A New Approach for Periodic Frequent Pattern Mining

Manjeet Samoliya[1], Akhilesh Tiwari[2]
[1, 2] Department of CSE & IT
[1, 2] Madhav Institute of Technology and Science, Gwalior (M.P), 474005, India

Abstract- *In this paper, present a new approach on rough set periodical mining that implements a new concept of the Genetic Algorithm. The use of the Genetic algorithm optimizes the patterns generated till date. The periodic mining gives the patterns in a synchronous way and thus, the patterns generated are synchronous. The use of evolutionary approach is done so as to evolve new patterns and new rules. In the proposed new approach, the use of genetic algorithm on the basis of the seasonal analysis gives the idea about the capability of the patterns generated. Thus, the results generated are assumed to be better in all the aspects that can take place in real world scenario.*

Keywords- Association rule mining, data mining, genetic algorithm, periodical pattern mining, rough sets.

I. INTRODUCTION

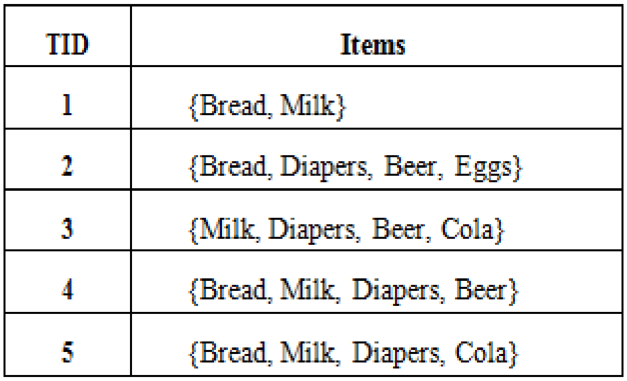
Data mining is a vast area of research now-a-days. It has been increasing at a very fast pace. It refers to extracting or “mining” knowledge from large amounts of data [1]. By performing data mining, interesting knowledge, reliabilities, or high-level information can be extracted from the database and viewed or browsed from different approaches. The discovered knowledge can be applied to decision making, process control, information management, and query processing [21].

Rapid advances in data collection and storage technology have enabled the organization to accumulate vast amounts of data. However, extracting useful information has proven extremely challenging. A part of data mining is association rule mining that gives the idea about the associations between the various items Association rule mining is a well-known method to discover interesting rules and relations between various items in large databases. Based on the concept of strong rules, Rakesh Agrawal et al.[2] introduced association rules for finding regularities between items in large-scale exchange information recorded.

Association rules are complete through examining information for frequent if/then patterns and utilizing the criteria support and confidence to identify the most important relationships. Support is an evidence of how frequently the items show up in the database. Confidence demonstrates the times quantity the if/then statements discovered to be valid [3].

Table 1 illustrates a case of such information, normally known as market basket transactions. In this table, every row corresponds to the transaction, which is a unique identifier labeled TID contains and also contains an itemset bought through an assumed client. Association rule mining has been explained with the help of an example:

Table 1: Transactions of Itemset



TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

For example, the following principle can be removed from the information set demonstrated in Table1: {Diapers}=>{Beer}

The rule suggests that a strong relationship exists between of the sale diapers and also beer because numerous clients who purchase diapers additionally purchase beer. Retailers can utilize this kind of principles to help them classify novel opportunities for the cross- selling their clients products.

The support sup(X) of an itemset X is defined as the database transactions proportion which contains the itemset. In the example database, the itemset {Bread, Diapers, Beer, Eggs} has a support of 1/5=0.2since it happens in 20% of every transaction (1 out of 5 transactions). The argument of supp() is a set of preconditions, and thus becomes more restrictive as it grows.

The confidence of a rule is defined as conf(X=>Y)=supp(X $\cup$Y)/ supp(X). For example, the rule {Bread, Milk}=>{Beer} has a confidence of 0.6/0.6 =1.0 in the database, which implies that for the 100% of transactions containing bread and butter the guideline is right (100% of the times a client buys {bread, milk}, {beer} is bought as well).Note that supp(X$\cup$Y) means the support of the items in

X and Y union[4].

