International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(4-1), pp.037-044
DOI: http://dx.doi.org/10.21172/1.841.06
e-ISSN:2278-621X

ASSOCIATION RULE MINING USING HBPSO

Amit Kumar Chandanan[1], Dr Kavita[2] and Dr M K Shukla[3]

Abstract— Association rules are one of the most researched areas of data mining. In the area of data mining finding association between data set or product in marketing field plays important role. Association rule mining is a way to find relations or co-relations among a set of information available. The aim to generate rules for giving multiple data from various databases. Analysis of data can be possible with the help of sequential access of data from database. In case of sequential access of data, it may cause multiple times same rules to be generated. It is desired to find a solution to get out of those unnecessary association rules due to the complex characteristics of serial data. Although many numbers of serial association rule with the use of either sequence or temporal constraint as prediction model, these two models did not consider with the repetition during the process of rule mining for the database. Duplicate data set generates redundant association rules with respect to support and confidence. In proposed method, remove redundant rules to improve efficiency of association rules with HBPSO. The experimental result comparison shows the improvement in quality of association rules.

Keywords—Redundant Rules, Association Rule, HBPSO, Genetic Alorithms

I.INTRODUCTION

Association Rule Mining [1, 2] is a technique in Data Mining that is used to reveal the hidden correlation among the different items of transactions exhibit in the database. An association rule can be described as any rule that involves association relationship among different objects (or itemset) such as an object implies to another or the occurrences of these objects, alone or with other objects. Association rules [1, 2] are, in general, if-then rules that work on some conditional probability. The two main parameters used for such conditions are support and confidence. The support can be concocted of as the percentage that all the items in the rules will satisfy. The confidence then again can be characterized as the degree of certainty that an association Let in a database D there are a number of transactions T. In each transaction there is number of items having a place with itemset I. If n is the distinct number of items in D then I = {i1, i2…in} is a set of all the items present in database. Also any transaction $t \in T$ may contain variable set of items over I, i.e., ii, ij, ik $\subset$ I. Every transaction is associated with an interesting identifier. T_ID. The association rule is of the shape of $X \Rightarrow Y$, where X, Y $\subset$ I and X Y = $\emptyset$, where X is the consequent of the rule.

The association $X \Rightarrow Y$ holds for any transaction T in D if its bolster S of any item is satisfied. Support s of an association rule R is the percentage of transaction t that contains XUY (both X and Y) which is the probability P(XUY) of the items in transaction.

Support ($X \Rightarrow Y$) = sup(R) = P(XUY).

The association rule R, of the form $X \Rightarrow Y$ has confidence C in transaction set T in D if the conditional probability satisfies, i.e., the transaction t containing X also contains Y. It is taken as P(Y/X).

Confidence(c) = confidence ($X \Rightarrow Y$) = conf(R) = P(Y/X) = support_count(XUY)/support_count(X) = sup(R)/sup(X).

An example of an association rule is as follows,

Cheese -> beer [sup = 10%, conf = 80%]

This rule says that 10% of clients purchase cheese and beer together, and those who buy cheese additionally purchase beer 80% of the time [3],[4].

II. ASSOCIATION RULE MINING(ARM)

ARM discovers the frequent patterns among the many item sets (IS). It pursuit to extract intriguing association, frequent pattern, and correlations amongst IS in the data repositories [4]. The formal statement of ARM problem turned into to start with unique [5]. Let I = I1, I2, … , Im be a collection of m precise attributes, T be the transaction that incorporate a group of items such that T □ I, D be a database with extraordinary transactions Ts. An association rule is an insinuation inside the shape of X □ Y, where X, Y □ I are sets of itemstermed ISs, and X □ Y= □ .X is known as antecedent. Y is known as consequent.

[1] *Jayoti Vidyapeeth Women's University,Jaipur,India*
[2] *Jayoti Vidyapeeth Women's University,Jaipur,India*
[3] *Jayoti Vidyapeeth Women's University,Jaipur,India*

The rule means X implies Y. The two significant basic measures of association rules are support(s) and confidence (c). Since the database is enormous in size, users concern about only the frequently bought items. The users can pre-define thresholds of help and confidence to drop the rules which aren't so useful. The two thresholds are named support and minimal confidence [6].

