Volume 4, Issue 11, November 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Mining Multilevel Association Rules for Data Streams with Time Horizon by Artificial Bee Colony Based Optimization for Fuzzy Logic

Mrs. K. Geetha
M.Sc., M.Phil Scholar,
Department of Computer Science,
Sri Ramalinga Sowdambigai College,
Vadavalli, Coimbatore, Tamilnadu, India

Mrs. B. Hemapriya
M.C.A.,M.Phil., Assistant Professor,
Department of Computer Application,
Sri Ramalinga Sowdambigai College,
Vadavalli, Coimbatore Tamilnadu, India

Abstract— Data mining becomes one of the most important frequently used methods in earlier days to mine frequent itemset for data stream applications. Extraction of multilevel fuzzy association rules for frequent itemset in the downstream data becomes a more difficult task, in order to overcome this problem in this presents a novel multilevel fuzzy association frequent itemset mining for downstream applications. In the proposed multilevel fuzzy association frequent itemset mining methods the frequent itemsets are mined inside time horizon and non-frequent itemsets are mined. In the proposed multilevel fuzzy association frequent itemset mining the fuzzy membership function values are tuned by proposing artificial bee colony optimization algorithm for the web anonymous transaction dataset and frequent itemsets are represented different hierarchical level. The efficiency of the proposed multilevel fuzzy association frequent itemset mining is measured based on the execution time and memory consumption, it becomes efficient in terms of together qualitative and quantitative terms. Experimental results also performed between three mining algorithms apriori algorithm, apriori window test algorithm, apriori window test shift algorithm, in terms of the execution time and memory use ,it shows that the proposed multilevel fuzzy association frequent itemset mining solution performs faster than the other methods.

Keywords—Data mining, Data stream mining, Mining methods and algorithms, Frequent itemsets, Artificial bee colony algorithm, fuzzy association rules.

I. INTRODUCTION

Investigation of the patterns and sequence, frequent items in the datastream application have possess new challenging and more interesting in several areas. As mined frequent itemsets can make known attractive and formerly unidentified facts [1–4] .Due to the development of the information technology (IT) quickly, the managing of larger transaction database and finding the importance of the each item based on the user perspective becomes more difficult and becomes more important. Several numbers of works have been designed to mine frequent itemset information and information from large databases. Some of the data stream related application are network monitoring data, telecommunication connection data, readings from sensor nets and stock quotes possess more examination area in now a day's [5-6] .

The objective of data stream mining, mines a best frequent itemset, pattern based on the creation of the learning methods, most of the existing data stream mining methods mainly focus on centralized approaches. It works well for frequent itemset mining and simultaneously increasing computational complexity and more power expenditure, and privacy disquiet.

Some of the applications which is under the categories of the data streams are as follows [7-8], Network monitoring in data stream, Intrusion detection in data stream, Sensor network analysis in data stream, Cosmological application in data stream and Environmental and weather data in stream. The number of items in the transaction is represented through record structure. In the data streams, some of the sequence in the transaction becomes frequently occurred and user by user other than the remaining items. So mining frequent itemset within in the time horizon becomes important for rainfall data to measure the level of rainfall per day to day. All of the existing frequent itemset mining methods frequent itemset are mined based on support count threshold value with sliding window and all the above mentioned work the importance of the attributes in the dataset may not considered as the multilevel, it may reduce the mining accuracy of the work.

The main objective of this work is to develop the frequent itemset mining result designed for finding the unidentified real time extraction of hidden information for datastream applications. It repeatedly investigates huge database to determine frequent patterns and extraction of motivating frequent itemset pattern with multilevel attribute condition for datastream applications at a high speed rate. The major contribution of the work as follows.

- In this work contribute multilevel fuzzy association rule mining method with time horizon for time series data.
- The fuzzy based association rule finding the membership value, becomes more difficult to solve this issue and tune the fuzzy membership value, in this work we contribute the artificial bee colony optimization method for fuzzy association rule mining time series with horizon data.

II. BACKGROUND STUDY

Young-Koo [9] developed a frequent itemset mining based on the creations of prefix-tree structure called Compact Pattern Stream (CPS) tree in dynamic manner for data stream data in transaction database .The propose CSP tree creates a tree structure based on the single pass scanning. The performance of the proposed CPS which is similar as Frequent Pattern (FP) growth technique. The created trees are automatically updated between time windows. The redesigned CPS tree is named as BSM method [11] and path adjusting method.

