ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2017.0221

ENHANCED ALGORITHMS FOR MINING OPTIMIZED POSITIVE AND NEGATIVE ASSOCIATION RULE FROM CANCER DATASET

I. Berin Jeba Jingle[1] and J. Jeya A.Celin[2]

[1]*Department of Computer Science Engineering, Noorul Islam University, India*

[2]*Department of Computer Applications, Hindustan College of Arts and Science, India*

Abstract

The most important research aspect nowadays is the data. Association rule mining is vital mining used in data which mines many eventual informations and associations from enormous databases. Recently researchers focus many research challenges to association rule mining. The first challenge is the generation of the frequent and infrequent itemsets from a large dataset more accurately. Secondly how effectively the positive and negative association rule can be mined from both the frequent and infrequent itemsets with high confidence, good quality, and high comprehensibility with reduced time. Predominantly in existing algorithms the infrequent itemsets is not taken into account or rejected. In recent times it is said that useful information are hidden in this itemsets in the case of medical field. The third challenge are to generate is optimised positive and negative association rule. Several existing algorithms have been implemented in order to assure these challenges but many such algorithms produces data losses, lack of efficiency and accuracy which also results in redundant rules. The major issue in using this analytic optimizing method are specifying the activist initialization limit was the quality of the association rule relays on. The proposed work has three methods which mine an optimized PAR and NAR. The first method is the Apriori_AMLMS (Accurate multi-level minimum support) this algorithm derives the frequent and the infrequent itemsets very accurately based on the user-defined threshold minimum support value. The next method is the GPNAR (Generating Positive and Negative Association Rule) algorithm to mine the PAR and NAR from frequent itemsets and PAR and NAR from infrequent itemsets. The third method are to obtain an optimized PAR and NAR using the decidedly efficient swarm intelligence algorithm called the Advance ABC (Artificial Bee Colony) algorithm which proves that an efficient optimized Positive and negative rule can be mined. The Advance ABC is a Meta heuristic technique stimulated through the natural food foraging behaviour of the honey bee creature. The experimental analysis shows that the proposed algorithm can mine exceedingly high confidence non redundant positive and negative association rule with less time.

Keywords:

Data Mining, Association Rule Mining, Apriori Algorithm, Accurate Multi Level and Multi Support, Advance ABC Algorithm, GPNAR

1. INTRODUCTION

Data mining is incredibly essential field in the case of knowledge discovery. Data mining deals with the process of mining unseen projecting information from massive databases. Recent technology helps medical, market, weather forecasting analysis to focus requisite information in their data warehouse. The data mining mines data's from different sources like image, text, and web. The mentioned data from various sources are nowadays exceptionally significant analyzing, which plays immense responsibility in shifting these data into valuable information and patterns. The data mining has several techniques like association rule, classification, decision tree, clustering, prediction, etc. for extracting the valuable information from this data sources. Commonly these procedures are mainly designed for the motivation of developing competent mining algorithm in order to extract patterns within reasonable and adequate time frame.

The association rule (AR) mining is a technique of data mining which is used to analyze high-dimensional relational data. The association rule mining discovers interesting relationship hidden in a large dataset. The association rule techniques are implemented effectively in application domain such as health informatics, image classification, network traffic analysis, market basket analysis, intrusion detection, telecommunication and diagnosis decision support. The major focal point of the association rule mining is the research community.

Usually the intention of association rule mining is extracting the frequent itemsets from the transactional database based on the user defined threshold value and the infrequent itemsets is ignored. But it is observed that the itemsets which has low support infrequent itemsets can also produce prospective significant negative association rule. The proposed work proves that from this infrequent itemsets both positive and negative association rule can be mined more effectively of the form $X\rightarrow\neg Y$, $\neg X\rightarrow Y$, $\neg X\rightarrow\neg Y$.

The proposed work chooses the medical dataset for the experimental analysis. The proposed system intention is 1) generate Frequent and infrequent itemsets very accurately, 2) mining both PAR and NAR from the frequent and infrequent itemsets, 3) Generate an optimized PAR and NAR.

