Implementation of Bio-Inspired Algorithms in High Utility Itemset Mining

Keerthi Mohan, J. Anitha, G. Nandini

Abstract: ***Utility based itemset mining hasevolved as an important research topic in data mining, having application in retail-market data analysis, stock market prediction, online advertising and so on. Bio-inspired computation attempts to replicate the way in which biological organisms and sub-organisms operate using abstract computing ideas from living phenomena or biological systems.This study focuses on the application of bio-inspired algorithms on high utility itemset mining. A detailed analysis on the performance of thesealgorithmswere conducted on various parameters such as execution time, memory usage and the number of high utility items identified. Experimental result suggest Particle Swarm Optimization excels in its efficiency in execution time and memory usage.When the number of high utility items identified are concerned, it is Genetic Algorithm which outperforms Particle Swarm optimization and Bats algorithm.***

Keywords: Utility based mining,bio-inspired algorithms,high utility itemset

I. INTRODUCTION

Data Mining, an advanced type of data analytics is a methodwhichbrings outrelevant information present in enormous databases. High Utility itemset mining(HUIM),determinesrelated itemsets with high combined profit in sales, is a relevant research topic in data mining which has received interest among researchers.Prominent applications of HUIM includes online shopping, online advertising, biological gene data analysis [6],web click stream analysis and so on.

A. Association Rule Mining(ARM)

One of the prominent ways in finding patterns in data is ARM ,which can be features which either occur together or features which are correlated. The thought of Association rule for mining related patterns was first presented by [1],[2]. Common steps in mining association rules are: 1. Generating frequent itemset 2. Generating association rules. Market basket analysis is one of the common applications of ARM. Manyresearch has been conducted in the field of ARM and generating frequent itemset.

B. Frequent itemset mining(F.I.M.)

Association rule searches for frequent items in dataset. Here the interesting associations and correlations in the transactional and relational databases are identified.

Revised Manuscript Received on October 20, 2019.

Keerthi Mohan, Dept Of Computer Science & Engineering, DSATM, Bengaluru, India. E-mail: keerthiimohan@gmail.com

Dr. J.Anitha,Professor, Dept Of Computer Science & Engineering, DSATM, Bengaluru, India. E-mail: anitha.jayapalan@gmail.com.

G.Nandini, Dept Of Computer Science &Engineering,RRCE,Bengaluru, India. E-mail: nanduamma@gmail.com

Agarwal in 1993 published the original algorithm for frequent item set mining, is still widely used. This algorithm iterates by first scanning the database for finding the first 1-itemset, proceeding to find 2-itemset and so on. By joining frequent itemset with a length of n-1,candidate itemset having length of n were generated during each iteration; The candidate itemset gets added to frequent itemset after evaluation of its frequency.

Consider dataset with given transactions.

transaction ID	items
1	{A,C,D}
2	{B,C,E}
3	{A,B,C,E}
4	{B,E}
5	{A,B,C,E}

Lets say minimum support count is 3. Relation hold is maximal frequent => closed => frequent. From the above example identified 1-itemset is as follows.

1-frequent:
{A}=3
{B}=4;

{C}=4;
{D} = 5;

2-itemset are identified as follows.

2-frequent:
{A,B}=2
{A,C}=3
{A,D}=3
{B,C}=3
{B,D}=4

{C, D} = 4

C. High Utility Mining(HUM)

In frequent itemset mining there are cases where infrequent but valuable items were ignored while mining. For example, in market analysis there may present more frequent but unprofitable items as well as less frequent but more profitable items. In frequent itemset mining these cases were not considered. Hence researchers came up with a concept calledhigh utility itemset mining in which the utility of the item is considered for mining rather than the frequency. High utility mining is considered as chronologically proceeding frequent item set mining. It is considered as a generalized FIM.

Analysing and generating items with high utility is called Utility mining. Usually, cost, profit and other user defined expressions defines the utility value of an item. In a database for transactions there are followingqualities of utility: 1) external utility 2) internal.Itemset'sutility can be generated by combining internal utility and external utility.An item is considered as having high utility if the utility is higher than the user set threshold. [3].