Pauray S.M. Tsai [12] develops a novel frequent itemset mining based on the window test is named as weighted sliding window (WSW) algorithm for datastream applications. Size of the window and the weight each window for transaction DataStream mining are specified by user. The threshold values of Size of the window and the weight each window also specified or defined by user. The frequent itemset transaction is divided into number of windows. New items are found in transaction it is checked by algorithm, whether it is frequent itemset or not, if it is frequent items it is added to frequent itemset mining results.

Problem of mining all frequent itemsets and maximal frequent itemsets also discussed in earlier works [13-14] along with landmark window test, both of works are generally based on the single pass scanning with more support count, it is named as summary frequent itemset forest (SFI-forest). Incremental updating is also carried out these work when new transaction will occurs update that result both frequent and infrequent itemset mining results. They conclude that the proposed SFI forest algorithm have consumes less memory and a timely manner.

The latter characteristic is also measured through Ao et al. [15]. They suggest FpMFI-DS [16]. In contrast through FpMFI, their result carries out a solitary pass over the data stream to construct the FP-tree. When the transactions go into the window, surrounded items are include into the FP-tree subsequent a lexicographical order, while at what time they come out, they are removed. They reduce infrequent itemset. From this step infrequent and frequent itemset are mined for data stream applications. Rough calculation can afford rough results but quicker solution, which can face real-time restriction in stream mining applications.

The memory consumption based on the data stream application problem also studied in earlier works. Chang and Lee [17] develop a stream for finding frequent itemsets (SWFI) inside a transaction-sensitive sliding window. The similar to SWFI-stream, some modification are done to improve the results of frequent itemset mining with prefix tree lattice structure new transaction data list are created it is named as the current transaction list (CTL). But all of the above mentioned work when new transaction added or existing transaction leaves is not supported by these work ,this problem is solved by our proposed work with multi-level frequent itemset mining for data stream applications .

III. PROPOSED METHODOLOGY

The mining of datastream possess new issues , because data streams application generally contains both discrete and continuous data which is changer over each a moment Mining frequent itemset for data stream application becomes also more challenge ,this problem is solved by proposing Window frequent itemset mining methods to mine all frequent item in the dataset at every moment .

To perform this process the frequent itemset the number of itemset in the database is defined as I, with frequent number of items occurrence vector is represented as v_I, with sliding block in W_S, for each itemset $I \in C$, the earlier position of the itemset in transaction database is maintained as last(I). When new items are added to position of the items in the transaction database is also updated based on the sliding window W_S. This procedure is performed from left to right in occurrence vector v_I for each one of the item in web anonymous dataset . Those items which is not added to frequent itemset mining results then the occurrence vector v_{I_r} is calculated through current sliding window size W,support count value of frequent items is calculated from v_I as:

$$Supp(I) = \sum_{t_i \in W} v_I(t_i)\, \mu(t_i) \qquad (1)$$

Algorithm 1 Window Itemset Shift (WIS)

C candidate list
F frequent itemset
S support of threshold
W frame of interest
Find optimal W_s
F ← frequents (W,S)
C ←candidates (W,W_S,S)
For all $I \in C$ do
Blind vector v_I last ← left most occurrence in W_S
end for
loop
move W forward of h slots
for all $I \in C$ do
last (I)+= h do
if lat(I) $>$ Length (Ws) then
remove I from C
end if

end for
$R_s \leftarrow$ Records entering W_s
For all $I_r I_r \neq 0$ $I_r \subseteq r, r \in R$ do
If $I_r \in C$ then
Last (I_r)=position (r)
Else
Insert I_r in C
Build vector v_r Last (I_r) ← leftmost occurrence in W_s
End if
End for
$F = \{I \in C \,|\text{Supp}(I) \geq S\}$
End loop

Proposed methods

Extraction of multilevel fuzzy association rules for frequent itemset in the datastream data becomes more difficult task, in order to overcome this problem, presents a novel multilevel fuzzy association frequent itemset mining for datastream application. During this mining association rules process the fuzzy membership function of the proposed multilevel fuzzy association rules are optimized using artificial bee colony optimization framework for testing window in multi level hierarchies in web anonymous transaction dataset. The original anonymous transaction data are frequently mined into several levels.