The association rule is mainly extracted from transaction databases but, in our proposed system our aim is to develop a high quality and high confidence association rule by using various measures. Then optimized algorithm is applied in the generated rule to remove all the inefficiency from it and mould it into unsurpassed one. Hence the advance ABC algorithm is used.

The advance ABC algorithm that impersonators the astonishing food foraging behavior of real honey bees contributes three key constraints: The first is a population that is considered as a number of food sources, the second constraint is a limit that is the number of tries following which a food source is rejected, and the third constraint is the norm to stop the process also known as maximum number of cycles.

2. RELATED WORKS

Idheba et al [21] suggested PNAR and IMLMS an approach for mining positive and negative association rule from transaction dataset. This approach is integrated by two algorithms. The positive negative association rule (PNAR) algorithm and the

Intersting multiple level minimum support (IMLMS) algorithm and the approach is PNAR_IMLMS. The IMLMS algorithm generates the frequent and infrequent itemsets. PNAR algorithm generates positive and negative association rule from the generated frequent and infrequent itemsets. Significantly better than the previous methodologies but lack in efficiency and accuracy also a time consuming process.

Dong [22] proposes an enhanced Aprior-IMLMS (interesting MLMS (Multiple Level Minimum Supports)) algorithm, which is designed for pruning uninteresting infrequent and frequent itemsets discovered by MLMS model. One of the pruning measures used in IMLMS model, interest, can be described as follows: to two disjoint itemsets A, B, if interest $(A,B) = |s(A \cup B) - s(A)s(B)| < m_i$, then $A \cup B$ is recognized as uninteresting itemsets and is pruned, where $s(.)$ is the support and mi a minimum interestingness threshold. This measure, however, is a bit difficult for users to set the value mi because interest (A,B) highly depends on the values of $s(.)$. This paper proposes a new measure, MCS (minimum correlation strength) as a substitute. MCS, which is based on correlation coefficient, has better performance than interest and it is very easy for users to set its value. The theoretical analysis and experimental results show the validity of the new measure.

Niu et al. [23] proposed a PNAR_MLMS algorithm to mine infrequent itemsets. Previous work, a MLMS model was proposed to discover simultaneously both frequent and infrequent itemsets by using multiple level minimum sup- ports (MLMS) model. In this paper, combines correlation coefficient and minimum confidence is proposed and a corresponding algorithm PNAR_MLMS is also proposed to generate PNARs correctly from the frequent and infrequent itemsets discovered by the MLMS model. The experimental results show that the measure and the algorithm are effective.

Soltan et al. [2] proposed an algorithm CARM (Confabulation-inspired Association rule mining) which generates frequent and infrequent itemset. The main achievement of knowledge of this model is finding association link between attainment and rule extraction. The extracted rule is then executed by deriving the weight age of these communication links this is done in the second phase. Li-Min Tsaiet al [3] proposed the GNAR (Generalized Negative Association Rule) which is an improved approach algorithm which shows negative rules are as imperative as positive rules. It helps user to make quick decision to analyse which is the best association rule. The advantage of this algorithm is cost and time reduction but lack in accuracy and efficiency.

Hidlern et al. [4] introduced an algorithm continuous association rule mining algorithm (CARMA) which compares online large datasets by two phase scanning of transaction in large itemset. This algorithm consist of 2 pass, the first pass continuously construct a lattice potentially for large itemsets. The second pass continuously removes itemsets which has less user specified threshold value. Xiangjun et al. [5] suggested another algorithm IMLMS model which generated the frequent and infrequent itemsets using the minimum correlation coefficient. This algorithm defines that the interestingness of the rule is high and the minimum support value is easy to set. But lack in accuracy in the generation of the frequent and infrequent itemsets.

