IJCTA, 10(03), 2017, pp. 81-89

Mining Useful Patterns from Text using Apriori_AMLMS-MGA Algorithm

J. Jeya A. Celin* and I. BerinJeba Jingle**

Abstract : Association Rule mining is one of the vital mining used in data mining which extract many prospective information and associations from large amount of databases. Many different existing methodologies are used in the case of Association rule mining for generating positive association rule from frequent item set and for generating negative association rule form infrequent association rule which result in lack of efficiency and accuracy. Associations rule is one of the momentous research fields which very much used in discovering frequent and infrequent dataset in text documents. In the Proposed work, more importance is given in discovering positive association rule from infrequent itemset and negative association rule from frequent association rule. Hence we use the Apriori_AMLMS-MGA algorithm for generating the frequent itemset and infrequent itemset because apriori alone can generate only the frequent itemset. When the apriori is combined with the AMLMS (Accurate Multi-Level and Multiple Support) both the frequent and infrequent item set are generated and from these itemset all possible positive and negative association rules are generated. The rules thus generated are more valuable and accurate and has high efficiency. The AMLMS algorithm mainly used to prune the discovered datasets and reduces the number of rules and increases the accuracy. The modified genetic algorithm (MGA) is the next optimization technique used to gear wide-ranging collection of optimization problem. The experimental result using these proposed techniques produces a valuable negative association rules and the rules generated are very accurate and does not lack in efficiency.
Keywords : Data mining, Text mining, Association Rule, Apriori algorithm, Accurate Multi level and Multi support, Modified Genetic Algorithm.

1. INTRODUCTION

Data mining is a very imperative field in the case of science. The process of data mining is mining of unseen prognostic information from huge databases. The data mining is recent technology which helps business management and market analysis to spotlight the indispensable information in their data warehouse. The data's are extracted from different sources such as Text, Image, and Web etc. Analyzing such data are very important process nowadays the knowledge discovery and the data mining has plays immense responsibility in changing these data into valuable information and patterns.There are many techniques available in data mining for extracting data from various data sources such as association rule, classification, decision tree, clustering, prediction, etc. Generally these techniques are mainly planned for the rationale of developing efficient mining algorithm in order to extract patterns within sensible and tolerable time frame.

The association rule is mainly extracted from transaction databases but, in our proposed system we used the competent apparatus for identifying the positive and negative association rule between medications, symptoms, laboratory results by this data mining technology. The process of extracting high-quality information from text data is text mining it is achieved through the statistical pattern learning and

* Assistant Professor, Department of Computer Application, Hindustan College of Arts and Science, Chennai, India. *E-mail : jjeyacelin@gmail.com*

** Assistant Professor, Department of CSE Noorul Islam University Nagercoil, India. *E-mail: berinjeba@gmail.com*

machine learning. Usually text documents are unstructured, noisy, formless, and difficult to covenant with algorithmically. Text mining also leads to learn structural element of text in order to find invisible useful text from the large text documents. Many existing techniques and methods are used in text mining in order mine efficient and accurate patterns or information from large databases. The method used is pattern taxonomy model (PTM) [1] it is a pattern based approach which has two process like pattern evolving and pattern deploying methods. This methods rectifies the problem like misinterpretation problem and low frequency problem but, the pattern discovered lacks in quality, efficiency and accuracy.