Multi level fuzzy association rule:

The proposed datastream frequent itemset mining creates a frequent itemset mining in multi-level taxonomy and group fuzzy membership are used to create fuzzy association rules in accord a known web anonymous transaction dataset. The proposed algorithm is

1. Use a series of numbers and the sign "*" to instruct the predefined categorization. The training is happening starting root through a zero value and sustained to next level beginning left to right through incrementing one importance"
2. D_i is the i-th web anonymous transaction dataset, where $1 \leq i \leq n$, add all of the frequent items through the equal foremost K number, calculate the k item count designed for every one of the groups in the web anonymous transaction dataset and remove the groups which is less than α predetermined support count value.
3. Regard as diverse fuzzy membership function intended for diverse web anonymous transaction dataset items. Each web anonymous transaction dataset data item has its own individuality and its individual fuzzy membership function. For each web anonymous transaction dataset D_i determination contain an anonymous frequent item state I_j^k , this is a j-th frequent item at point k, its value is represented as Q_{ij}^k is converted into f_{ij}^k with fuzzy regionh_j^k. R_j^k $(1 \leq l \leq h_j^k)$. The Q_{ij}^k is defined as,

$$\left[\frac{f_{ij1}^k}{R_{j1}^k} + \frac{f_{ij2}^k}{R_{j2}^k} + \cdots . + \frac{f_{ijh}^k}{R_{jh}^k}\right] \quad (2)$$

Calculate the importance of each one of the fuzzy region R_{il}^k from the web anonymous transaction dataset as,

$$Count_{il}^{k} = \sum_{i=1}^{n} f_{ijl}^k \quad (3)$$

Discover the greatest count importance articulate $MaxCount_{il}^{k}$ between Countilk values $(1 \leq l \leq h_j^k)$, as

$$MaxCount_{j}^{k} = max_{l=1}^{h_j^k}(Count_{il}^{k}) \quad (4)$$

If $MaxCount_{j}^{k}$ of a fuzzy region R_{jl}^k is equivalent to the smallest amount of support threshold value , subsequently place $MaxCount_{j}^{k}$ addicted to one frequent itemset.

4. If L_{lk} is null then enlarge k through one. If $r = 1$ *t*hen go to step 2 or else go to subsequently step.

Blind vector v_l last ← left most amount in W_S
end for
loop
move W frontward of h slots
for all $I \in C$ do
$last\ (I) += h\ do$
if $lat(I) > Length\ (Ws)$ then
Eliminate I from C
end if
end for
$R_s \leftarrow$ Records entering W_s
For all $I_r I_r \neq 0$ $I_r \subseteq r, r \in R$ do
If $I_r \in C$ then
Last (I_r)=position (r)

Else

Insert I_r in C

Build vector v_r Last (I_r) ← leftmost occurrence in W_s

5. The subsequent process is carried out in favor of diverse values.

i) If $r = 2$ create the new two level candidate frequent itemset set C_{2k} ,

ii) If $r > 2$ then create the new candidate frequent itemset set C_{rk}, with r-items on level k from L_{r-1k}

6. For every one attain candidate r-frequent itemset is specified as S through $(S_1, S_2, \ldots, S_r)$ items in C_{rk}

i) Calculate the fuzzy value of S with lowest amount operative of fuzzy reason, $f_{is} = min(f_{is1}, f_{is2}, \ldots, f_{isr})$ .Counts is the sum of $f_{is}, 1 \leq i \leq n$,

$$Count_s = \sum_{i=1}^{n} f_{is} \quad (5)$$

If Counts is better than then insert S into L_{rk}.

6. If value of the L_{rk} is null frequent itemset k is increased by and move to next step , or else augment r through one and move to step 8.