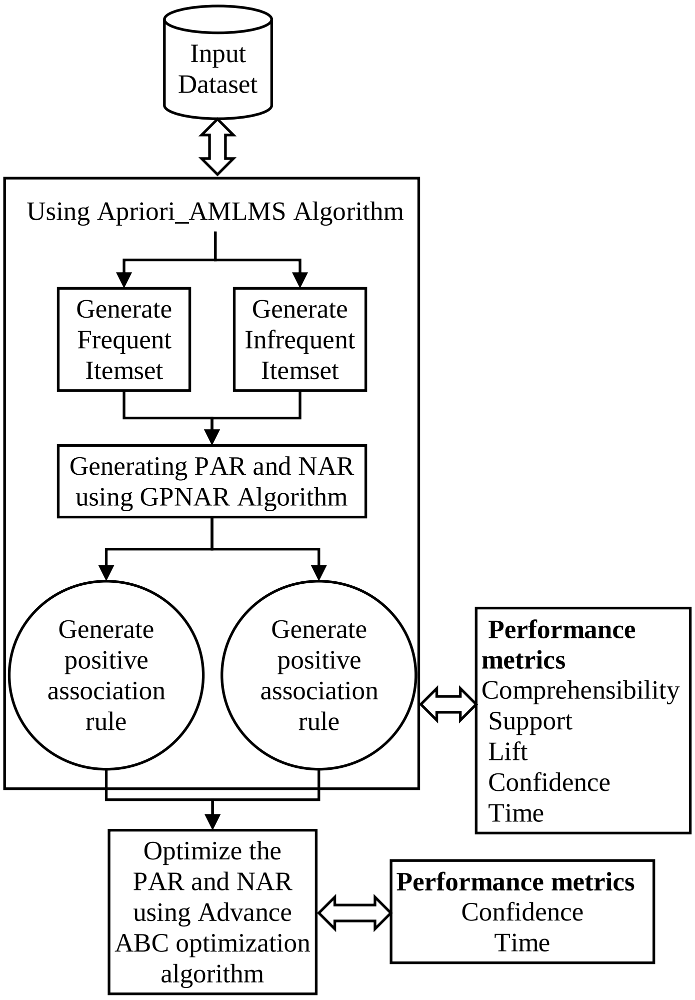

Fig.1. Overall Proposed architecture

The MLMS algorithm mines the FIS and inFIS itemsets, the discovered patterns are not much interesting and are raucous and hence it requires pruning. So the existing method used the modified wu's pruning strategy with IMLMS [7] an algorithm was designed to discover interesting frequent and infrequent patterns. This method has some difficulty to set the threshold value. It uses interest measure to calculate the interestingness of the itemsets the interest (A,B) depends on values of support $s(.)$. This method mainly prunes uninteresting patterns. The discovered patterns lack in efficiency and accuracy. The next existing method rectifies the measure interest and uses another measure Minimum correlation Strength (MCS) [4] based on correlation coefficient the performance is better than the measure interest here the users finds easy to set the values here $\rho(A,B)$ is calculated instead of interest(A,B). However the performance improves but this method still lack in accuracy and efficiency. The AMLMS-GA [18] generates the accurate frequent and infrequent itemsets and latter mines Positive and negative association rule from the generated Frequent and infrequent itemsets. Then the optimizd rule is generated based on the Genetic Algorithm (GA) which classifies the generated itemsets based on their relevancy. The optimized rule is not properly generated and lack in efficiency. The proposed Apriori_AMLMS [20] algorithm is proposed to improve efficiency of the optimized rule. The Modified Genetic Algorithm (MGA) is proposed to improve the optimized algorithm.

3. THE PROPOSED SYSTEM

The research proposes three algorithms in order to generate the optimized Positive and negative association which is eventually applied in area of association rule mining. The proposed work chooses the UCI machine medical dataset for the analysis of the proposed algorithm. There are three phases 1) apriori_AMLMS algorithm 2) GPNAR algorithm 3) Advance ABC algorithm.

3.1 THE PROPOSED ARCHITECTURE

The architecture shows the proposed work (shown in Fig.1) has three contributions. The first contribution generates the accurate frequent and infrequent itemsets based on the user defined threshold value. The second contribution generates the Positive and the negative association rule from the generated frequent and the infrequent itemsets. The third contribution generates an optimized positive and negative association rule.