The Temporal Data Mining part of the Knowledge Discovery in Temporal Databases procedure is concerned with the algorithmic means through which temporal patterns are removed and counted from temporal data. Temporal data mining is concerned with the analysis of temporal data and for discovering temporal patterns and regularities in sets of temporal information. Temporal data mining has led to a novel way of interacting with a temporal database [5].

Rough Sets represent a different mathematical approach to vagueness and uncertainty. The rough set methodology is based on the premise that lowering the degree of precision in the data makes the data pattern more visible. The results of the rough set approach are presented in the form of classification or decision rules derived from a set of examples [6]. The concept of approximation in rough sets can explain as below:

A. *Lower approximation* of the set X with respect to R is the set of each objects, which can be for the certain categorized as X with respect to the R (are certainly X with respect to R).

$$R(X) = \{x \in U : R(x) \subseteq X\}$$

B. *Upper approximation* of a set X with respect to R is the set of the each object which can be probably categorized as X, with respect to R (are possibly X in the vision of R).

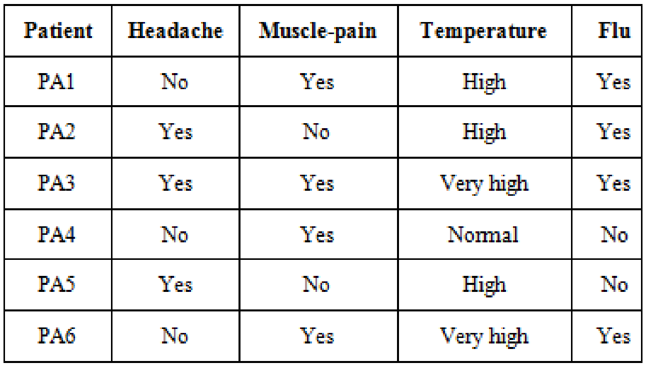
$$R^*(X) = \{x \in U : R(x) \cap X \neq \emptyset\}$$

C. *Boundary region* of a set X with respect to R is the set of the each object, which can be categorized neither as X nor as not-X with respect to R.

$$R(X) = R^*(X) — R^*(X)$$

Thus, a set is rough (imprecise) if it has a nonempty boundary region; otherwise the set is crisp (precise). Information is frequently presented as a table, the columns of which are labeled through attributes, rows through interest objects and entries of the table are attribute values. For example, in a table concluding data about patients suffering from a various disease objects are patients (strictly speaking their ID's), attributes can be, for example, blood temperature, body pressure etc., whereas the entry corresponding to object Smith and the attribute blood pressure can be usual. Such tables are called as data systems, attribute-value tables or data tables. Below an example of knowledge table is presented. We are providing information about patients, in Table 2 [7].

Table 2: Patients Info.



Patient	Headache	Muscle-pain	Temperature	Flu
PA1	No	Yes	High	Yes
PA2	Yes	No	High	Yes
PA3	Yes	Yes	Very high	Yes
PA4	No	Yes	Normal	No
PA5	Yes	No	High	No
PA6	No	Yes	Very high	Yes

Columns of the table are labeled by attributes (symptoms) and rows by objects (patients), whereas entries attribute values in the table. Thus, table row can be seen as data about the particular patient.

For instance, patient pa2 is categorized in the table through following attribute-value set (Flu, yes), (Headache, yes), (Muscle-pain, no), (Temperature, high), which form the information about the patient. In the table patients pa2, pa3 and pa5 are mixed up concerning the property Headache, patients pa3 and pa6 are ambiguous regarding qualities Muscle-pain and the Flu, and patients pa5 and pa2 are indiscernible with respect to the attributes Temperature, Headache and Muscle-pain. Hence, for example, the attribute Headache creates two elementary sets {pa1, pa4, pa6} and {pa2, pa3, pa5}, whereas the attributes Muscle-pain and Headache form the following elementary sets: {pa1, pa4, pa6}, {pa2, pa5} and {pa3}. Essentially one can describe elementary sets produced through any attributes subset.