7. If $K > p$, move to step 11 or else set $r = 1$ move to step 1

8. Fuzzy association rules are created for r frequent itemset $S = (S_1, S_2, \ldots, S_r), r > 2$ with confidence value. Find all association rules $X \rightarrow Y$ where $X \subset S$ and $Y \subset S$ and $X \cap Y = \emptyset$ and $X \cup Y = S$, Calculate the confidence assessment for all association rules through:

$$confidence = \frac{\sum_{i=1}^{n} \min(f_{is})}{\sum_{i=1}^{n} \min(f_{ip})} \quad (6)$$

Artificial bee colony for fuzzy membership rules

To above mentioned fuzzy association rules are optimized based on the fuzzy membership value by proposing ABC algorithm [19-20].In generally ABC consists of three major categories of the bees such as employed bees, onlookers and scouts. The selected fuzzy membership value for association rules half of the fuzzy membership value are selected as employee bee and remaining is considered as for onlookers. In other words, the number of employed bees is equal to number of fuzzy rules with fuzzy membership value. The selected fuzzy membership value of the employee bee is rejected if the best frequent itemset are not mined and send to scout bee phase. The position of the each fuzzy membership value is initially defined and the nearest frequent itemset results are determined based on fuzzy membership function based on the fitness function solution. In initial step of the ABC algorithm the fuzzy membership function values are initialed through SN size value. Each fuzzy membership function is a D-dimensional vector. Here, D is the number of optimization parameters results for association rules in the multilevel. After initialization of fuzzy membership function then membership function results are optimized through maximum number of iteration, then best fuzzy membership function results are saved in memory and worst cases are removed, then highest probability value of the fuzzy membership value is chosen by onlooker bee using the following expression:

$$p_i = \frac{fit_i}{\sum_{n=1}^{SN} fit_n} \quad (7)$$

where fit_i is the fitness value of the optimized fuzzy rules i . The position values of the current fuzzy membership value are updated using following expression,

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) \quad (8)$$

where $k \in \{1, 2, \ldots, SN\}$ and $j \in \{1, 2, \ldots, D\}$ are randomly chosen indexes.. $\phi_{ij} \in [-1,1]$ If the value of the best fuzzy membership value reaches the predetermined fuzzy membership limit value then it is considered as the best fuzzy membership value for multi level fuzzy association rules ,the position value of the currently selected fuzzy membership value updated and defined in ,

$$x_i^j = x_{min}^j + rand(0,1)(x_{max}^j - x_{min}^j) \quad (9)$$

Detailed pseudo-code of the ABC algorithm is given below:

1. Initialize the number of population as fuzzy membership values for solutions
2. Evaluate the fuzzy membership values population
3. cycle=1
4. repeat
5. Produce new fuzzy membership values solutions for the employed bees by using fitness value and asses them
6. Apply the greedy assortment procedure
7. Calculate the probability values for the fuzzy membership used in equation (7)
8. Produce the new fuzzy membership values from onlookers and assess them
9. Apply the greedy assortment procedure
10. Determine the removed fuzzy membership values for the scout and substitute it with a new randomly selected fuzzy membership values solution
11. Memorize the best fuzzy membership solution achieved so far
12. cycle=cycle+1
13. until cycle=MCN

IV. EXPERIMENTATION RESULTS

To perform frequent itemset mining for stream data, in this work prefer MSNBC.com Anonymous Web Data dataset which is available machine learning UCI Machine Learning Repository. Anonymous dataset is directly chosen from Internet Information Server (IIS) logs for msnbc.com. The sequences of user log files from msnbc.com communicate to page views of a user log files at 28th September 1999. The event for each one of the sequences is selected based on the pages requested by user and it is recorded under different page categories. To perform this data stream process in consistent manner the page categories which is requested by user is grouped as ten number of clusters and the remaining duplicates page records are removed. The general categories of the pages which are requested by user categories are frontpage,news, tech, local, opinion, on-air, health, living etc.. To measure the performance accuracy of the above selected anonymous web server log data , four major existing apriori algorithm, apriori window test algorithm, apriori window test shift algorithm and proposed multi level fuzzy logic window test shift algorithm for frequent itemset mining , performance assessment is carryout based on the execution time and memory use results are shown in Fig.5and Fig.6.