3.2 APRIORI_AMLMS ALGORITHM

The proposed Apriori_AMLMS [19][20] algorithm uses the user defined minimum support threshold value for generating the frequent and infrequent itemsets. The transaction medical dataset is first transformed into the decision Table.in the preprocessing step. The data's are arranged in the decision table, which contains the conditional and decisional attributes. The attributes are compared with the neighbour attributes and arranged in priority wise hierarchal manner.

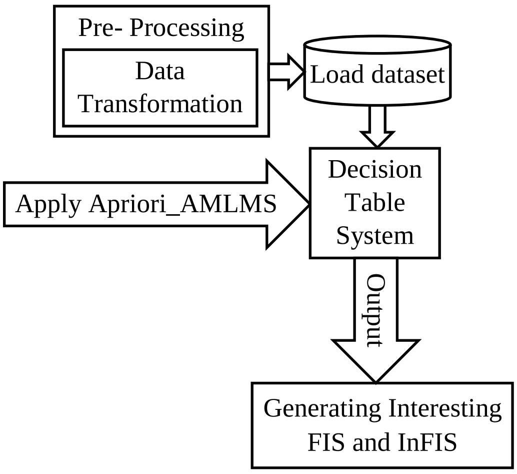

Fig.2. Apriori_AMLMS Architecture

The input of the proposed algorithm is the dataset and the user defined threshold value. The output generated is the Frequent and the infrequent itemsets. For each itemsets the support is calculated. If suppo(X) ≥ minisuppo (number(X)),then the X is frequent itemsets. If suppo(X) < minisuppo (number(X)), then the X is infrequent itemsets. The generated frequent and infrequent itemsets are collected in the hash table. The hash map stores each data with index and value. The index value denotes weather the itemsets is frequent or infrequent itemsets. The generated FIS and inFIS generated through this algorithm is very accurate.

3.3 GPNAR ALGORITHM

The second contribution is the GPNAR [20] (Generating Positive and Negative Association Rule). This algorithm mines the Positive association rule (PAR) and negative association rule (NAR) based on the user defined minimum confidence (miniconfi) threshold value.

Suppo $(\sim X)$ = 1-suppo(X)

Suppo $(X \rightarrow Y)$ = P$(X \cup Y)$

Suppo $(\sim X \rightarrow Y)$ = Suppo(X)-P$(X \cup Y)$

Suppo$(X \rightarrow \sim Y)$ = Suppo(Y)-P $(X \cup Y)$

Suppo $(\sim X \rightarrow \sim Y)$ = (((1-Suppo(X) - (1-Suppo(Y)) + P$(X \cup Y)$)

Confi $(X \rightarrow Y)$ = Suppo$(X \cup Y)$ / Suppo(X)

Confi $(X \rightarrow \sim Y)$ =1-(Suppo$(X \cup Y)$ / Suppo(X))

Confi $(\sim X \rightarrow \sim Y)$ = Confi$(\sim X \rightarrow Y)$ / Suppo$(\sim X)$

Lift$(X \rightarrow Y)$ = ((suppo$(X \cup Y)/N$) / (suppo(X)/N × suppo$(Y)/N$) > 1,

$$Comprehensibility = \frac{\log\left(1+|Y|\right)}{\log\left(1+|X\grave{E}Y|\right)} \qquad (1)$$

The comprehensibility describes the clarity of rule. The main measurement of the association rule mining is the confidence, support, correlation, comprehensibility and time. The correlation gives the interestingness of the Positive and Negative association rule. Consider let Suppo$(X \rightarrow Y)$, Suppo(X), Suppo$(Y) \geq$ minisuppo and Lift$(X \rightarrow Y)$ >1

For experimental analysis the UCI machine cancer dataset is chosen which has 18 classes with 22 attributes and 1582 instances. This a primary tumor domain which was obtained from the University Medical Centre.

Suppo (tumor) =0.3,

suppo (¬tumor) =0.7

Suppo (cancer) =0.7,

suppo (¬cancer) =0.3

Suppo (cancer ∪ tumor) =0.056

Minisuppo =0.3

Miniconfi =0.7

From the above following recognized values the PAR and the NAR can be identified for the itemsets.

If Suppo (cancer ∪ tumor) =0.056 < minisuppo(0.3) then this items are said to be infrequent item set.