Support (s) is defined as the proportion of records that contain X □ Y to the whole records in the database. The amount for each item is augmented through one, whenever the item is crossed over in extraordinary transaction in database in the course of the direction of the scanning.

Support (XY)= support sum(XY)/

$$\text{Support(XY)}= \frac{\text{support sum XY}}{\text{Overall records in database D}}$$

Confidence (c) is defined because the proportion of the no. of transactions which contain X □ Y to the whole files that include X, the place, if the ratio outperforms the threshold of self assurance, an association rule X □ Y may also be produce.

Confidence (X/Y) $=\frac{\text{support(XY)}}{\text{support (X)}}$

Confidence is a degree of strength of the association rules, if the confidence of the association rule X = Y is 80 percent, it infers that 80 % of the transactions which have X also incorporate Y together, likewise to verify the interestingness of the rules distinct minimal self assurance can be pre-outlined by using users. Association rule mining is to note association rules that satisfy the pre-described minimal support and self-belief [7]. The drawback is sub-separated into sub trouble. The first is to search out the ISs which existences surpass a predefined threshold, most often mention to as IS. The following is to generate association rules from huge IS with the obstacles of minimal self-belief. If probably the most huge IS is Lk, Lk= {I1, I2…Ik-1, Ik}, then organization ideas are generated with these IS. Checking the arrogance with the rule of thumb {I1, I2, …, Ik-1} {Ik}, it may be made up our minds for interestingness. Via deleting the last items, the opposite rules are created within the antecedent and inserting it to the ensuing, then the confidences of the new rules are checked to come to a decision the interestingness. The procedures iterated till the antecedent turns into empty. The important sub problem will also be two folded into candidate huge IS generation process and FIS generation procedure. Those IS whose aid exceeds the aid threshold called as large or FIS, these IS which can be predictable to be huge or frequent are identified candidate IS. An efficient model has classification rules with high confidence and large support [8].

III SEQUENTIAL PATTERN MINING

SPM proposed by way of Agrawal[16] on analyzing big data from supermarket, is an vital branch of data mining. Sequential pattern mining, crucial branch of information mining. Consecutive example mining,… popular in web get to pattern analysis, market basket analysis, fault detection in network, DNA sequences etc, which needs to find all of the sequential pattern that surpasses the base support threshold[17]. Conventional algorithm on sequential pattern mining are classified categories: successive example that outperforms the base support threshold[17]. Traditional calculation on SPM are characterized classes: Apriori, GSP, projection and SPADE[18].

Apriori use codes generating-testing methods and is simple and easy to implement. However, Apriori generates a massive amount of items-sets and scans the database frequently, for that reason wastes a massive amount of time GSP [19] is based on the frequency-item mining algorithm of Apriori and uses time limitations, sliding window to improve the efficiency while it needs traverse the database multiple times.SPADE[20] by Zaki transforms the data into a vertical form,but generates masses of items-sets. Generating item-sets and branch trimming consumes great amount of time. Based on projection, Freespan[21] and Prefix span[22] use "divide-conquer" to divide the raw database into smaller projection databases, and then mine the sequential pattern in smaller databases. Divide conquer increases the efficiency and has excellent expansion. However, this method spends incredible measure of time in dividing database into projection databases and has the bottle neck in constructing projection databases and scanning data [23].

IV FREQUENT PATTERN MINING(FPM)

FPM is the method of mining data in a IS or certain patterns from huge databases, that must chain the least support threshold. A frequent is a pattern which befalls typically in a dataset. These frequent patterns may present in different forms like frequent IS, sequential pattern or substructure. Frequent IS generally suggests that a fixed of gadgets that often occurs together in a transactional data set. As an example, milk and sugar. The patterns which purchaser have a tendency purchase in subsequences point out to frequent sequential pattern. For example, customer generally first purchases the laptop and then thinks for purchasing anti-virus software for it. A substructure can take many structural types like graph, trees or lattices and these varieties may also be mixed with IS or subsequences. And if these substructures show certain frequency in operations, then it is called as a frequent structure pattern.