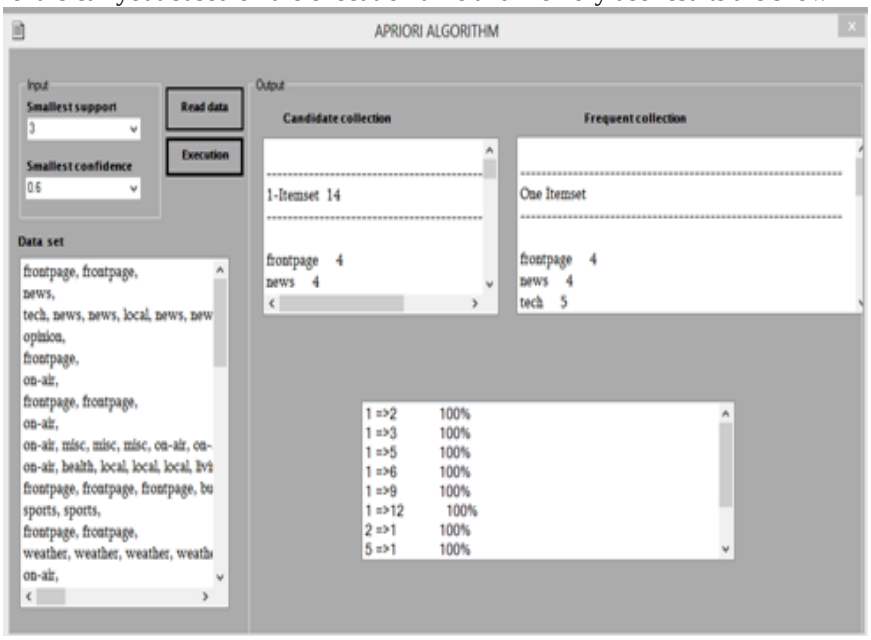

Fig.1. Frequent apriori itemset mining results for anonymous dataset

Fig.1 shows the frequent apriori itemset for anonymous dataset is exposed through minimum support value which is above than 3 and the minimum confidence value which is also greater than 0.6.

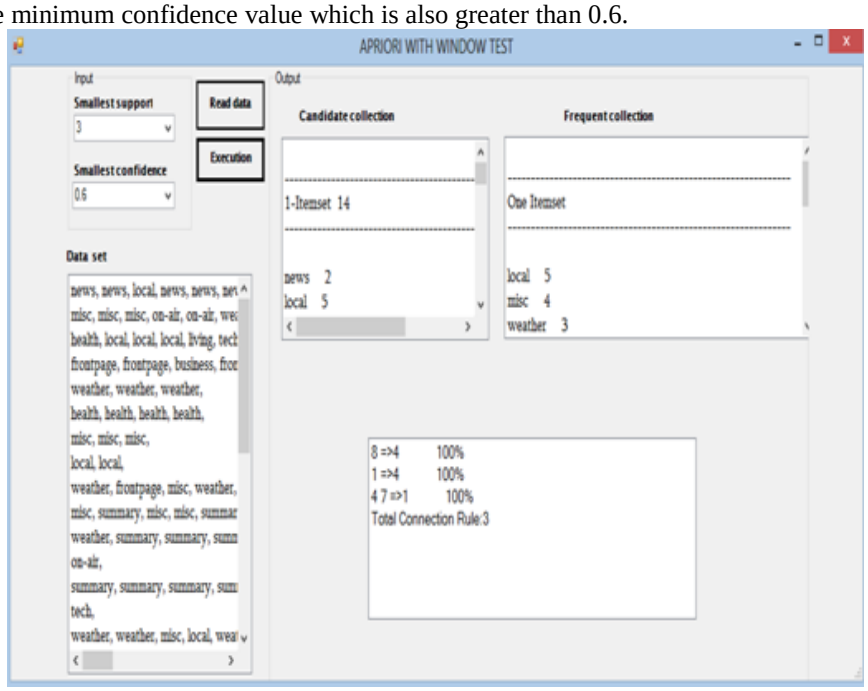

Fig.2.Frequent apriori window test itemset mining algorithm for anonymous dataset

Fig.2 shows the Frequent apriori window test itemset for Anonymous Web Data dataset is exposed with minimum support which is which is above than 3 and the minimum confidence value which is also greater than 0.6. Fig.3 have best itemset mining results by additionally supporting window test for online learning.

