithms are used

- Assoc outlines' (Association Outliers).
- Frequent Item Set Mining.
- Supervised Association Rule.

2.1.1 Association Outliers

An association outlier algorithm is used to build rules from an attribute value dataset.Important terms used in this algorithm are,

- $A_1, A_2, \ldots, A_m$ are attributes.
- $D_1, D_2, \ldots, D_m$isdata items.
- Let $z^{(i)}$to be aithoccurrence of z . A is the value on the get attribute of the eventi. $z^{(i)}$ can be represented as, $z^{(i)} = (z_1^{(i)}, z_2^{(i)}, \ldots, z_m^{(i)})$, where $z_k^{(i)} = z^{(i)}$. $A_k \in D_k$,$k \in \{1, \ldots, m\}$. Z is the set of all events.

Table 1.Pseudo Code for Association Outliers

Step 1- Get input of the record set is contained database DB and a rule set is belong to R

Step 2-

1. Initializes I is 0 (NULL) value
2. For each transaction t belongs to DB. i.e., t ∈ DB
3. Candidate Generation for Association outliers with the transaction is C_t^0;
4. For (i=0; R'=R; i++)
5. Until the candidate generation is growing
6. Temp is NULL;
7. For each transaction t is equal to X -> Y belongs to R'
8. If $X == C_1^i$then
9. Append Y to temp and delete t from rule generation R';
10. The sum of the candidate generation is $C_t^{i+1} = C_t^i$ union by temp, i++;
11. Transaction t = mod of $C_t^i - C_t^0$ divide by mode of C_t++;
12. Return NULL;

2.1.2 Frequent Item Set Algorithm

A description of frequent item set mining algorithms are instinctive, a set of items that emerge in many containers is assumed to be "frequent". Frequent items to be formal; there are a number of us, entitled the support threshold. If, I is a place of items, the support for I is the amount of containers for which I is a subset. Applications of frequent item sets are used in supermarkets, and unique purpose of thisis used for analysis of true market baskets. That is, superstores and chain stores, record the contents of each market basket (physical shopping cart) brought to the list for checkout. At this time the "items" are the unlike products that the store sells, and the "containers" are the sets of items in market-basket. A most important chain might sell 100,000 different items and accumulate data about millions of market baskets. Through finding frequent item sets, a merchant can find out the items which are frequently purchased. [12][16].

Table 2. Pseudo Code for Frequent Item Sets

Pseudo code

Step 1- Ck: Candidate itemset of size k L_k: frequent itemset of size k

Step 2-

1. L1 = {frequent items};
2. for (k = 1; L_k !=Æ; k++) do
3. C_{k+1} = candidates generated from L_k;
4. For each transaction to inthe database do
5. Increment the count of all candidates in C_{k+1} that are contained in it;
6. Endfor;
7. L_{k+1} = candidates in $C_{k+1}$with min_support
8. Endfor; return $\cup_k L_k$;

2.1.3 SPV Assoc Rule (Supervised Association Rule)

This algorithm was originally developed tothe relational variables with constant position. The predictive association rules explore the associations between the items that differentiate a dependent attribute. Thisalgorithmisused in supervised learning framework.The algorithm is not truly customized. Looking at the association rules is just limited to item sets that consist of the dependent variable. The computation time is reduced after that, there are two components of Tanagra are devoted to this mission: SPV Assoc Rule and SPV Assoc Rule Tree. To compare the predictable approaches, the machinery of Tanagra has an additional specificity, it can denote the class value "dependent variable = value" that desire to forecast. This is decisive for occurrence when the preceding probability of the dependent changeable values is very dissimilar. However, it was in the perspective of multivariate characterization of collections of individuals. These individualsare compared to the group characterization component. [18].