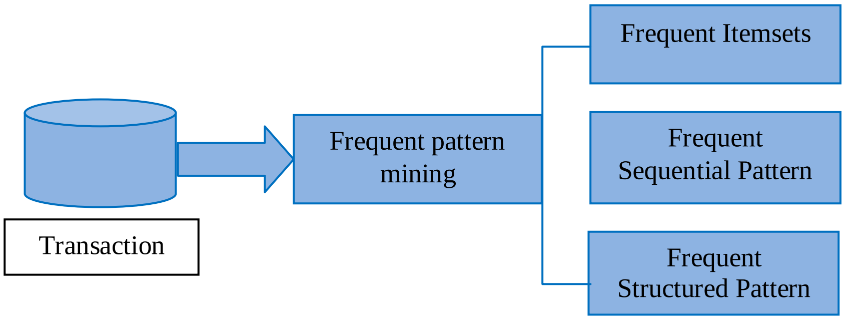

Fig. 1 FPM framework

Frequent IS mining plays vital part in numerous data mining arenas as association rules, warehousing, correlations, clustering of high-dimensional biological data, and taxonomy. Problems occurring during market basket analysis are one main motives of frequent IS mining. A set of object bought through buyer in market basket analysis in any transaction is known as tuple. An association rule extracted from market basket database depend on the principle which if certain items are bought in transaction, then it is possible which certain other items are also get buy. A very significant thing while mining the association rule is the IS mining. Therefore, several techniques are present to produce frequent IS, with the aid of that we can efficiently mine the association rules. Large number of algorithms is present to mine the frequent IS [9].

Sequential Pattern Mining PM proposed by way of Agrawal [10] on analyzing big data from supermarket, is an vital branch of data mining. SPM, crucial branch of data mining. Consecutive example mining, popular in web get to pattern analysis, market basket analysis, fault detection in network, DNA sequences etc, which needs to find All the sequential pattern that surpasses the base support threshold [11]. Conventional algorithms on sequential pattern mining are classified categories: successive example that outperforms the base support threshold [11]. Traditional calculation on SPM is characterized classes: Apriori, GSP, projection and SPADE [12]. Apriori use codes generating-testing methods and is simple and easy to implement. However, Apriori generates a massive amount of items-sets and scans the database frequently, for that reason wastes a massive amount of time GSP [13] in view of the frequency-item mining algorithm of Apriori and uses time limitations, sliding window to improve the efficiency while it needs traverse the database multiple times. SPADE[14] by Zaki transforms the data into a vertical form, but generates masses of items-sets. Generating item-sets and branch trimming consumes extraordinary measure of time. Based on projection, Freespan [15] and Prefix span [16] use “divide-conquer” to divide the raw database into smaller projection databases, and then mine the sequential pattern in smaller databases. Divide conquer builds the productivity and has excellent expansion. However, this method spends incredible measure of time in dividing database into projection databases and has the bottle neck in constructing projection databases and scanning data [17].

This research work combines the concept of binary particle swarm optimization to generate rule and mutation concept from genetic algorithm to represent association rules without redundant association rules. No change in quality of dataset [22].