And if Confi (tumor ∪ cancer) =0.244 < miniconfi(0.7), hence the positive rule has less confidence.

So the negative rule is derived. Consider for example.

Suppo (tumor ∪ ¬cancer) = suppo (tumor) – suppo (cancer ∪ tumor)

= 0.3-0.056

= 0.244 > miniconfi

Confi (tumor ⇒ ¬cancer) = (Suppo (cancer ∪¬tumor)) / suppo (tumor)

= 0.244/0.3= 0.8 > miniconfi,

Hence (tumor ⇒ ¬cancer) is a negative rule.

Lift (tumor ⇒ ¬cancer) = (Suppo(tumor∪¬cancer))/(Suppo(tumor)suppo(¬cancer))

= 0.244/(0.3×0.3)

= 2.244 >1,

in strong negative rule and it has a valid 87% of which tells strong presence of act and absence of cancer

3.4 ADVANCE ABC ALGORITHM

The Advance ABC algorithm is the third contribution in the case of the proposed system. The main goal is to extract an optimized PAR and NAR. The ABC algorithm is a recent swarm intelligence ABC algorithm is categorized into four phases. 1) initialization Process 2) employee Bee 3) onlooker bee 4) scout bee. Here the extracted PAR and NAR is given as input the. The rules are taken as food source. Steps involved in Advance ABC algorithm.

Input: PAR and NAR, minicofi.

Output: Optimized PAR and NAR

Step 1: Initialize population using ABC on selected members to discover associations

Step 2: Find each association rule fitness function.

Step 3: Check following condition: If (fitness function> miniconfi).

Step 4: Set $Q = Q \cup (x => y)$ /* the rules are added to the temporary variable Q */

- 5.1 In memory, employed bees are placed on food sources;
- 5.2 Generate new offspring from older offspring after applying onlooker bee phase.
- 5.3 For finding new food sources, send scout bee to search space.

Step 5: Until (requirements are not met).

The proposed advance ABC algorithm is given in the Fig.3. The process is initialized first. The food sources are the rule.

3.4.1 Initialization Phase:

The beginning process is the initialization process. The positions of 3 food sources (CS/2) of employed bees is initialized first, (50, 500) are the range of uniform distribution and they are randomly utilized.

$$y_j^k = y_{\min}^k + rand[0,1] \times \left(y_{\max}^k - y_{\min}^k \right) \tag{2}$$

where, $y_{\min}^k$ the upper is bound for y_j^k and $y_{\min}^k$ is the lower bound for y_j^k, where j = 1,2,...,M and k = 1,2,...,D. y_j^k is a restriction to be optimized for the j^{th} employed bee on the dimension k of the D-dimensional space. Number of employed bee denoted as M.

3.4.2 Employee Bee Phase:

Next the objective function of each rule is defined. The object (obj) value is based on support, comprehensibility and confidence of each rule.

$$\text{ObjVal} = (\text{suppo.*confi}).\text{*comp} \tag{3}$$

The fitness value depends on object function. The fitness function is calculated using

$$F(j) = \begin{cases} \dfrac{1}{\left(1+obj(i)\right)} & if \left(obj(i)\right) \geq 0 \end{cases} \tag{4}$$

The obj value is rule based on the support, confidence, and comprehensibility. This obj value is considered as the food. In order to calculate fitness value number of iteration is used by comparing with the neighbouring food sources. Until the best optimized value is obtained the iteration goes on. The best fitness value rules are stored in a new memory space. The new position is given as $v_j^k = y_j^k + \varphi_j^k \left(y_j^k - y_i^k \right)$ Where, k = 1,2,...,D and i = 1,2,...,M. In the above equation, y_j^k is the j^{th} employed bee, v_j^k is the new solution for y_j^k, y_i^k is the neighbour bee of y_j^k in employed bee population, [-1, 1] is the range of φ and it is randomly selected, from the above equation the k and i are selected and memorized as best solution.