Itemset Utility (iu) = UtilityInternal (ui) * Utility External (ue)

D. Algorithms

The concept called High utility itemset mining was first introduced by [4] . A detailed study on utility mining is done by [5]wherea mathematical model was proposed which was on the foundation of the two properties, utility bound and support bound. To prune itemsets we can use a heuristic measure obtained from the utility bound property.

A candidate generating algorithm called 2 phase algorithm which is proved to have maintained downward closure property was proposed by [5].This algorithm competently prune down the number of candidates and find the complete set of high utility items precisely.

Ahmed et al in[6],suggested an Incremental High Utility Pattern Mining(IHUP) algorithm. Previous mining results are used to avoid unwanted calculations during an update to the database or during a change in the user defined threshold.

To discover utilities without candidate generation HUI-Miner algorithm was introduced by [7]. To signify the utility information, a utility list is defined for items of length 1 of qualified items.The algorithm iterates recursively constructing a utility list of item sets of length k utilising a pair of utility list having itemsets which have alength of k-1. To create a utility list a single scan of the database would be enough. By performing join operation of utility list of lesser length, a larger itemset can be obtained.

An effective strategy for pruning candidate itemset for mining high utility itemset was suggested by [8]. The algorithms such as. Utility Pattern Growth(UP-Growth) and UP-Growth+ were introduced by this author. A tree based data structure,termedas Utility Patter tree(Up-Tree) is used to maintain the high utility itemset information. Hence two database scan is required to generate candidate itemset. This proposed technique was proved to performsignificantly in terms of runtime particularly when there is larger database computations.

Fast High Utility Miner(FHM) proposed by [9] is proved to have improved HUI-Miner. It integrates a newapproach called Estimated utility co-occurrence pruning(EUCP).This approach uses a utility list data structure for reducing the number of join operations when mining HUI's and evades costly joins.FHM is 6 times faster than HUI-Miner.

A notableadvancement in high utility itemset mining was a one phase algorithm,Efficient high utility itemset mining(EFIM) by [10].For reducing the cost of database scans an effective database projection and transaction merging technique was proposed here.

Another HUI-Miner based algorithm by [11] is a MHUI miner where a tree structure is introduced to avoid unnecessary utility list construction in HUI miner for itemset expansion.

E. Bio Inspired Algorithms

Traditional algorithms of HUIM have to handle exponential problem of a huge search space. Also manual setting of minimum utility threshold is not an effective way. These problems were addressed by bio inspired algorithms.

[12] proposes two genetic algorithms(GA) named $HUPE_{umu}$-GA and $HUPE_{wumu}$-GA to address the above mentioned problems.$HUPE_{umu}$-GA is selected when our primary concern is search space and memory usage. Here data analyst inputs minimum utility threshold.Optimal HUI's without specifying minimum utility threshold were generated in $HUPE_{wumu}$-GA.

A swarm intelligence approach called Binary PSO(Particle Swarm Optimization) approach has been suggested by[13] to address the below given problems of Genetic Algorithm.

1. PSO requires few parameters compared to genetic algorithms. 2. In Genetic algorithm, in the evolution process crossover and mutation operations were essential to randomly generate next solutions. A good amount of computations were required here to find the satisfied HUI in the initial step, which will be an overhead when there are large number of distinct items.

Contributions of [13] are 1. A discrete PSO based algorithm called HUIM-BPSO was defined. It integrates sigmoid updating strategy an TWU(Transaction weighted utility) model[14]. 2.An OR/NOR tree structure is developed to reduce the numerous database scans,by early pruning the invalid combinations of the particle.

[15] proposes HUIM-ACS algorithm based on Ant Colony Optimization(ACO) to find HUI's through a precise routing graph and TWU model. Here positive pruning rule and recursive pruning rules were used to decrease search space and enhance an efficient itemset search.