The other method used for discovering patterns is the PTM and Naive Bayes classifier [2]. This method mainly solves the major problem of misinterpretation and low-frequency problem. This method leads complexity of the system hence direct to slow operation. The discovered patterns experience lack of performance problem. The novel Pattern mining approach [3] is the existing methodology used for pattern discovery which mines patterns from positive and negative responses next classifies the patterns in order to remove noise data at last the novel pattern deploying approach is used for improving the performance of the frequent patterns in text documents. This approach is a time consuming process with low speed. The MLMS discovers patterns, the discovered patterns are not interesting and are noisy and hence it requires pruning. So the existing method used the modified wu's pruning strategy with IMLMS [5] an algorithm was designed to discover interesting frequent and infrequent patterns. This method uses a measureinterestmito discover pattern, because *interest(A,B)* depends on values of support *s(.)* which is bit difficult to set the value for users. 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. Another existing method used is the hidden markov model which appraises the optional ethics among the noticed and unnoticed datasets. Still noisy data occurs in this method and the accuracy and performance lacks. The AMLMS-GA [18] is the another existing method used which uses the association rule for generating rules and the AMLMS (Accurate MLMS) is the algorithm which generate frequent and infrequent dataset from text documents. The GA (genetic algorithm) is an optimization algorithm it classifies the generated dataset based on their relevancy. The AMLMS-GA works better the performance, accuracy and the efficiency is better but, the performance and accuracy can be improved much better with our proposed methods and technique. The proposed system uses the association rule for generating the rules next the *Apriori_AMLMS* which generates frequent and infrequent item set which is also pruned by this algorithm and the modified genetic algorithm is later applied. In order to use genetic algorithm these measurements are needed (*i.e.*) fitness value, crossover and mutation hence it produces and optimized solution.

Association Rule

The association rule generates the frequent itemset which has high frequency of dataset and infrequent itemset which has minimum support which is very essential in generating negative association rules with very lofty confidence. The discovering of negative association rule has created very sky-scraping interest in many researchers nowadays. Hence when we take a text document huge numerous amount of frequent and infrequent itemset is generated but only few amount of data's are important for generating the interesting association rules. Here the challenging depends on identifying useful itemset. Association rule mainly deals with finding positive association rule. The rules that are generated by association rule has two quality measurements support(suppo) and confidence(Confi). But in our system we have concentrated more on generating negative association rule. But it is avery big challenge in mining text document. When we analyse with text document more event handling deals with the infrequent itemset because in textual documents infrequent item set arises more. The AMLMS algorithm helps in generating the frequent and infrequent itemset. The association rule mining is divided into two phases.

1. Identifying frequent and infrequent itemset from database MD
2. Next it identifies positive and negative association rule from these frequent and infrequent itemset.

Algorithm Flow Diagram

The flow diagram describes about the proposed method. A set of documents are taken from the medical database each individual documents are taken for this method. Collection documents are taken for the pre-processing process which is the early process and makes the data gleaming and now the data become more efficient.