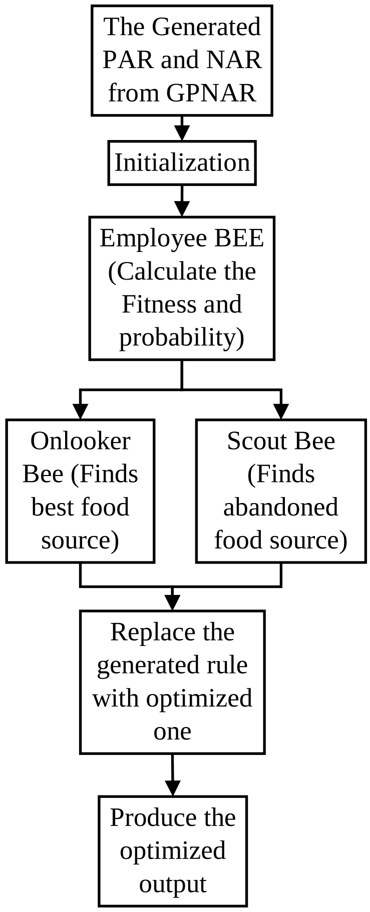

Fig.3. Architecture of Advance ABC Algorithm

3.4.3 Onlooker Bee Phase:

This comes under unemployed honey bee. The employee bee completes the phase by calculating the optimized fitness value. The onlooker bee probabilistically picks the food source relying on the data. In advance ABC, an onlooker honey bee picks a food source communicate upon probability values figured utilizing fitness values gave by utilized honey bees. For this reason, a fitness based fortitude strategy can be utilized, the algorithm uses the roulette wheel selection technique for this purpose. The probabilistic is given as $F(j)$, which denotes fitness value, SN denotes swarm size.

3.4.4 Scout Bee Phase:

The scout bee comes under the unemployed bee these bees pick their foods randomly. This unemployed bee searches new food sources randomly depending on an internal motivation or promising external clues. For instance, if solution y_j^k has been abandoned, the new solution discovered by the scout who was the employed bee of y_j^k can be defined. The rules which are initially poor or made poor by exploitations are abandoned by the scout bees and initialize the next set of rules to be optimized.

4. EXPERIMENT RESULT

The proposed methods choose the UCI machine cancer dataset for experimental and performance analysis. This dataset has 18 classes and 22 attributes and 1582 instances and 20 transactions. The experimental analysis first shows how the frequent and infrequent itemset are generated by the support (minsuppo) value as input for the dataset DS. Then from the FIS and inFIS how the PAR and NAR is mined and from this NAR and PAR using optimized ABC algorithm how the optimized rule is mined from various minisuppo.

Table.1. Accurate FIS and inFIS

Minisuppo	Frequent Itemset	Infrequent Itemset
0.2	479	521
0.5	431	569
0.7	372	628
1	254	746

From the experimental result it is clearly seen in Table.1 that there is a gradual decrease in the frequent itemset set as the user defined threshold value minisuppo is increased. This shows that the candidate itemset generation is decreased and database scanning is also decreased. Hence automatically the space is also decreased. The time to generate the FIS and inFIS is also reduced using the proposed Apriori_AMLMS algorithm when compared with the existing algorithms. The generated PAR and NAR from frequent and infrequent itemset are of the form

Table.2. Generated Rules for PAR and NAR

Rule Support Confidence Comprehensibility Lift
{tumor, bone}->{treatement} 25% 85.5% 0.6309 1.432
{chemo,radiation}->{tumor} 40% 93% 0.7253 1.342
{bone,tumor}->{treatement} 40% 95.3% 0.6342 2.54
{tumor,CT scan, brain}->{cancer} 65% 92.5% 0.6309 1.43

PAR from Frequent Itemset

Rule Support Confidence Comprehensibility Lift
{tumor, bone}->{¬cancer} 45% 94% 0.6459 1.432
{tumor, radiation}->{¬temodar} 33% 92% 0.723 1.342
{tumor, brain}->{¬glioblastoma} 40% 92.5% 0.683 2.54
{¬tumor, brain }->{¬cancer} 35% 89.5% 0.6309 1.43