Table 3. Pseudo Code for Supervised Association Rules

Supervised Association _ Rule _APRIORI

Step 1- Input candidate item set 1 and 2

Step 2-

1. If supervised item sets k=2
2. For each frequent item set f∈ F_1 do
3. Candidate generation item sets are inserted to frequent item set fins C_1
4. End for
5. $C_{1\_class\_label,}$ C_{1_other} is equal to the split of C_1 is groups of class label into the $C_{1\_class\_label}$ and the other frequent item sets into C_{1_other}, CL.
6. For each candidate item sets C_1∈ C_{1_Label} do
7. Generate the item set of class_label items and non_class_label items.
8. For each candidate itemset c2 ∈C_{1_other} do {
9. Now Σ(c) = form of c_1 and $c_{2.}$
10. Class_Label candidate item sets c is inserted into $C_{2.}$
11. } }
12. For each candidate item set c_1 ∈C_{1_label} do {
13. Identify all the class labels in the array of C_{1_label} that is after c1
14. For each candidate item sets c_2 ∈C_{post} do {
15. Now Σ (c) = form of c_1 and $c_{2.}$
16. Insert the c into the C_2
17. }}Else
18. For each i_1 is count C_i {
19. For each i_2 is frequent item set F_{k-1} {
20. If (the same item sets are included by k-2 items of i_1, i_1) ^ (different fromthe last item set are i_1, i_2) {
21. Candidate generation C= the form of first k-1 items of i_1 and last items of i_2
22. Insert c into the C_k
23. } } } }
24. Return C_k

Table 4. Rule Generation for Association Outliers

Window Size	Threshold	1000 Ds	2000 Ds	5000 Ds	10,000 Ds
		Rules			
W1	σ = 25, C = 55	231	231	328	359
W2		199	57	187	359
W3		234	125	156	421
W4		231	251	312	499
W5		241	297	297	484

3.EXPERIMENTAL RESULTS

The connect data set is used in this work. It is extorted from http://fimi.ua.ac.be/data/connect.dat. It consists of 67,558 instances and 48 attributes. In this work, 1K, 2K and 5K instances are used.The continuous arrival of data is partitioned into five windows with a fixed size, i.e. W_1, W_2, W_3, W_4, W_5[16].A number of rules generated and execution time is considered as the performance factors.

Table 5. Execution Time for Association Outliers

Window Size	Threshold	1000 Ds	2000 Ds	5000 Ds	10,000 Ds
		Rules			
W1	σ = 25, C = 55	2220	2240	2290	2375
W2		2234	2250	2210	2241
W3		2315	2311	2342	2386
W4		2936	2913	2918	2954
W5		2940	2932	2948	2979

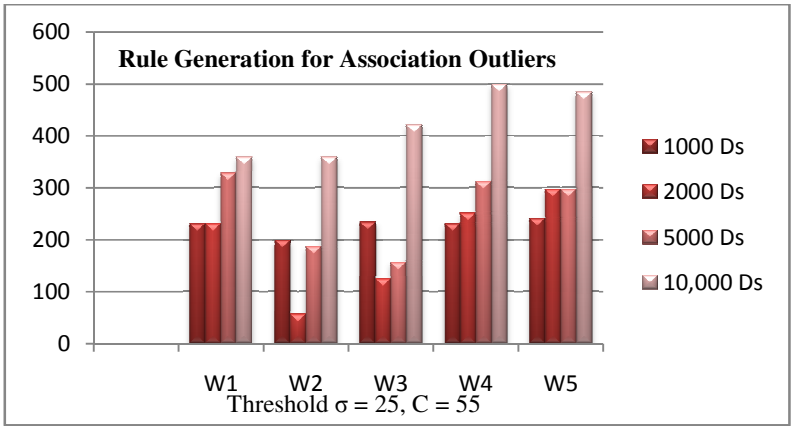

Figure 2.Association Outliers for Rule Generation.

Figure 2 gives the information about the number of association rules generated by the association outlier algorithm for 1K, 2K, 5K and 10K of datasets with two different thresholds like support and confidence values. i.e for five windows.