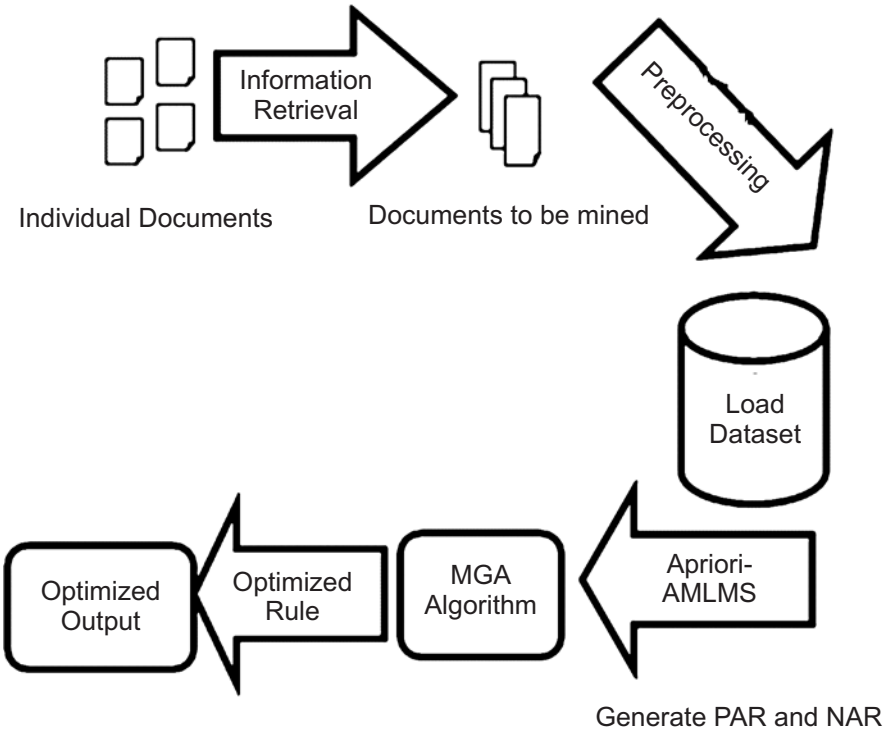

Figure 1: Proposed Algorithm flow diagram

The entire document is converted to lower-case and the non-alphanumeric contents are removed by the stop-word elimination and the grammatical stemming techniques and the noisy data's are removed using the stop-word list. In the grammatical stemming each word is said to be stem and are identified by the stemmer algorithm which removes suffix from the word this is done because each word should be identical in order to repeated words and hence the pre-processing produces a noise free data.

The next step is the load set these are the data which are ready to be processed by the proposed method. It has number of small and bid text documents with single and multiple classes. Each dataset are tested in the proposed method for performance, efficiency and accuracy. The apriori_ MLMSalgorithm now generate the frequent and infrequent itemset from the dataset and from the generated itemset the positive and negative association rules are engendered. The Modified genetic algorithm(MGA) suits well for minimum support value and minimum confidence value for generating positive association and negative association rule. The algorithms is well enhanced for processing optimization rule even for large datasets.

Proposed Apriori-AMLMS Algorithm

The apriori-AMLMs algorithm is used to identify the frequent and infrequent item set in the dataset. The Apriori is ancompitent algorithm that uses bottom up search to identify all the frequent itemset from the database. Proposed method uses Apriori-AMLMS algorithm in order to identify not only the frequent itemset but also all the interesting infrequent item set. Let I = $\{i_1, i_2, i_3.....i_m\}$, where m is the

distinct literals or we call as terms. Let DS be the transaction database (in the form of documents). Each transaction S is set of items S is a subset of I.Each transaction S has a unique identifier S_{ID}. Consider X(antecedent),Y(consequent)are the two set of items where $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, $X \cap Y = \varnothing$. Here the two eminence measurements are used to Support(suppo) and Confident (confi) along with this measurement the Apriori-AMLMS algorithm also provide another measurement named *MLMS* is denoted asmit in order to identify the interesting positive and negative association rule. The lift value let know where the relationship among X and Y generates what type of association. Once the Support value is high than the user's suggestion then it is defines as minsuppo, hence only minimum threshold transaction is presented. Once the confidence is greater than user's suggestion then it is defined as minconfi these conditions are possible between X and Y then it is a valid rule. Once the association is a valid one the *mit* measurement identifies the positive and the negative association rule. If the value of mit is greater than 1 then it is said as positive rule and if it is less than one then it is negative rule and if the value equals 1 then recognise no association rule generated.

Suppo(X) defines no transaction containing X, suppo($X \Rightarrow Y$) = Prob(X,Y), denotes the number of transaction where X and Y coalesce. The confidence

$$\text{Confi}(X \Rightarrow Y) = \frac{\text{Prob}(X \Rightarrow Y)}{\text{Prob}(X)},$$

confidence measure X and Y occurs in the transaction. = $\dfrac{1 + \text{Prob}(XY) - \text{Prob}(X)\text{prob}(Y)}{\text{Prob}(X)\text{Prob}(Y)}$, each item is denoted $X \subset I$

$$\text{mit}(X \Rightarrow Y) = \frac{\text{Prob}(X, Y)}{\text{Prob}(X)\text{Prob}(Y)},$$

the lit measures the strength of the relationship between the X and Y. Consider minsuppo(*n*) is the minimum support of *n* items where *n* = 1,2,3...and minsuppo(0) is the threshold value.

If suppo(X) $\geq$ minsuppo(number(X)),then the X is frequent itemset
If suppo(X) $<$ minsuppo(number(X), then the X is infrequent itemset
If mit($X \Rightarrow Y$) $>$ 1, then it is positive association rule
If mit($X \Rightarrow Y$) $<$ 1, then it is a negative association rule
If(mit($X \Rightarrow Y$) = 1 , then no association rule is generated.