NAR from Frequent Itemset

Rule Support Confidence Comprehensibility Lift
{esophasus, tumor}->{cancer} 15% 85.5% 0.6309 2.523
{tumor, surgery, breast}->{cancer} 14% 94.3% 0.652 1.58
{CTscan, brain}->{treatment} 15% 96.5% 0.573 1.87
{tumor, brain,mri }->{cancer} 27% 94.5% 0.723 1.76

PAR from inFrequent Itemset

Rule Support Confidence Comprehensibility Lift
{chemo, treatment}->{¬surgery} 35% 95.5% 0.754 1.623
{treatment, brain }->{¬glioblastoma} 30% 90.3% 0.634 1.512
{chemo,¬treatement}->{¬glioblastoma} 35% 93.5% 0.672 2
{ glioblastoma }->{¬radiation} 31% 92.5% 0.743 2.87

NAR from inFrequent Itemset

The Table.2 has four different sets of tables, here the rules generated from the proposed methods are explicitly shown. The first two tables are the set of PAR and NAR generated from the frequent item sets and the last two sets of tables show the PAR and NAR generated from the infrequent item sets with various measures like support, confidence, and comprehensibility.

Table.3. Overall result analysis of the proposed algorithm

Minisuppo		0.2	0.5	0.7	1
Miniconfi		0.6	0.65	0.7	0.8
FIS	PAR	225	229	161	113
	NAR	254	202	211	132
inFIS	PAR	241	254	283	253
	NAR	280	315	345	493
Comprehensibility		0.6309	0.6732	0.5342	0.6309
Confidence		0.85	0.95	0.92	0.85
Lift		1.5	1.3	1.4	2
Support		0.3	0.25	0.3	0.25

The rules generated by the proposed algorithm are very huge and many rules are redundant which is clearly shown in Table.3 so to achieve optimized high confidence rule, the optimized Advance ABC algorithm is applied. The result analysis shown in Table.4 proves that the achieved rule is well optimized rule. The optimized rule is based on the fitness or optimized value of each rule after several iteration.

Table.4. Optimized PAR and NAR

Miniconfi	Fitness value	Confidence	Apriori_AMLMS and GPNAR		Adv.ABC	
			PAR	NAR	PAR	NAR
0.6	0.6983	0.7	243	655	132	455
0.65	0.6943	0.75	162	432	102	301
0.7	0.7412	0.9	223	563	84	183
0.8	0.8453	0.95	212	435	56	230

5. CONCLUSION

This paper concentrates on three proposed contributions. The research was made by analyzing the exiting algorithms. The first contribution generates the frequent and the infrequent itemsets more accurately in less time. The second method mines the PAR and NAR from the frequent and infrequent itemsets. The third method is the Advance ABC algorithm which is used as optimized algorithm which generates optimized PAR and NAR. The generated rules has high confidence, more support, high comprehensible, and the quality is good compared with the existing algorithm. The proposed algorithms mine rules for the cancer dataset very accurately. Any hidden useful data are mined though these contributions. The experimental analysis shows the proposed algorithm is promising and efficient

REFERENCES

[1] Nikky Suryawanshi Rai, Susheel Jain and Anurag Jain, “Mining Interesting Positive and Negative Association Rule Based on Improved Genetic Algorithm”, *International Journal of Advanced Computer Science and Applications*, Vol. 5, No. 1, pp. 160-165, 2014.

[2] Azadeh Soltani and M.R. Akbarzadeh, “Confabulation-Inspired Association Rule Mining for Rare and Frequent Itemsets”, *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 25, No. 11, pp. 2053-2064, 2014.

[3] Li-Min Tsai, Shu-Jing Lin and Don-Lin Yang, “Effective Mining of Generalized Negative Association Rules”, *Proceedings of IEEE Conference on Granular Computing*, pp. 113-117, 2010.

[4] Christian Hidber, “Online Association rule mining”, *Proceedings of International Conference on Management of Data*, pp. 145-156, 1999.

[5] Li Min Tsai, Shu Jing Lin and Don Lin Yang, “Effective Mining of Generalized Negative Association Rules”, *Proceedings of International Conference on Granular Computing*, pp. 163-167, 2010.