Patient pa2 has the flu, whereas patient pa5 does not, and they are indiscernible with respect to knowledge acquisition properties Temperature, Headache and Muscle-pain, hence flu cannot be characterized in forms of properties Temperature, Headache and Muscle-pain. Hence, pa5 and pa2 are the boundary-line cases, which cannot be accurately grouped in perspective of the accessible learning. The remaining patients pa1, pa3 and pa6 show symptoms which empower us to categorize them with certainty as having the flu, patients pa5, and pa2 cannot be avoided as having flu and also patient pa4 for sure does not have the flu, in showed of the displayed symptoms. Consequently the lower estimate of the arrangement of patients having flu is the set {pa1, pa3, pa6} and the higher approximation of this set is the set {pa1, pa2, pa3, pa5, pa6}, whereas the boundary-line cases are patients pa2 and pa5. Likewise pa4 does not have the flu and pa2, pa5 cannot be rejected as having flu, thus the lower estimate of this idea is the set {pa4} whereas - the upper approximation – is the set {pa2, pa4, pa5} and the limit region of the idea “not flu” is the set {pa2, pa5}, the same as in the past case.

Genetic Algorithm is an adaptive heuristic search algorithm which based on the concepts of genetics and natural selection. As such they signify an intelligent random exploitation search used to resolve improvement issues. Although randomized, Genetic Algorithms are through no means random, instead they exploit historical data to direct the search into the region of improved execution inside the search space [8].

The genetic algorithm is a method for solving both constrained and unconstrained optimization problems that are based on natural selection. The genetic algorithm then genetic algorithm simulates the survival of fittest among individuals over the consecutive generation for problem-solving purposes[12].

```
Selection of initial population
Evaluation of the fitness of the individual
Defining the average fitness of the population
Repeat
    Best suited individuals are selected for reproduction
    Mating is performed at random
    Crossover is applied
    Mutation is performed
    Evaluation of the fitness of all individuals
    Determining average fitness of the population
Until terminating condition is reached
```

In all generation, the fitness of each individual in the population is evaluated, multiple individuals are selected from the present population (based on their fitness), and modified (recombined and possibly mutated) to form a new population. Fitness function was defined as follows [9, 10, 11]:

Fitness= α1*Support +α2*Confides – α3*NA 1

Where NA is the number of characteristics participating in the produced rule and coefficients α control effect of each parameter inside fitness function. The selection of the fittest individual is done with the help of various methods. Various methods to select the best chromosomes, for example, roulette wheel selection, Boltzmann selection, tournament selection, rank selection, steady state selection and some others [13]:

A. Roulette wheel selection:

The basic part of the selection procedure is to stochastically choose from one generation to make the basis of the other generation. The necessity is that the fittest individuals have a larger survival chance than weaker ones. This replicates nature in that the fitter individual will tend to have a perfect probability of the survival and will go forward to form the mating pool for the next generation. Weaker individuals are not without any chance. In nature, such individuals may have the genetic coding that may prove valuable to future generations [14,15,16].

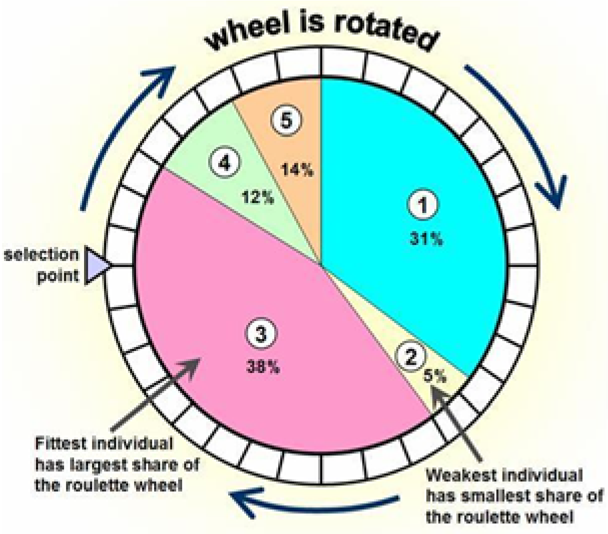

wheel is rotated
5
14%
4
12%
1
31%
selection point
3
38%
2
5%
Fittest individual has largest share of the roulette wheel
Weakest individual has smallest share of the roulette wheel