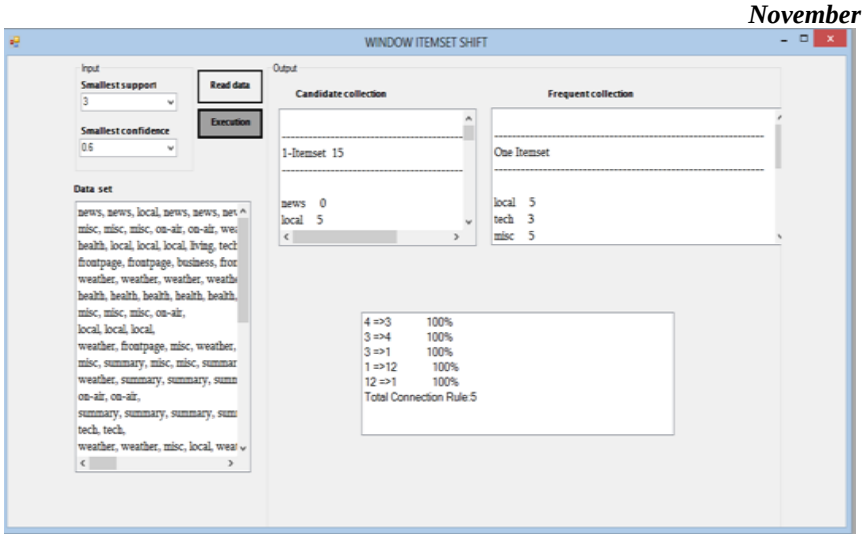

Fig. 3. Frequent apriori window test shift itemset mining algorithm for anonymous dataset

Fig.3 shows the Frequent apriori window test shift itemset for Anonymous Web Data dataset is exposed with minimum support which is which is above than 3 and the minimum confidence value which is also greater than 0.6. The Fig.3 results is differ from existing apriori window test algorithm, since the existing methods if new transaction is added in the dataset is not supported ,this problem is overcome by proposed apriori window test shift if newly added user transaction also added to frequent itemset mining.

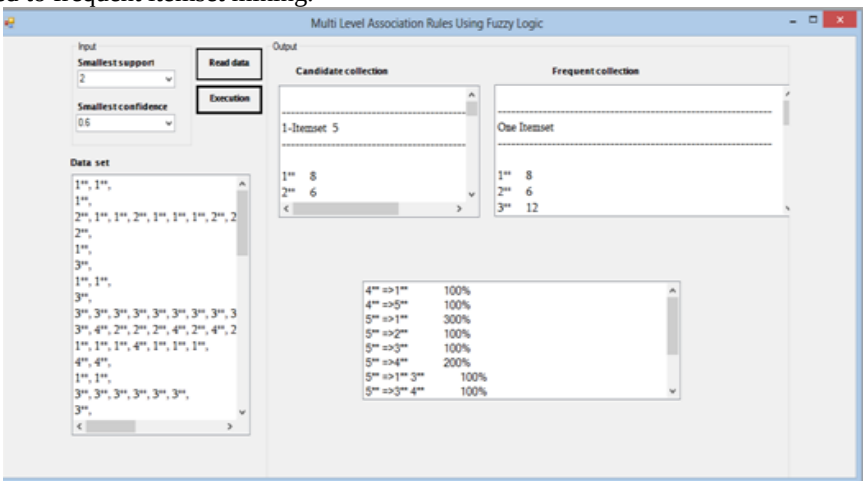

Fig.4.Frequent multi level fuzzy logic window test shift itemset mining algorithm for anonymous dataset

Fig.4 shows the Frequent multi level fuzzy logic window test shift itemset for Anonymous Web Data dataset is exposed with minimum support which is which is above than 3 and the minimum confidence value which is also greater than 0.6. It shows that proposed frequent itemset mining results have accurate results than the earlier methods since proposed methods performs frequent itemset in multi level and fuzzy rules are optimized using ABC algorithm.