[6] Xiangjun Dong, Zhendong Niu, Donghua Zhu, Zhiyun Zheng and Qiuting Jia, “Mining Interesting Infrequent and Frequent Itemset based on MLMs Model”, *International Conference on Advanced Data Mining and Applications*, pp. 444-451, 2008.

[7] Xiangjun Dong, “*Ming Interesting Infrequent and Frequent Itemset Based on Minimum Correlation Strength*”, Springer, 2011.

[8] K. Mythili and K. Yasodha, “A Pattern Taxonomy Model with New Pattern Discovery Model for Text Mining”, *International Journal of Science and Applied Information Technology*, Vol. 1, No. 3, pp. 115-119, 2012.

[9] Charushila Kadu, Praveen Bhanodia and Pritesh Jain, “Hybrid Approach to Improve Pattern Discovery in Text Mining”, *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 2, No. 6, pp. 2477-2481, 2013.

[10] Zhen Hai, Kuiyu Chang, Jung-Jae Kim and Christopher C. Yang, “Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance”, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 6, No. 6, pp. 623-634, 2012.

[11] Spyros I. Zoumpoulis, Michail Vlachos, Nikolaos M. Freris and Claudio Lucchese, “Right-Protected Data Publishing with Provable Distance-based Mining”, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 19, pp. 2014-2028, 2012.

[12] K. Aas and L. Eikvil, “Text Categorisation: A Survey”, Technical Report, Norwegian Computing Center, 1999.

[13] R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases”, *Proceedings of 20th International Conference on Very Large Data Bases*, 1994, pp. 478-499, 1994.

[14] Dipti S. Charjan and Mukesh A. Pund, “Pattern Discovery For Text Mining using Pattern Taxonomy”, *International Journal of Engineering Trends and Technology*, Vol. 4, No. 10, pp. 22-26, 2013.

[15] J. Han and K.C.C. Chang, “Data Mining for Web Intelligence”, *Computer*, Vol. 35, No. 11, pp. 64-70, 2002.

[16] J. Han, J. Pei and Y. Yin, “Mining Frequent Patterns without Candidate Generation”, *Proceedings of ACM SIGMOD International Conference on Management of Data*, pp. 1-12, 2000.

[17] Rupesh Dewang and Jitendra Agarwal, “A New Methods for Generating All Positive and Negative Association Rules”, *International Journal on Computer Science and Engineering*, Vol. 3, No. 4, pp. 1649-1657, 2011.

[18] I. Berin Jeba Jingle and J. Jeya A. Celin, “Markov Model in Discovering Knowledge in Text Mining”, *Journal of Theoretical and Applied Information Technology*, Vol. 70, No. 3, pp. 459-463, 2014.

[19] I. Berin Jeba Jingle and J. Jeya A. Celin, “Discovering Useful Patterns in Text Mining using AMLMS-GA Algorithm”, *International Journal of Applied Engineering Research*, Vol. 10, No.18, pp. 39763-39767, 2015.

[20] Idheba Mohamad Ali O. Swesi, Azuraliza Abu Bakar and Anis Suhailis Abdul Kadir, “Mining Positive and Negative Association Rules from Interesting Frequent and Infrequent Itemsets”, *Proceedings of IEEE Conference Publication*, pp. 650-655, 2012.

[21] Xiangjun Dong, Zhendong Niu, Xuelin Shi, Xiaodan Zhang and Donghua Zhu, “Mining Both Positive and Negative Association Rules from Frequent and Infrequent Itemsets”, *International Conference on Advanced Data Mining and Applications*, pp. 122-133, 2007.

[22] X. Xing, Yao Chen and Yan-En Wang, “Study on Mining Theories of Association Rules and its Application”, *Proceedings of International Conference on Information Technology and Ocean Engineering*, pp. 343-347,

[23] I. Berin Jeba Jingle and J. Jeya A. Celin, “Mining Optimized Positive and Negative Association Rule using Advance ABC Algorithm”, *Journal of Theoretical and Applied Information Technology*, Vol. 95, No. 24, pp. 6846-6855, 2017.