Another kind of database is emerging both in the research community and in the commercial market place. This new sort of database allows developers to represent hierarchical data in XML form while providing query, transaction and security services similar to commercial relational database software. (In fact, hierarchical databases are not new; some earlier databases used hierarchical data models such as CODASYL. The hierarchical version is taking part in a rebirth with the arrival of XML [24]). While it will most likely not replace all relational databases, it is being aimed at just the sort of problem we have defined in implementing the NMP missions and technology database. The database canonically implements the data hierarchy. A hierarchy, in this case, can be thought of as a tree structure. An example of such a a structure that is natural to the majority of the community structure that is recognizable to the greater part of the group …is the file system directory, as seen in the Windows or Macintosh stack of folders metaphor. Here, parent or higher level folders may contain child, or lower-level folders, which, in tum, contain folders of yet a lower level. Each folder may also contain a specific type of data or file. If used as intended (hut not enforced), "child folders", i.e., those contained within their "parent folder" contain data that is a subset of the types of data contained in the parent folder. In addition, pointers (aliases or "shortcuts") are provided to allow linking logically connected, but non-adjacent, folders. In most cases, the folders are displayed, not as a tree, but as an indentured list. While the indentured list is a convenient and space-efficient format, it does, unfortunately, tend to bide the tree structure. Still it is a familiar construct with which most individuals are familiar, and for which the underlying structure is readily grasped.

The upsides of a hierarchical database are basically the inverse of the disadvantages of the relational database, i.e. With a hierarchical database, the hierarchy is the native structure. There is no need to craft custom interfaces to hide the actual database structure or to interpret it for the user. Hierarchical data is stored in a hierarchical format (XML). A simple display interface allows the user guide access to the structure as implemented and the data as stored. System maintenance and debug efforts are much reduced. In this business, surprises are not good, and this ability to view the structures as they are tends to minimize surprises [25].

V LITERATURE SURVEY SUMMARY

Table 1. Summary of liturature survey in association rule

Author Name	Algorithm	Work
Ruilin Liu [2016] et. al [18]	SLAM algorithm	An efficient rare association rule mining algorithm called spark-based rare association rule mining (SRAM) which leverages not only the efficiency of FP-growth algorithm but also the powerful big data processing mechanism of spark platform. We have implemented our algorithm on the start platform and tested with various of data sets.
Morteza Zihayat [2016] et. al [19]	BigHUSP	Another structure for mining HUSPs in tremendous data. A dispensed and parallel algorithm referred to as Big HUSP is proposed to discover HUSPs efficiently. At its heart, Big HUSP makes use of multiple Map Reduce-like steps to process information in parallel. We also propose some of pruning techniques to reduce seek area in disbursed surroundings, and consequently decrease computational and communique charges, whilst nevertheless preserving correctness
Hong-Yi Chang [2015] et. al [20]	Apriori algorithm and FP-Growth algorithm	We developed a method that combines the Apriori and FP-Growth algorithms with MapReduce to rectify this problem. In experiments carried out, we varied the block length of the Mapper to obtain execution performance higher than the ones of the Apriori and FP-Growth algorithms
Masome sadat Hoseini [2015] et. al [21]	FP-growth algorithm	A new approach is presented for mining Cantree, and it's evaluated to reveal its development over the FPgrowth technique that mine FP tree.

VI PROPOSED WORK

In this research work authors are using binary particle swarm optimization with mutation (HBPSO Hybrid Binary Swarm Particle). With the help of mutation operator introduced in this approach, authors provide solution of problem of premature convergence into a local minimum, which occur in particle swarm optimization. Mainly the mutation operator provide sharpness in convergence and it provides the best possible solution. The particle that have high mutation probability are taken for generating rules. After getting the best solution, rules are represented by using Michigan rule representation approach.

VII RESULT ANALYSIS

In this section authors present in detail the experiments undertaken to prove that the previouslyproposed techniques work.For each proposed piece of work we outline the experimentalenvironment that was set up and the results that were obtained. We also analyse the results andprovide further discussionsThe experiments for nonredundant association rule m ining, the proposed HBSPO (for multi-level datasets) and for various dataset i.e books,grocery etc.

We used 5 randomly self-bult dataset which were composed 50,100,500,1000 and 2000 transactions on it . the statistic(support count ,confidence) for this data is detailed in table 2. the Elapsed time using base algorithm with the dataset is 5.544824 seconds. The association rule generated in the case is 5(R1 to R5).