Table 6.Rule Generation for Frequent Item Set

Window Size	Threshold	1000 Ds	2000 Ds	5000 Ds	10,000 Ds
		Time (ms)			
W1	σ = 25, C = 55	190	234	278	240
W2		256	125	121	199
W3		44	74	184	202
W4		220	256	120	240
W5		303	375	109	183

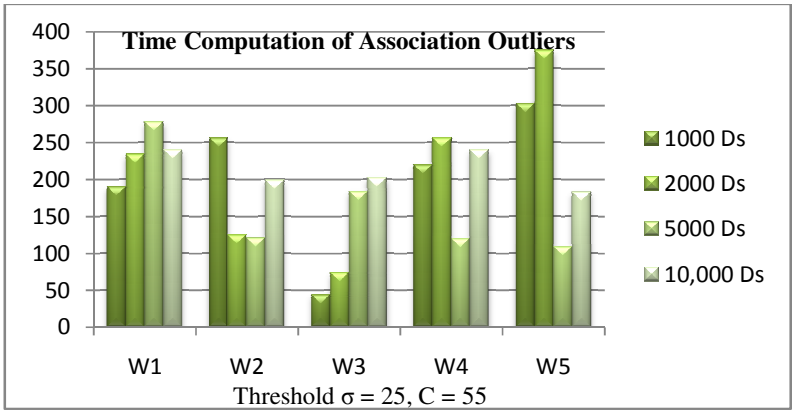

Figure 3.Execution time for Association Outliers.

Figure 3 gives the information about the time computation by the association outlier algorithm for 1K, 2K, 5K and 10K of datasets with two different thresholds like support and confidence values. i.e for five windows.

Table 7 Time Computation for Frequent Item Set

Window Size	Threshold	1000 Ds	2000 Ds	5000 Ds	10,000 Ds
		Time (s)			
W1	σ = 25, C = 55	0.03	0.02	0.03	0.04
W2		0.01	0.1	0	0.02
W3		0.01	0.01	0.09	0.02
W4		0.02	0	0.01	0.03
W5		0.04	0.01	0.01	0.09

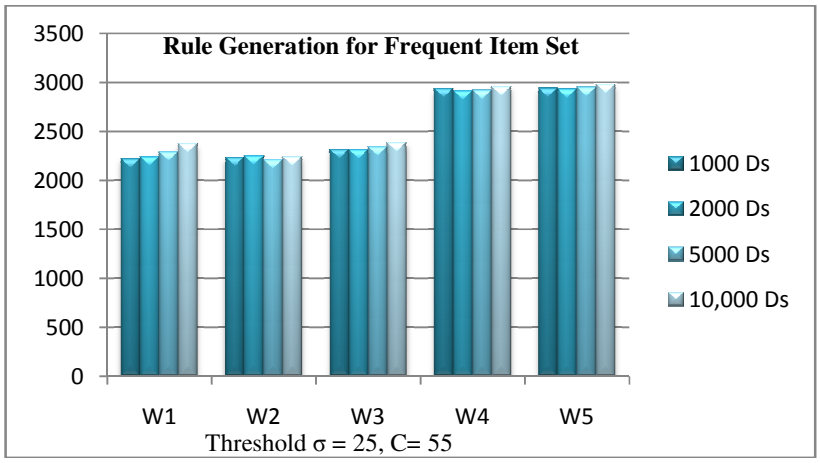

Figure 4 provides the information about the number of association rules generated by the frequent item set algorithm for 1K, 2K, 5K and 10K of datasets with two different thresholds like support and confidence values. i.e. for five windows.

Table 8. Rule Generation for SPV Association Rule

Window Size	Threshold	1000 Ds	2000 Ds	5000 Ds	10,000 Ds
		Rules			
W1	σ = 25, C = 55	216	218	212	220
W2		122	135	131	120
W3		145	156	167	144
W4		202	199	256	193
W5		181	201	210	198