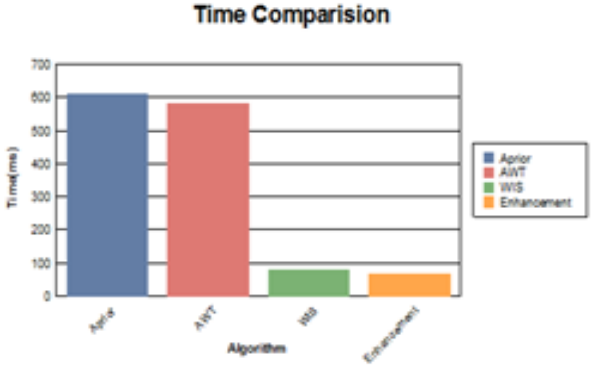

Fig.5. Time comparison results of the methods

Fig.5 shows the time comparison results of the frequent itemset mining methods such as the apriori, apriori with window itemset, window itemset shift and proposed multi level fuzzy association logic have taken less time than the existing methods, since fuzzy rules in the proposed system are optimized using ABC algorithm and multi level are considered for each attributes in the frequent itemset mining for anonymous web dataset.

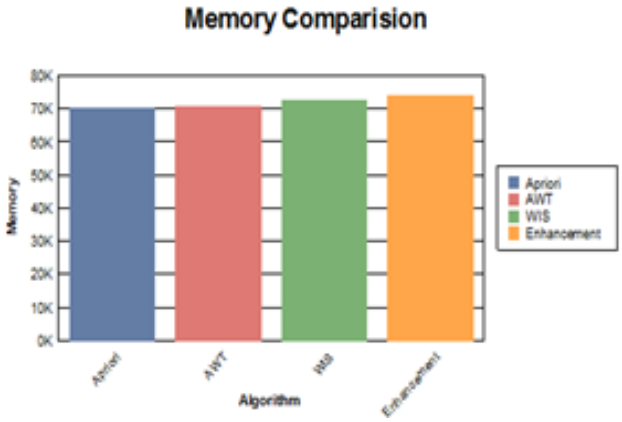

Fig.6 Memory comparison results of the methods

Fig. 6 shows the memory comparison results of the frequent itemset mining methods such as the apriori, apriori with window itemset, window itemset shift and the multi level fuzzy association logic, have take more memory consumption than the existing methods, since fuzzy rules in the proposed system are optimized using ABC algorithm and multi level are considered for each attributes in the frequent itemset mining for anonymous web dataset.

TABLE 1 PEFORMANCE COMPARISON RESULTS

Metrics	Apriori	Window test algorithm	Window itemset shift	Multi level fuzzy association
Time (Ms)	590	574	86	72
Memory consumption (KB)	68	70	73	74

The performance comparison for anonymous web server log data between four major existing apriori algorithm, apriori window test algorithm, apriori window test shift algorithm and proposed multi level fuzzy logic window test shift algorithm for frequent itemset mining , performance assessment is carryout based on the execution time and memory use results are tabulated in table 1.

V. CONCLUSION AND FUTURE WORK

In this paper, we studied the problem of data stream mining for frequent itemset. To solve the problem of frequent itemset mining in data stream application in this paper presents a novel multilevel fuzzy association rules for each frequent itemset in the data stream of anonymous data with different membership function. The multi level fuzzy association rules for data stream applications are mined at specific time interval. The degree of membership value for each attributes can be articulated on each slot of the window. The fuzzy membership function for association rule is tuned through artificial bee colony optimization (ABC) algorithm .Proposed the multi level fuzzy association rule with fuzzy ABC as an alternative solution, which maintain a memory of flowing candidates inside a concentrated test window. In the proposed work, minimum support value are defined to perform frequent itemset mining for web anonymous transaction dataset to obtain fuzzy rules for each user requirements. Experimental results also performed between three mining algorithms apriori algorithm, apriori window test algorithm, apriori window test shift algorithm, in measured in terms of the execution time and memory use. To speedup frequent itemset mining for web anonymous transaction database Master-slave parallel architecture is added to present algorithm. In the future, we will constantly effort to improve the ABC -based framework in favor of further complex mining problems.

REFERENCES

[1] L. Troiano, G. Scibelli, C. Birtolo, "A fast algorithm for mining rare itemsets, Intelligent Systems Design and Applications", *Ninth International Conference on, IEEE Computer Society*, 2009, pp. 1149–1155.

[2] L. Troiano, L.J. Rodríguez-Muñiz, J. Ranilla, I. Díaz, "Interpretability of fuzzy association rules as means of discovering threats to privacy", *Int. J. Comput. Math*. 89 (2012) 325–333.