Table 2**.** Association Rule generated by BPSO for 50 transactions

Rule Number	Antecedent	Consequent	Support	Confidence
R1	Youth Books	Geog Books	66.00	84.62
R2	Player Book	Youth Books	74.00	100.00
R3	Child Books	ItalCook	58.00	72.50
R4	Child Books	Youth Books	60.00	75.00
R5	Youth Books	Child Books	60.00	76.92

The experimental result for same data set using our method (name the method) is show in table 3. the Elapsed time is 1.386758 seconds. The redundant association rule has been remove by our proposed (name the method) technique.

Table 3. Association Rule generated by H BPSO for 50 transactions

Rule Number	Antecedent	Consequent	Support	Confidence
R1	Cook Books	ItalCook	44.00	80.00
R2	Cook Books	DoItY Books	20.00	36.80
R3	Child Books,Geog Books	Cook Books	39.00	58.50
R5	Cook Books	Art Books	25.00	45.92

The experiment shows that the performance of the proposed method has been improved by 20% with respect to number rule generation. So this is better than previous approach.

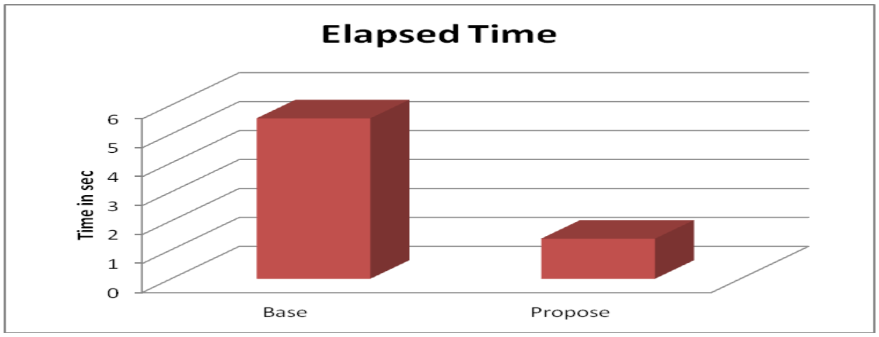

Fig 2. Run time comparision of BPSO and modified BPSO with 50 data set

Same base method has been tested with 100 transactions in dataset, the run time for this time is 30.868013. The support and confidence for association rules are shown in table 4

Table 4. Association Rule generated by BPSO for 100 transactions

Rule Number	Antecedent	Consequent	Support	Confidence
R1	Player Book	Horror Book	76.00	98.62
R2	Child Books	Ref Books,Geog Books	68.00	81.93
R3	DoItY Books	DoItY Books	30.00	60.00
R4	Geog Books	Youth Books	45.00	60.00
R5	Ref Books,Player Book	Art Books	49	73.72
R6	Ref Books,Geog Books	Youth Books,Player Book	67.00	91.92

When proposed method used to tested with 100 transactions in dataset, the run time for this time is 1.323444. The support and confidence for association rules are shown in table 5.

Table 5. Association Rule generated by HBPSO for 100 transactions

Rule Number	Antecedent	Consequent	Support	Confidence
R1	Cook Book	Art Book	25.00	45.62
R2	DoItY Books	Cook Book	20.00	40.93
R3	Horror Book	Italcook	70.00	98.50
R4	DoItY Books	Ref Books,Geog Books	44.00	88.02

Six rules have been generated using previous method (base paper) and Rules R5 and R6 are redundant. Now we run our algorithm, we find 4 rules without any redundancy. The performance of our work is increase by 34% with respect to rule generated.