A frame work for mining high utility itemsets using Bats Algorithm(BA),geneticalgorithm(GA), and Particle swarm optimization(PSO),was suggested by [16]. The suggested approach uses a roulette wheel selection method of selecting the optimal selection. The defined framework is compared with the state of art bio inspired algorithms and proved to have substantial enhancement in the number of HUI identified, memory usage and execution speed.

II. ALGORITHM COMPARISON

A. Comparison of $HUPE_{umu}$_GRAM and HUIM PSO

[12] proposed a bio inspired genetic algorithm for high utility itemset mining.An elaborate study on the performance of HUIM-BPSO and $HUPE_{umu}$_GRAM was done here.The algorithm is tested against two real time dataset such as chess and mushroom[15]. Chess is a small but dense dataset containing a large number of transactions.

The dataset mushroom includes descriptions of hypothetical samples corresponding to 23 species of mushrooms. To have consistent testing methods all the below analysed algorithms are tested against above mentioned two datasets derived from the UCI Machine Learning Repository[15]. Performance analysis of this algorithm is done based on the execution time taken and number of HUI's discovered.

A. 1. Execution time.

In[13] HUPE$_{umu}$_GRAM algorithm is improvised by constructing OR/NOR tree and termed as HUPE$_{umu}$_GRAM^{+} and HUPE$_{umu}$_GRAM algorithm without tree is termed as HUPE$_{umu}$_GRAM^{-}.Similarly HUIM-BPSO is also improvised as HUIM-BPSO^{+} having the construction of OR/NOR tree and HUIM-BPSO^{-} without tree. All the four algorithms are designed with 10000 iterations and populations size is set as 20. In chess dataset the execution of the 4 algorithms gave the result as shown in Table1

Table 1. Execution time

ALGORITHM	Execution Time
HUPE_GRAM^{+}	10,795 sec
HUPE_GRAM^{-}	10,764sec
HUIM-BPSO^{+}	10,568sec
HUIM-BPSO^{-}	9913sec

(a) Chess dataset

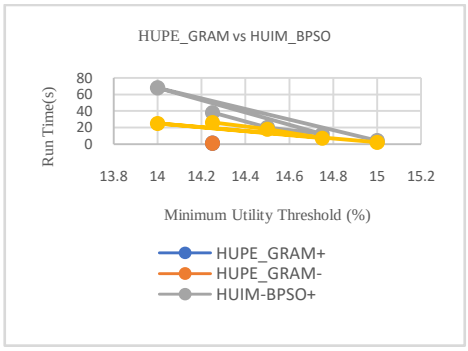

Fig.1HUPE_GRAM vs HUIM_BPSO(Chess dataset)

(b)Mushroom Dataset

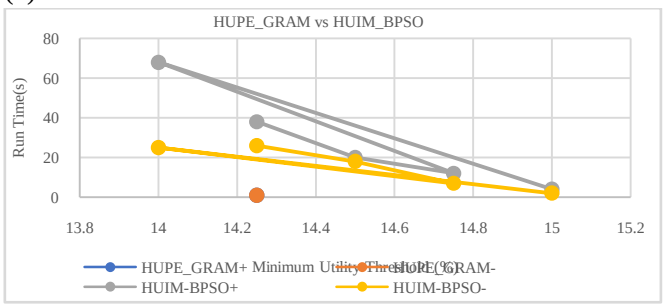

Fig 2.HUPE_GRAM vs HUIM_BPSO(Mushroom dataset)

It is proved from Fig2 that the construction of OR/NOR tree in HUPE_GRAM^{+} has affected in its performance compared to HUPE_GRAM^{-}.HUIM-BPSO has an efficient execution time in both the datasets compared to its counterpart.

A.2. Number of HUI's

The TWU model proposed in[16] has proved to discover the actual and complete HUI's from the quantitative database .In[13] the result obtained is compared with the number of HUI's discovered through TWU model. HUIM-BPSO^{+} algorithm generates nearly the same number of HUI's as in TWU model. A sample result at threshold value 25% in chess data set is shown in Table2.