Fig.1: Roulette wheel

II. RELATED WORK

Omari et al., [17] developed a new temporal measure for interesting frequent itemset mining. Frequent itemset mining helps in searching for powerfully associated items and transactions in large transaction databases. This measure is based on the fact that interesting frequent itemsets are typically covered by several recent transactions. This minimizes the cost of searching for frequent itemsets by minimizing the search interval. Additionally, this measure can be used to enhance the search approach implemented with the Apriori Algorithm.

An algorithm called R_Apriori is created by Chen Chu-xiang et al., [18] for the problems with the decision-making domain. Initially, the cores situations are mined with Algorithm of the Rough Attribute Reduction, 1-frequent itemsets and corresponding sample set is then created with use mining cores set through the Apriori Algorithm. After the above-mentioned stage, the multi-stage frequent itemsets and the corresponding confidence and support can be gained through the sample set intersection operator.

Market Basket Analysis is the most important function of information discovery in databases. All things considered, market creates databases, for the most part, contain temporal coherence, which couldn't catch by means of standard association rule mining. Therefore, there is a requirement for creating calculations that show such transient cognizance's inside the information. Schluter et al., accumulates various thoughts of temporal affiliation principles and presents a system for mining a large portion of these sort (calendar-based, cyclic - and lifespan)in a business basket database, improved by two new tree structures. These two tree structures are called as EP-and ET-Tree, which are obtained from available systems enhancing standard affiliation guideline mining. They are utilized as a representation of the database and henceforth make the revelation of temporal association rules which are extremely proficient. There have been at various modern ponders in the field of periodic pattern mining.

There have been at various modern studies in the field of periodic pattern mining. Ozden et al.[19] problem defined the which is based on discovering cyclic association rules as finding cyclic connections between items presence within transactions. In their research, the input information was a transaction set, in which all consisted of a set of items. In addition, every transaction was tagged with a time of execution.

Han et al.[20] displayed a few algorithms to effectively mine partial periodic patterns, by exploring few features partial periodicity related, such as an Apriori feature and also max-sub pattern hit sets a feature, and through shared mining of multiple times. In order to time the restriction cyclic association rule, Han, et al. Used confidence in the measure how is a periodic pattern important.

Periodic patterns are categorized based on several conditions. Periodic patterns are classified as frequent periodic patterns and statistically important patterns which are based on the frequency of occurrence. Frequent periodic patterns are in turn categorized as imperfect and perfect periodic patterns, partial and full periodic pattern, asynchronous and synchronous periodic patterns, approximate periodic patterns, dense periodic patterns. A pattern which happens periodically without any misalignment is known as a pattern of synchronous periodic. We have used the synchronous periodic data mining in our proposed because most of the festivals are coming in the same months.

III. PROPOSED WORK

We have proposed an algorithm which uses temporal association rule mining using the concepts of rough sets so as to optimize the results and also the periodical mining tells about the occurrences of the events in a synchronous or asynchronous mode. The whole proposed process is explained here with the help of a flowchart and pseudo code that represents the flow of the proposed algorithm.

The proposed algorithm contains the concept of periodical mining using rough sets. The whole proposed algorithm is explained below with the help of a flow chart. Also, a pseudo code has been written for the algorithm designed.