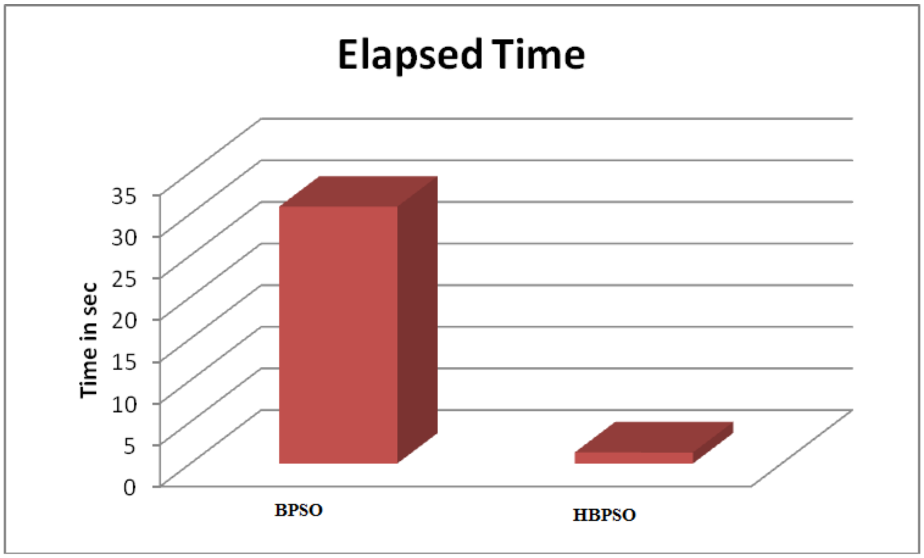

Fig.3. Run time comparision of BPSO and HBPSO with 100 data set

Same base method has been tested with 500 transactions in dataset, the run time for this time is 5.342488, the support and confidence for association rules are shown in table 6.

Table 6. Association Rule generated by BPSO for 500 transactions

Rule Number	Antecedent	Consequent	Support	Confidence
R1	Ref Books	Cook Books Youth Books	39.20	52.62
R2	Geog Books	Child Books	69.00	92.93
R3	Cook Books, Child Books	Ref Books Youth Books	34.00	76.58
R4	Ital Cook	Cook Books	44.40	56.35
R5	Cook Books Ref Books Child Books	Ital Cook, Horror Book	32.60	83.16
R6	Child Books	DoItY Books, Horror Book	38.86	46.08
R7	Ref Books	Cook Books, Child Books	39.20	42.69

The experimental result for 500 transactions in dataset with our proposed method is shown in table 7 with association rules generates, the run time for this time is 1.436457 and run time comparison shown in figure 4.

Table 7. Association Rule generated by HBPSO for 500 transactions

Rule Number	Antecedent	Consequent	Support	Confidence
R1	Ital Cook	DoItY Books	44.20	56.09
R2	Horror Book	Cook Books	43.60	56.19
R3	Horror Book	Child Books	62.20	80.15
R4	Geog Books	Child Books	69.90	92.49
R5	Child Books, Geog Books	Ital Cook	63.00	91.30
R6	Ital Cook	DoItY Books	44.20	56.09

Seven rules have been generated using previous method (base paper). Now we run our algorithm, we find six rules without any redundancy. the performance of our work is increase by 14.28% with respect to rule generated