Table 2:No of HUI's

ALGORITHM	Noof HUI's discovered
HUPE_GRAM^{+}	11
HUPE_GRAM^{-}	12
HUIM-BPSO^{+}	75
HUIM-BPSO^{-}	83
TWU Model	98

At threshold higher than 25% , the same number of HUI's were generated on both HUIM-BPSO^{+} and HUIM-BPSO^{-} as that of TWU.

(a) Chess Dataset

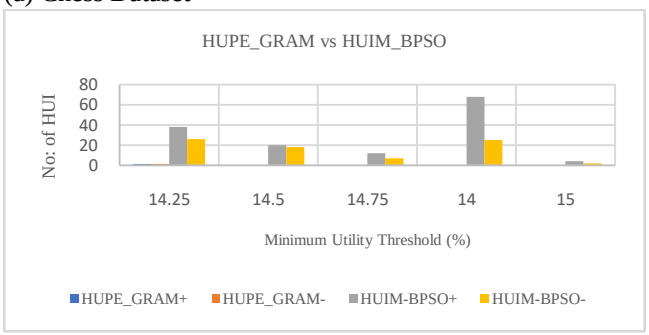

Fig 3:HUPE_GRAM vs HUIM_BPSO(Chess dataset)

(b)Mushroom Dataset

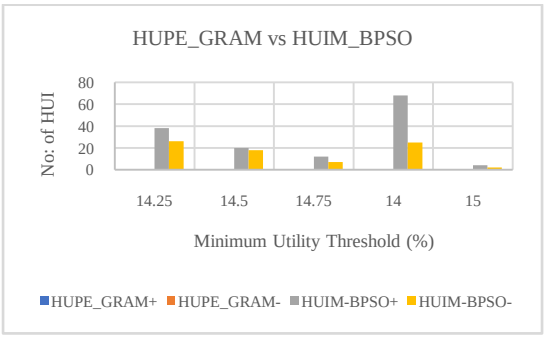

Fig 4:mushroom dataset on noof HUI

B.Comparison of HUIM-BPSO and HUIM-ACS

An Ant colony system(ACS) based algorithm for mining HUI's was proposed by[17]. It is proved to have outperformed other existing evolutionary algorithms for mining HUI's. the reason behind ACS better performance is because it does not calculate the utility of the same itemset again. The process here does not allocate memory to a node till the ant arrives.Using HUIM-ACS the number of HUI's generated are almost same as that of TWU model in Chess dataset.

Even the computation time of HUIM-ACS is less compared to other evolutionary algorithms.

B.1.Execution Time

(a) chess dataset

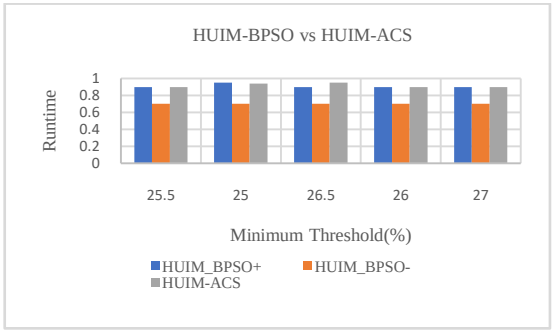

Fig.5:Execution time on chess dataset

(b) Mushroom dataset

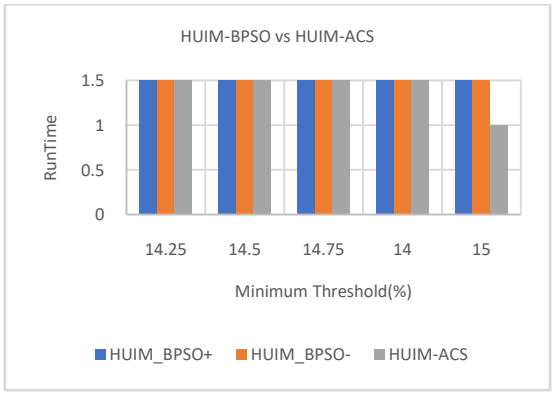

Fig 6:Execution time on mushroom dataset

B.2 Number of HUI's

Experimental results shows that HUIM-ACS generate more of less equivalent number of HUI's as that of TWU model. Well statedheuristic function and routing graph in HUIM-ACS algorithm enhances the generation of HUI in the proposed approach.