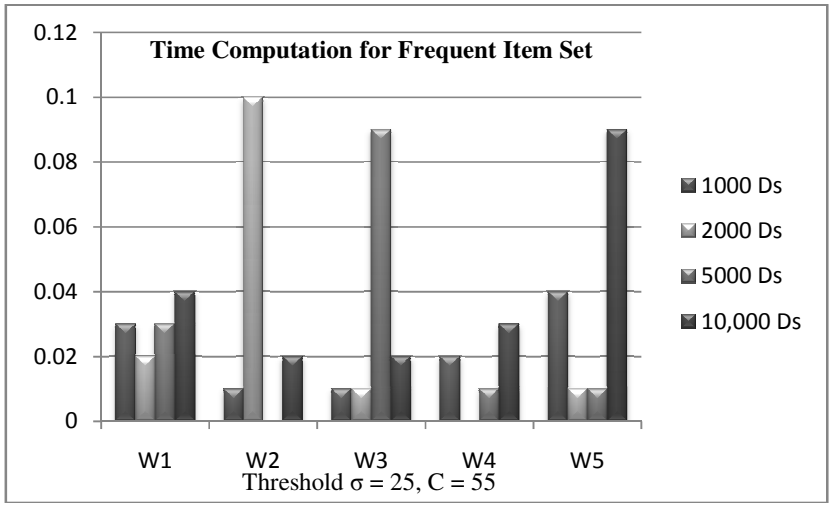

Figure 5.Time Computation for Frequent Item Set.

Figure 5. Provides the information about the time computation by the frequent item set algorithm for 1K, 2K, 5K and 10K of datasets with two different thresholds like support and confidence

values. i.e. for five windows. An experimental result of the frequent item set algorithm is better than the Association outlier algorithm and SPV Association rule algorithm.

Table 9. Time Computation for SPV Association Rule

Window Size	Threshold	1000 Ds	2000 Ds	5000 Ds	10,000 Ds
		Time (ms)			
W1	σ = 25, C = 55	62	64	51	53
W2		31	33	31	30
W3		46	47	46	61
W4		46	23	46	44
W5		33	31	33	64

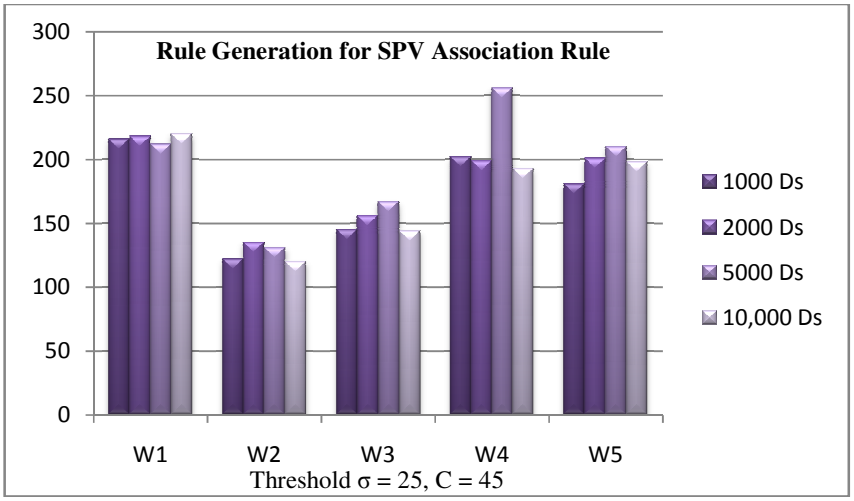

Figure 6.Rule Generations for SPV Association Rule.

Figure 6 shows the association rules generated by the Supervised Association Rule algorithm for 1K, 2K, 5K and 10K of datasets with two different thresholds like support and confidence values. i.e. for five windows.

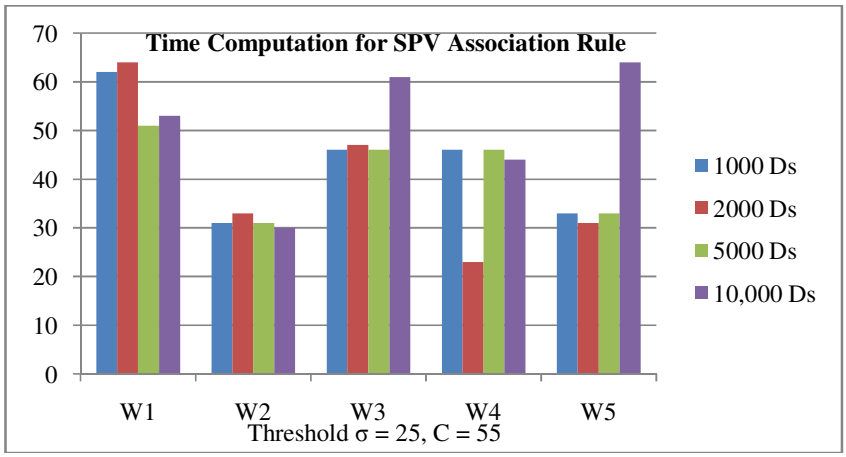

Figure 7.Time Computation for SPV Association Rule.