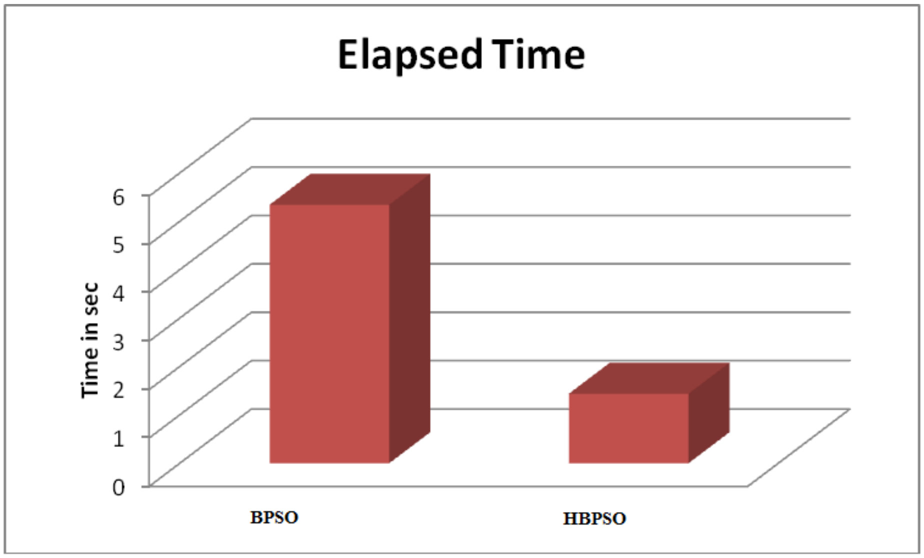

Fig.3. Run time comparision of BPSO and HBPSO with 500 data set

Tree is built with the help of a MATLAB programming. This is the second tree which is made on the result table of propose result on hundred records of which elapsed time is 1.436457 seconds. The above tree is consisting of the root node and their sub nodes which are named as the parent nodes; here parent node also consist of the sub nodes which are called the child nodes. There are different nodes which are consisting of different numbers in the ratio.

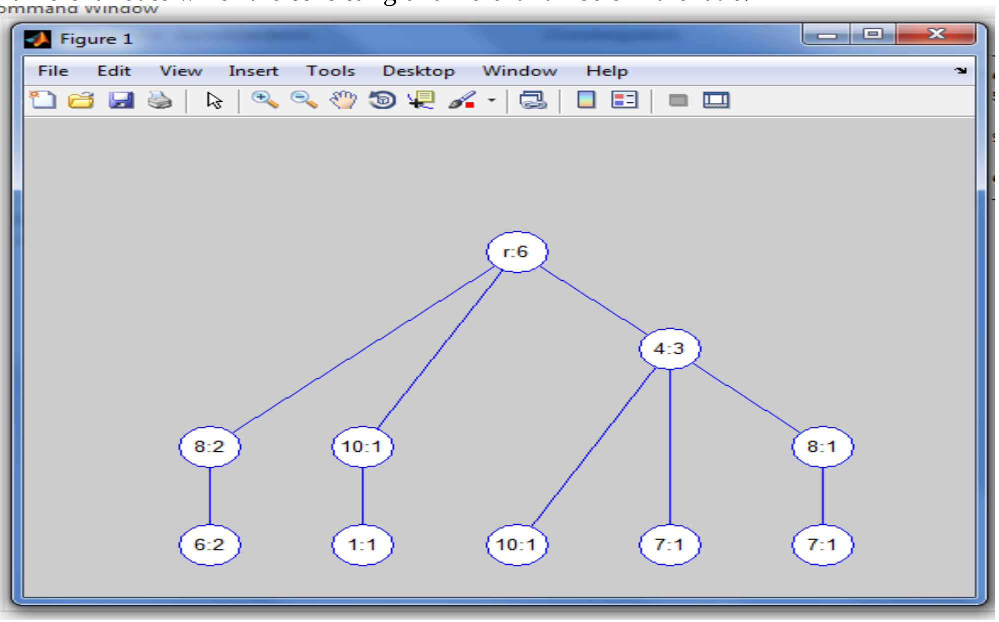

Fig.4 Binary tree for propose result on 500 records

In this section we have discussed the result in details of various datasets which showed the comparison of base result and the propose result of the following:

i. Book dataset (on 50 records, 100 records, 500records, 1000 records, 2000 records)
ii. Food dataset (on 50 records, 100records, 250records,500records, 1000records)
iii. Grocery dataset (on 50records, 100records, 250 records, 500 records, 1000records)

These dataset records have comparison based on base paper and propose paper, showing the result having the figure and also by tree which is built with the help of MATLAB programming. Consisting of root node, sub node, and the parent node. Here it is shown the elapsed time with base paper and propose paper as well.