The flowchart can be shown as below:

A. Flowchart of Proposed Algorithm:

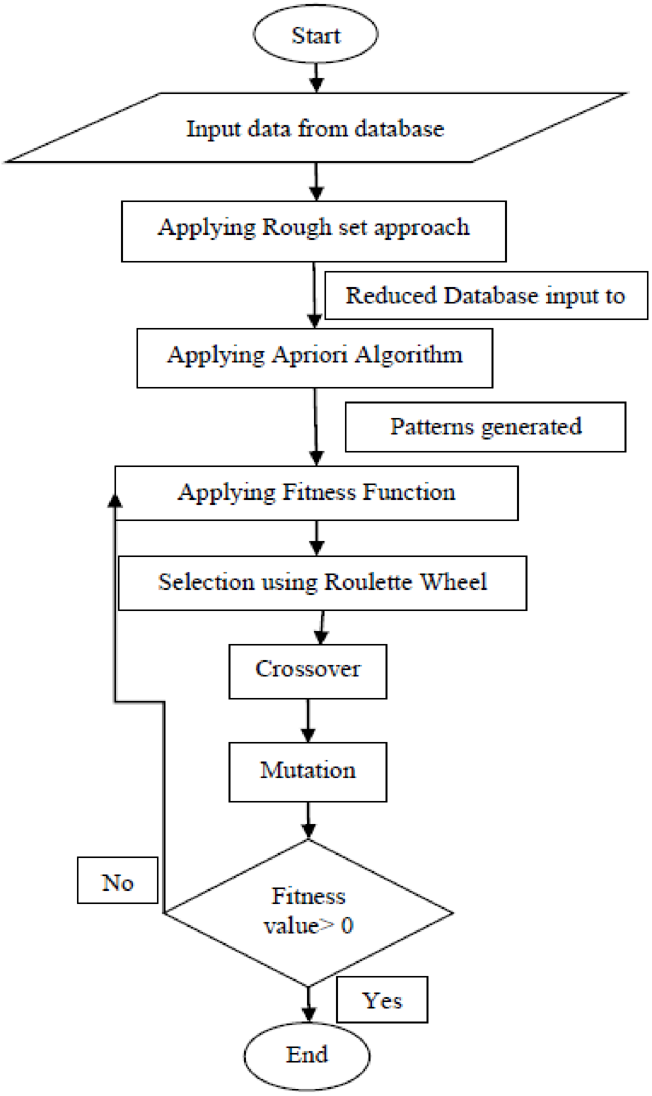

Start
Input data from database
Applying Rough set approach
Reduced Database input to
Applying Apriori Algorithm
Patterns generated
Applying Fitness Function
Selection using Roulette Wheel
Crossover
Mutation
No
Fitness value> 0
Yes
End

Fig.2: Flowchart of Proposed Algorithm

This flowchart explains the whole process of the proposed algorithm. New improved algorithm proposes the implementation of rough sets theory is used so as to reduce the number of unnecessary itemsets. The features of Genetic algorithm like crossover, fitness function, mutation and selection.

The above figure gives a clear idea about proposed algorithm. The pseudo code for this algorithm has been explained below:

B. Proposed Algorithm:

Input: Items from the database Db; crossover point cp;I: Set of n items, I = {i_1, i_2, … , i_n}, in the database; Roulette wheel selection rws; fitness value FitnV; cumulative fitness comfit; No. of chromosomes selected Nsel;