[3] I. Díaz, L.J. Rodríguez-Muñiz, L. Troiano, "On mining sensitive rules to identify privacy threats*", in: J.-S. Pan, M.M. Polycarpou, M. Wozniak, A.C.P.L.F. Carvalho, Hybrid Artificial Intelligent Systems Lecture Notes in Computer Science* ,vol.8073, pp 232-241,2013.

[4] LeFevre, K., DeWitt, D.J., Ramakrishnan, R , "Workload-aware anonymization techniques for large-scale datasets", *ACM Transaction on Database System*,vol. 33,no.3,pp.1–17,2008.

[5] Aggarwal C (ed.) ,*Data Streams – Models and Algorithms*. Springer,2007.

[6] Gaber M, Zaslavsky A, Krishnaswamy S, *A Survey of Classification Methods in Data Streams*. In C Aggarwal (ed.), Data Streams – Models and Algorithms, Springer,2007

[7] L. Troiano, G. Scibelli, "A time-efficient breadth-first level-wise lattice-traversal algorithm to discover rare itemsets", *Data Min. Knowl. Disc.* (2013) 1–35.

[8] Anushree Gowtham Ringe, Deeksha Sood and Turga Toshniwal, "Compression and privacy preservation of data streams using moments", *Information journal of machine learning and computing*,2011.

[9] Syed Khairuzzaman Tabeer, Chowdary Farha ahmed, Byeong-Soo Jeong, Young Koo Lee, "Efficient frequent pattern mining over data streams" , *CIKM '08 Proceedings of the 17th ACM conference on Information and knowledge management*, pp.1447-1448,2008

[10] Tanbeer, S. K., Ahmed, C. F., Jeong, B.-S., and Lee, Y.-K. ,"CP-tree: a tree structure for single-pass frequent pattern mining" ,*S. In Proc. of PAKDD, Lect Notes Artif Int*, pp.1022-1027,2008.

[11] Koh, J.-L., and Shieh, S.-F ,*An efficient approach for maintaining association rules based on adjusting FP-tree structures*, Springer-Verlag, Berlin Heidelberg New York, pp.417-424,2004.

[12] Pauray S.M.Tsai ,*Mining frequent item sets in data streams using the weighted sliding window model*, Elsevier publication.2009.

[13] H.-F. Li, S.-Y. Lee, M.-K. Shan, "An efficient algorithm for mining frequent itemsets over the entire history of data streams", *1st International Workshop on Knowledge Discovery in Data Streams, in conjunction with 15th European Conference on Machine Learning, Pisa (Italy)*, 2004.

[14] H.-F. Li, S.-Y. Lee, M.-K. Shan, "Online mining (recently) maximal frequent itemsets over data streams", *Proceedings of the 15th International Workshop on Research Issues in Data Engineering: Stream Data Mining and Applications, RIDE'05*,pp. 11-18,2005.

[15] F. Ao, Y. Yan, J. Huang, K. Huang, "Mining maximal frequent itemsets in data streams based on fp- ee", *Proceedings of the 5th international conference on Machine Learning and Data Mining in Pattern Recognition, MLDM'07*, pp. 479-489, 2007.

[16] G.S. Manku, R. Motwani, "Approximate frequency counts over data streams", *Proceedings of the 28th international conference on Very Large Data Bases*, pp. 346-357,2002.

[17] C.-H. Lee, C.-R. Lin, M.-S. Chen, "Sliding window filtering: an efficient method for incremental mining on a time-variant database", *Information System*, vol. 30 ,pp. 227-244,2005.

[18] J.H. Chang, W.S. Lee, "A sliding window method for finding recently frequent itemsets over online data streams", *Journal of Information science Engineering* ,vol. 20 ,pp. 753-762,2004.

[19] D. Karaboga, *An Idea Based On Honey Bee Swarm For Numerical Optimization*, Technical Report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department, 2005.

[20] B. Basturk, D.Karaboga, "An Artificial Bee Colony (ABC) Algorithm for Numeric function Optimization", *IEEE Swarm Intelligence Symposium* , Indianapolis, Indiana, USA, 2006.