REFERENCES

[1] Agrawal, R., Imielinski, T., Swami, A. "Mining association rules between sets of items in large databases." pp. 207-216, ACM SIGMOD (1993)

[2] Agrawal, R. and Srikant, R. "Fast algorithms for mining association rules." Proceedings of VLDB(1994)

[3] Feng Tao, Fionn Murtagh & Mohsen Farid, "Weighted Association Rule Mining using Weighted Support and Significance Framework", ACM SIGKDD (2003)

[4] Satpal Singh, Vivek Badhe, "Profit Association Rule Mining with Inventory Measures", 978-1-5090-0076-0/15 IEEE(2015)

[5] Agrawal, R., Imielinski, T. and Swami, A. Mining association rules between sets of items in large databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data, 207-216, (1993)

[6] Qiankun Zhao and Sourav S. Bhowmick. Association rule mining: A Survey. Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003116(2003)

[7] Agrawal, R. and Srikant, R. Fast Algorithms for Mining Association rules. Proc. 20th VLDB conference, Santiago, Chile(1994).

[8] Viet PhanLuong and Lif, "Mining normal and abnormal class association rules" IEEE 27th International Conference on Advanced Information Networking and Applications. 968-975, Barcelona, (25- 28 Mar 2013)

[9] Dhanabhakyam, M and Punithavalli, M. A Survey on Data Mining Algorithm for Market Basket Analysis. Global Journal of Computer Science and Technology, Vol. 11(11), (2011)

[10] R. Agrawal, R. Srikant. Mining Sequential Pattern[C]//Pro. of the 11st Int. Conf. on Data Engineering , Taipei,3:3-14(1993)

[11] HAN J. KAMBER M. Concept and Technology of Data Mining [M] . Fan Ming,Meng Xiao-feng Translation . BeiJing ： Machinery Industry Press. 320—336 (2001).

[12] Chen Zhuo,Yang Rui-ru,Song Wei,Song Ze-feng. Survey of sequential pattern mining[J]. APPLICATION RESEARCH OF COMPUTERS.,07:1960-1976(2008).

[13] Srikant R,Agrawal R. Mining sequential pattems: generali-zations and performance improvements [C]. EDBT 96:Proceeding of the 5th International Conference on Extending Database Technology: Advance in Database Technology .UK,London :Springer-Verlag,1996:3-17

[14] Zaki M. SPADE: An Efficient Algorithm for Mining Frequent Sequence [J] .Macheine Learning,42(1):31-60(2001)

[15] Han J， Pei J， Mortazavi-Asl B， et al. Freespan : freguent patternprojected seguential pattern minin［A］.In : Proceedings of the International Conference on Knowledge Discovery and Data Mining ACM［C］.Montreal， Canada 355 -359(2000).

[16] Pei J,Han J,Mortazavi-Asl B,et al. Mining sequential patterns by patterngrowth: the prefixspan approach [J]. IEEE Transaction On Knowledge and Data Engineering ,16(11):1424-1440(2004).

[17] XUE Fei and SHAN Zheng, "A Improved Sequential Pattern Mining Algorithm Based on PrefixSpan", xue(2016)

[18] Ruilin Liu, Kai Yang, Yanjia sun, Tao Quan, Jin Yang, " spark based Rare Association Rule Mining for big Datasets", 978-1-4673-9005-7/16, IEEE(2016).

[19] Morteza Zihayat, Zane Zhenhua Hu, Aijun An, Yonggang Hu, "Distributed and Parallel High Utility Sequential Pattern Mining", 978-1-4673-9005-7, IEEE(2016).

[20] Hong-Yi Chang , Yih-Jou Tzang, Jia-Chi Lin, Zih-Huan Hong, Ting-Yun Chi, Chun-Yen Huang, "A Hybrid Algorithm for Frequent Pattern Mining Using MapReduce Framework", 978-1-4673-8600-5/15 IEEE(2015).

[21] Masome sadat Hoseini , Mohammad Nadimi Shahraki, Behzad Soleimani Neysiani, "A new algorithm for mining frequent patterns in CanTree",978-1-4673-6506-2/15 IEEE(2015).

[22] A K Chandanan, Kavita, M K shukla, "Association Rule Mining Using Modified BPSO" International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR), ISSN(P): 2249-6831; ISSN(E): 2249-7943 Vol. 7, Issue 2, Apr 2017, 29-36