Output: Patterns generated from periodical mining

Algorithm:

```
1. Loading the data into the database.

2. Rough set theory implementation
        A,H,L values are set for the itemsets.
        % H denotes high, L denotes low and A
          denotes average values
        Then, LA and UA are find out
        Accuracy of approximation = (LA/UA) * 100
        %LA and UA are lower and upper
          approximation for the database table.
        Reduced database is created.

3. Apriori Algorithm
        Ck: Candidate itemset of size
        k Lk : frequent itemset of
        size k L1 = {frequent items};
         for ( k = 1; L k != ∅; k++) do
        begin Ck+1 = candidates
        generated from Lk;
            for each transaction t in database do
            increment the count of all candidates in
            Ck+1 that are contained in t
    Lk+1 = candidates in Ck+1 with
    min_support end

4. The data is calculated for the consecutive 10
   years, and periodical mining of the rules/ patterns
   is performed.

5. Genetic Algorithm
   5.1. Calculating fitness function using the formula:
        Fitness= $1 * + 2 * \quad + 3 * ( \quad - 1)$
   5.2. Applying Roulette Wheel Selection
                %Syntax:
            %NewChr = rws(FitnV, Nsel)

            %This function selects a given number of
        individuals Nsel from a population. FitnV is a
        column vector containing the fitness values of
        the individuals in the population. The new
        population, ready for mating, can be obtained
        by calculating OldChrom(NewChr, :).

              i.  function NewChr =
            rws(FitnV,Nsel); Identify the population
            size

            %Perform Stochastic Sampling
        with Replacement
             ii.  cumfit = cumsum(FitnV);

            iii.  trials = cumfit(Nind) .* rand(Nsel,
                  1);
             iv.  Mf = cumfit(:, ones(1, Nsel));
              v.  Mt = trials(:, ones(1, Nind))';

   5.3. Crossover is performed considering a single-
        point crossover.

            If( No._of_chr ==
                even) Cp = mid;
            Else
                Cp = mid + 1;

   5.4. Mutation is performed. Random single bit
        mutation is performed over the chromosome.

   5.5. Repeat steps 4.1 to 4.4 till
            Min_fitness > 0

6. End
```

IV. RESULT ANALYSIS

We have done experiments on Rough Set Theory For Mining Periodic Frequent Patterns (RSPFPM) and our proposed Improved Frequent Pattern mining (IFPM) algorithms.

Simulation tool: MATLAB R2013.

The dataset contains data for past 10 years based on the requirement of the algorithm. The database with all the values is stored in Microsoft Office Excel 2013.All experiments were performed well and fully on Dell workstation with 4 GB RAM and 32-bit operating system, running windows 7.

A. Result Obtained:

The Steps in the process of the implementation:

Step 1: When clicking on run, a GUI is displayed which is shown below:

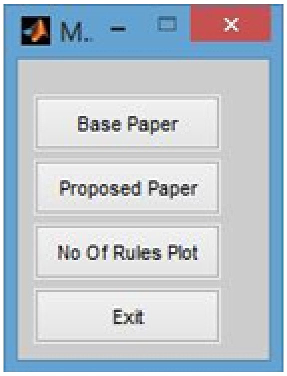
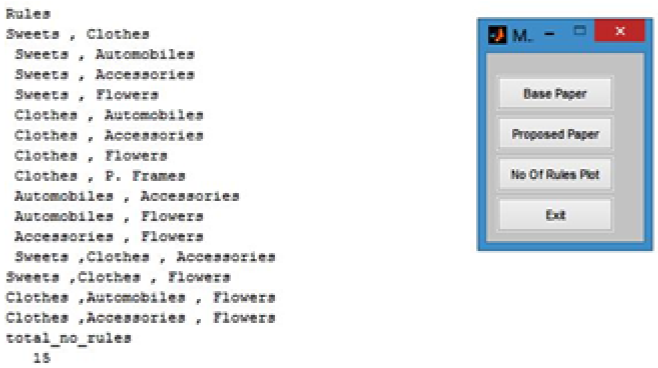


Fig.3: GUI

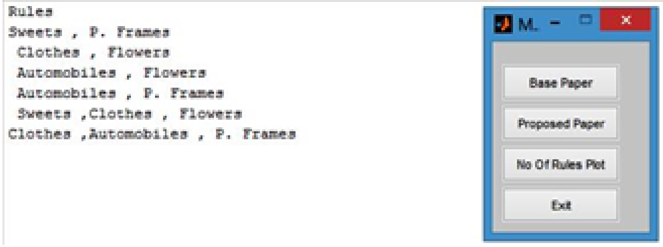
Step 2: On clicking the Base Paper button, the algorithm behind the base paper runs and gets its database from the excel sheet where the database has been created. The patterns are generated which are shown below:



```
Rules
Sweets , Clothes
 Sweets , Automobiles
 Sweets , Accessories
 Sweets , Flowers
 Clothes , Automobiles
 Clothes , Accessories
 Clothes , Flowers
 Clothes , P. Frames
 Automobiles , Accessories
 Automobiles , Flowers
 Accessories , Flowers
 Sweets ,Clothes , Accessories
Sweets ,Clothes , Flowers
Clothes ,Automobiles , Flowers
Clothes ,Accessories , Flowers
total_no_rules
    15
```