(a) chess dataset

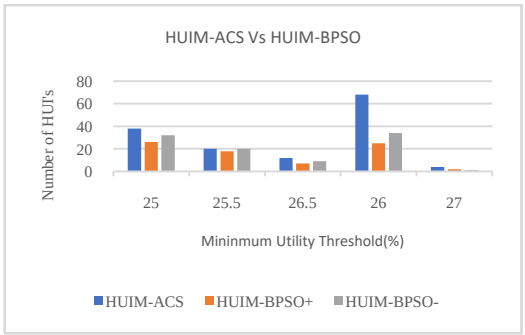

Fig 7:HUI identified on chess dataset

(b)Mushroom dataset

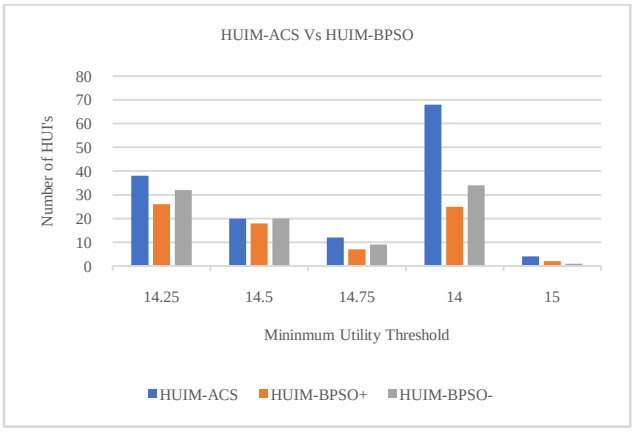

Fig 8.HUIM-ACS Vs HUIM-BPSO(mushroom dataset)

C. Comparison of Bio-HUIF-BA, Bio-HUIF-GAand Bio-HUIF-PSO

A new framework for high utility itemset mining using bio inspired algorithms proposed by[18]. In the proposed technique three bio inspired algorithms are compared on the dense datasets chess and mushroom. The efficiency of these algorithms on varied minimum utility threshold for each data set is analysed. The result shows that Bio-HUIF-PSO attains the finest performance over the other two.Fig 9 shows a comparative result of the above mentioned three algorithms with HUIM-PSO. Fig 9 clearly shows that Bio-Huif-PSO an Bio-Huif-BA takes less compilation time compared to others analysed.

C.1 Execution Time

(a) Chess dataset

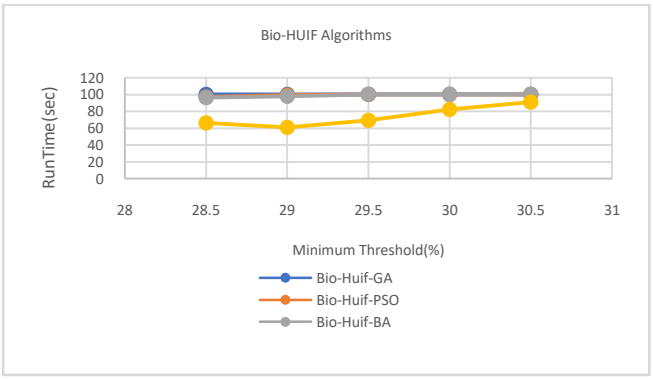

Fig 9.Bio-HUIF Algorithms on execution time(Chess dataset)

(b)Mushroom Dataset

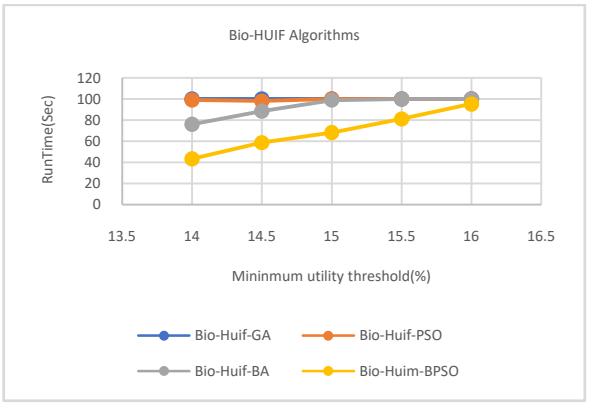

Fig 10. Bio-HUIF Algorithms on execution time(Mushroom dataset)