Figure 7offers the information about the time computation by the Supervised Association Rule algorithm for 1K, 2K, 5K and 10K of datasets with two different thresholds like support and confidence values. i.e. for five windows.

4. CONCLUSION

This main objective of this work is to compare the traditional association rule mining algorithms for generating association rules in data streams. From the experimental results, it is observed that the performance of frequent item set mining algorithm is good and it has produced better results than association outliers and SPV association rule mining algorithms. These algorithms scanned the database more than once and hence it needs more execution time. In future new algorithms are to be developed in order to reduce the number of scans and execution time.

REFERENCES

[1] Aggarwal C (2003). A Framework for Diagnosing Changes in Evolving Data Streams.ACM SIGMOD Conference.
[2] Agrawal, R. and Srikant, R. Fast Algorithms for Mining Association rules. Proc. 20th VLDB conference, Santiago, Chile, 1994.
[3] A. Savasere, E. Omiecinski, and S.B. Navathe, “An efficient algorithm for mining association rules in large databases,” Intl. Conf. on Very Large Databases, pp. 432–444, 1995.
[4] Charu C. Aggarwal “Data Stream Models and algorithms”-Data streaming book 2009, Springer.
[5] Christian Hidber. Online Association rule mining. SIGMOD ’99 Philadelphia PA. ACM 1-58113-084-8/99/05, 1999.
[6] CharanjeetKaur, Association Rule Mining using Apriori Algorithm: A Survey ISSN: 2278 – 1323 International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 6, June 2013.
[7] “Data mining techniques “by Arun k Pujari.
[8] “Data Streams: An Overview and Scientific Applications” Charu C. Aggarwal.
[9] “Data Mining: Introductory and Advanced Topics” Margaret H. Dunham.
[10] Frequent item set mining data set repository, http:// fimi.cshelsinki.fi/data/
[11] Han, J., Kamber, M.: “Data Mining Concepts and Techniques”, Morgan Kaufmann Publishers, 2006.
[12] Kamini Nalavade, B.B. Meshram, “Finding Frequent Item sets using AprioriAlgorihm to Detect Intrusions in Large Dataset”, International Journal of Computer Applications & Information Technology Vol. 6, Issue I June July 2014 (ISSN: 2278-7720). Page | 84

[13] "Mining frequent patterns across multiple data streams" Jing Guo, Peng Zhang, Jianlong Tan and li Guo, 2011.
[14] Nan Jiang and Le Gruenwald, "Research Issues in Data Stream Association Rule Mining"- SIGMOD Record, Vol. 35, No. 1, Mar. 2006.
[15] RakeshAgrawal, RamakrishnanSrikant; Fast Algorithms for Mining Association Rules; Int'l Conf. on Very Large Databases; September 1994.
[16] S.Vijayarani et al, " Mining Frequent Item Sets over Data Streams using Éclat Algorithm" , International Conference on Research Trends in Computer Technologies (ICRTCT - 2013) Proceedings published in International Journal of Computer Applications® (IJCA) (0975 – 8887) 27.
[17] Website: Tanagra.software.informer.com.
[18] Website:http://data-mining-tutorials.blogspot.com/2008/11/supervised-association-rules.html.

AUTHORS

Dr. S. Vijayarani, MCA, M.Phil, Ph.D is working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy and security issues in data mining and data streams. She has published papers in the international journals and presented research papers in international and national conferences.

Ms. R.Prasannalakshmihas completed M.C.A in Computer Applications. She is currently pursuing her M.Phil in Computer Science in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of interest are Data Streams and privacy preserving in Data mining.