Fig. 4: Patterns generated from the base algorithm

Step 3:- In this step, we now click on second button i.e. Proposed paper. The new improved algorithm runs on clicking this button and generates the optimized patterns. The improved algorithm optimizes the patterns on the basis of the genetic algorithm implemented on the periodical data. The patterns generated are shown as below:-



```
Rules
Sweets , P. Frames
 Clothes , Flowers
 Automobiles , Flowers
 Automobiles , P. Frames
 Sweets ,Clothes , Flowers
Clothes ,Automobiles , P. Frames
```

Fig.5: Patterns generated from new proposed Algorithm

Step 4: The comparison between the numbers of patterns generated by both the algorithms i.e. Rough Set Theory for Mining Periodic Frequent Patterns (RSPFPM) and our proposed improved Frequent Pattern mining (IFPM) has been shown with the help of the graph. The graph can be displayed like this:-

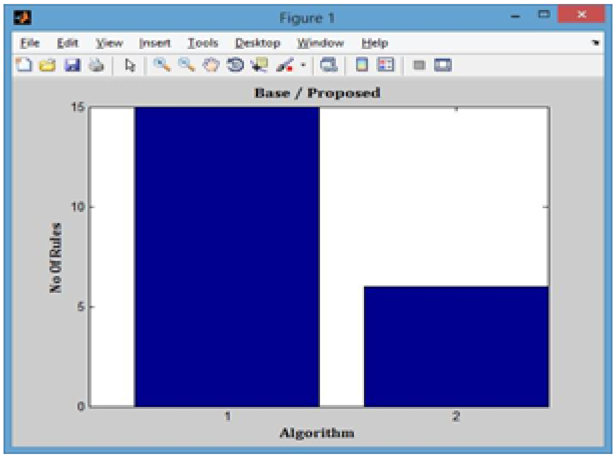


Fig.6: Comparison between the base and proposed algorithms

V. CONCLUSION AND FUTURE WORK

The improved algorithm has been proposed which has various measures that are better as compared to the previously designed algorithm. The improved algorithm uses the genetic algorithm which makes it optimized and the patterns generated are in optimized form. This paper proposes a new rough set based approach for extracting periodic frequent patterns. Due to the use of rough set based concept and genetic concept, proposed approach considers only the reducts of the initial database. Hence, it is clear that the proposed approach works on the reduced dataset which leads to the enhancement in the performance and generates optimized rules. So, based on this and the periodical mining algorithm, the common rules are generated for 10 years shows the frequent rules and thus give the idea about the festivals that have been occurring periodically in the past 10 years.

The future scenario for this improved algorithm can be that it may be implemented on real time scenarios as we have experimented on the synthetic dataset. Other than this, the other factors that can affect the optimized rule generations can be improved on CPU time utilization basis.

REFERENCES

[1] J. Han and M. Kamber, “Data Mining: Concepts and Techniques”, 2nd ed., The Morgan K\\aufmann Series in Data Management Systems, Jim Gray, Series Editor 2006.

[2] Rakesh Agrawal, SaktiGhosh, Tomasz Imielinski, BalaIyer, and Arun Swami, An Interval Classier for Database Mining Applications", VLDB-92, Vancouver, British Columbia, 1992, 560-573.

[3] J. Han and M. Kamber, Data Mining: Concepts and Techniques. Second edition Morgan Kaufmann Publishers.

[4] Dixit et al.," A Survey of Various Association Rule Mining Approaches" International Journal of Advanced Research in Computer Science and Software Engineering 4(3), March - 2014, pp. 651-655.

[5] Pradnya A. Shirsath, Vijay Kumar Verma, "A Recent Survey on Incremental Temporal Association Rule Mining", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-3, Issue-1, June 2013.

[6] Yiyu Yao, "Rough Set Approximations: A Concept Analysis Point of View", University of Regina, Regina, Saskatchewan, Canada, 2015.