C.2 Number of HUI's

(a)Chess dataset

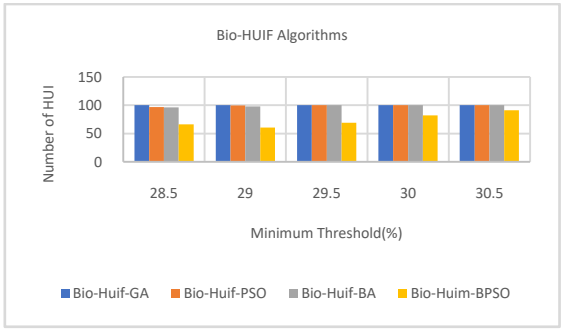

Fig 11. Bio-HUIF Algorithms no: of HUI's(Chess dataset)

(b) Mushroom dataset

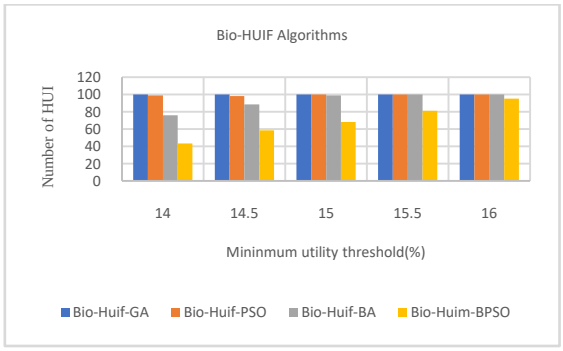

Fig 12: Bio-HUIF Algorithms on Noof HUI's(Mushroom dataset)

Fig 11 and 12 clearly show the effectiveness of the projected algorithm in generating HUI's.

III. CONCLUSION

In this work, a detailed study on the performance of bioinspired algorithms for mining high utility itemset is performed and the result is summarized. Bioinspired algorithms were efficient in mining high utility itemset on various utility threshold values.Roulette wheel selection proposed by[18] made an important improvement on the following bioinspired algorithmsGA,PSO and BA in mining HUI's. As a further study,other bioinspired algorithms will be analysed on their efficiency on mining HUI considering sparse dataset also.