[7] En-Bing Lin and Yu-Ru Syau, "Comparisons between Rough Set Based and Computational Applications in Data Mining", International Journal of Machine Learning and Computing, Vol. 4, No. 4, August 2014.

[8] Richa Garg and Saurabh mittal, "Optimization by Genetic Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering 4(4), April - 2014, pp. 587-589.

[9] Dewang Rupesh, Agarwal Jitendra. "A New Method for Generating All Positive and Negative Association Rules":International Journal on Computer Science and Engineering,(2011) ,Vol. 3, pp:1649-1657.

[10] Abdoljabbar asadi, Mehdi Afzali. "Providing a new method for detecting positive and negative optimal performance association rules in very large databases using Binary Particle Swarm Optimization": The sixth Iran Data Mining Conference / IDMC, Dec 01,01 / 2102, Tehran, Iran,(2102).

[11] Abdoljabbar Asadi, Mehdi Afzali, Azad Shojaei, Sadegh Sulaimani. "New Binary PSO based Method for finding best thresholds in association rule mining". Life Science Journal; 9(4).pp:260 - 264,(2012).

[12] David E Goldberg, Holland John H, "Genetic Algorithms in Search, Optimization and Machine Learning", Reading,MA:Addison-Wesley.

[13] Dewang Rupesh, Agarwal Jitendra. "A New Method for Generating All Positive and Negative Association Rules": International Journal on Computer Science and Engineering,(2011) ,Vol. 3, pp:1649-1657.

[14] D. E. Goldberg, Genetic algorithms in search, optimization, and machine learning, Addison Wesley Longman, Inc., ISBN 0-201- 15767-5, 1989.

[15] D. E. Golberg and K. Deb, "A comparative analysis of selection schemes used in genetic algorithms", Foundations of Genetic Algorithms, San Mateo, CA, Morgan Kaufmann, pp 69-93, 1991.

[16] K. A. De Jong, An Analysis of the behavior of a class of genetic adaptive systems (Doctoral dissertation, University of Michigan) Dissertation Abstracts International 36(10), 5140B University Microfilms No. 76/9381, 1975.

[17] Omari et.al. new temporal measure for association rule mining. Second International Conference on Knowledge Discovery and Data Mining, 1997.

[18] Chen Chu-xiang, Shen Jiang-jing, Chen Bing, et al., "An Improvement Apriori Arithmetic Based on Rough Set Theory", In proceeding of the 2011 Third Pacific-Asia Conference on Circuits, Communications and System (PACCS). pp.1-3, 2011.

[19] B. Ozden, S. Ramaswamy, and A. Silberschatz. Cyclic association rules. In Proc. of the 14th International Conference on Data Engineering,, pages 412–421, 1998.

[20] J. Han, G. Dong, and Y. Yin. Efficient mining partial periodic patterns in time series database. In Proc. of the15th International Conference on Data Engineering, pages 106–115, 1999.

[21] E. Simoudis, J. W. Han, and U. Fayyad, editors. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96).AAAI Press, 1996.

AUTHORS

Manjeet Samoliya is currently pursuing the M.tech degree in computer science and engineering from CSE/IT deptt., MITS Gwalior (M.P), India. He has received B.E. degree from Gwalior Institute of Technology and Science, Gwalior (M.P), India. His areas of interest are data mining, Association rule mining, frequent pattern mining, rough set and their applications.

Dr. Akhlesh Tiwari has received the Ph.D. degree in Information Technology from Rajiv Gandhi Technological University, Bhopal, M.P. (India). He is currently working as Associate Professor in the Department of CSE & IT, Madhav Institute of Technology & Science (MITS), Gwalior, India. He has guided several students at Master and Under Graduate level. His areas of current research include Knowledge Discovery in Databases and Data Mining, Wireless Networks. He has published more than 20 research papers in the journals and conferences of international repute. He is also working as a reviewer & member in the editorial board of various international journals. He is having the memberships of various Academic/ Scientific societies including IETE, CSI, GAMS, IACSIT and IAENG