REFERENCES

1. Agarwal R., and Srikant R., "Fast algorithms for mining association rules", In the Proceedings of 20th International Conf. Very large Data Bases, pp.487-499, 1994.
2. Agrawal R., Imieliński T, Swami A,"Mining association rules between sets of items in large databases", Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207, 1993.
3. Vincent S.Tseng,Bai-Enshie,Cheng-Wei Wu and Pjillip S.Yu, "Efficient Algorithms for Mining High Utility Itemset from Transactional Databases",8 August 2013,IEEE Transactions on Knowledge and Data Engineering ,Vol 25 pp 1172-1786,2013.
4. Chan Q.,YangY., Shen D., "Mining high utility itemsets", in:Proceedings of the 3rd IEEE International Conference on Data Mining , Melbourne , Florida, pp.19-26, 2003.
5. H.Yao, H.J.Hamilton, C.J.Butz, "A foundation approach to mining itemset utilities from databases", in: Proceedings of the Third SIAM International Conference on Data Mining, Orlando, Florida , pp.482-486, 2004.
6. Ahmed C.F. , Tanbeer S.K., JeongByeong-Soo, Lee Young-Koo, "Efficient tree structures for high utility pattern mining in incremental databases", in: IEEE Transactions on Knowledge and Data Engineering 21(12) ,2009.
7. Liu M. and Qu J., "Mining High Utility Itemsets withoutCandidateGeneration",CIKM'12,Maui,HI,USA, ACM, October29-November 2,2012.
8. Vincent S.Tseng,Bai-Enshie,Cheng-Wei Wu and Pjillip S.Yu, "Efficient Algorithms for Mining High Utility Itemset from Transactional Databases",8 August 2013,IEEE Transactions on Knowledge and Data Engineering ,Vol 25 pp 1172-1786,2013.
9. Philippe Fournier Viger ,Cheng-Wei Wu,SouleymaneZida,Vincent S. Tseng, "FHM: Faster High-Utility Itemset Mining using Estimated Utility Co-occurrence Pruning", Proc. 21st International Symposiumon methodologies for Intelliegnet Systems (ISMIS 2014),Springer,LNAI,pp 83-92,2014.
10. Souleymane Zida, Philippe Fournier-Viger, Jerry Chun-Wei Lin,Cheng-Wei Wu,Vincent S. Tseng, "EFIM: A Highly Efficient Algorithm for High-Utility Itemset Mining", 30 December 2015,Mexican International Conference on Artificial Intelligence Advances in Artificial Intelligence and Soft Computing pp 530-546,2015.
11. Peng.A.Y.Koh,Y.S,&Riddle P(2017)."mHUIMiner.A fast high utility itemset mining algorithm for sparse datasets" In *Proceedings of the Pacific-Asia conference on knowledge discovery and data mining*(pp.196-207)
12. Kannumuthu.S,&Premalatha K(2013)."Discovery of high utility itemsets using genetic algorithm" International Journal of Engineering and technology(IJET),4866-4880
13. Lin.J.C,W.YangL,Fourier-VigerP,Wu.J,M T Hong(2016),"A Binary PSO Approach to mine high utility itemsets".Soft Computing,3,1107-1135.
14. [14] Vincent S.Tseng,Tzung-Pei Hong,Guo-Cheng Lan(2010)"Mining High Transaction Weighted Utility Itemsets"second International conference,IEEE,2010.
15. [15]Lichman.M(2013).UCIMachinelearningrepository.http://archive.ics.uci.edu/ml.
16. [16]LiuY,LiaoChouhary A(2005)A two phase algorithm for fstdiscovery of high utility itemsets.Lecture Notes ComputerSci:689-695
17. [17]Jimmy Ming-Tai,JustinZhan,Jerry Chun-Wei Lin,"An ACO-based approach to mine high-utility itemsets",Knowledge-Based systems 116(2017) 102-113.
18. [18]W Song,C.Huang:"Mining HUI Using Bio-Inspired Algorithms:A Diverse Optimal Value Framework"IEEE open Access Journel,2018

AUTHORS PROFILE

Keerthi Mohan: Is presently working as Assistant Professor in the department of ComputerScience and Engineering at DayanandaSagarAcademy of Technology and Management,Bangalore,India.She received herBachelor of Engineering(B.E) and Master of Engineering(M.E) in Computer Science and Engineering from Anna University,Chennai. She is pursuing her Ph.DinVishvesarayaTechnological University,Belagavi,Karnataka. Her research interests are big data analytics,Utility mining,Bio-inspired algorithms.

Dr J Anitha , is currently working as a Professor, Department of Computer Science andEngineering, DSATM, with an experience of 17 years. Graduated BE from KIT, Tipturunder Bangalore University in Computer Science and Engineering. She obtained herMaster degree ME in Computer Science and Engineering from Anna University, Chennaiand PhD degree from Anna University, Chennai in the area of Knowledge Based systems.She has 50 publications in International and National conferences, 20 publication inNational journals and International Journals. Life Member of ISTE, CSI.

Nandini G, Is currently working as Assistant Professor in the Dept. Of Computer Science and Engineering at Rajarajeshwari College of Engineering , Bangalore, Karnataka, India. She received her Bachelor of Engineering (B.E), Master of Technology (M.Tech) from Visvesvaraya Technological University (VTU). Currently she is Pursuing Ph.D in VTU, Belagavi, Karnataka, India. Her research interests include Wireless Sensor Network, Internet of Things. Actively involved in professional organizations like CSI, ISTE.