Proposed Algorithm Apriori-AMLMS-MGA

This algorithm derives the frequent and infrequent itemset

Input : DS (transaction document database)
Minsuppo, minimum support threshold
Step 1: initialize FIS = $\varnothing$ and inFIS = $\varnothing$
Step 2: temp1 = $\forall$ X|X*c* $\in$1–itemset, suppo
2.1. FIS_1 = {X|X $\in$ temp1 and suppo(X) $\geq$ minsuppo);
2.2. inFIS1= temp1-FIS1
Step 3: while ($temp_{n-1} \neq \varnothing$) do begin
3.1. C_k = apriori_generate(tempn-1,minsuppo(0));
3.2. for each transaction $d \in$ DS
Do begin
C*d* = subset (C_n, *d*);
For each candidate $c \in C d$

c.count++

end

c.suppo = $\left(\frac{\text{c.count}}{|\text{DS}|};\right)$

$temp_n$ = {c|c∈C_n and (c.suppo ≥ minsuppo)}

Step 4: FIS_n = {X∈$temp_n$ and X.suppo ≥ minsuppo)};

Step 5: $inFIS_n$ = $temp_n$ – FIS_n

Step 6: $FIS_n \cup_n inFIS_n$

Step 7: $inFIS_n = \cup_n inFIS_n$

Step 8: *n*++

End

Step 9: Return FIS and inFIS

Algorithm 1: Proposed algorithm for identifying frequent and infrequent items

The apriori_AMLMS algorithm engender all the frequent and infrequent itemset from a given database DS. As we look out the algorithm the Step 1 contains the initialization part and in Step 2 engender temp1 for all itemset of dimension 1. The step 2.1 and step 2.2 engender FIS1(frequent itemset) and inFIS1 (infrequent itemset) during the first surpass of DS. The step 3 engender the, where $n \geq 2$ by the while loop, the FIS_n is generated for *n*-itemset the support is FISnamdinFIS_n greater than the user defined threshold in the n^{th} surpass of DS. And again the $inFIS_n$ is also generated with minimum support than the user defines minimum threshold. This is an iterative process which continues till $temp_{n-1} = 0$. The step 3 has three substeps the Step 3.1 the candidate itemset C_n for all *n*-item set in DS is engender. Step 3.2 uses a loop to count candidate itemset*Cn* in DS next the Step 3.3 derives the support for each candidate in C_n in Step 3.4 the generated items are stored in temporary data structure. Now the FISnamdinFIS_n, are engender in Step 4 and Step 5. FIS_n is very useful pattern which has maximum value than the minsuppo value and the infrequent items $inFIS_n$ which has minimum value than the minsuppo value. The Step 6 and Step 7 raise the itemset size. The FISnamdinFIS_n are added to the FIS and inFIS. The Step 9 ends the process and presents the output by discovering the frequent and infrequent itemsets.

Proposed Algorithm for Generating Positive and Negative Association Rule

The next is the proposed algorithm for producing positive and negative association rule from the discovered frequent itemset (FIS) and infrequent itemset (inFIS) which is derived from Algorithm 1.

Input

Minisuppo; miniconfi; FIS; inFIS;

Output

NAR (Negative Association Rule)

PAR (Positive Association Rule)

/*algorithm for generating PAR and NAR from FIS

Step 1: initialize NAR = ∅; PAR = ∅;

Step 2 : For everyitemset I in FIS

Do begin

Each itemset X ∪ Y = I, X ∩ Y = ∅

Do begin

2.1: if confi(X ⇒ Y) ≥ miniconfi && mit(X ⇒ Y) ≥ 1

Then output(X ⇒ Y); PAR ∪ (X ⇒ Y)

Else

2.2: ifconfi($X \Rightarrow \neg Y$) ≥ miniconfi && mit($X \Rightarrow \neg Y$) ≥ 1
Then output($X \Rightarrow \neg Y$); NAR ∪ ($X \Rightarrow \neg Y$)
ifconfi($\neg X \Rightarrow Y$) ≥ miniconfi && mit($\neg X \Rightarrow Y$) ≥ 1
Then output($\neg X \Rightarrow Y$); NAR∪($\neg X \Rightarrow Y$)
ifconfi($\neg X \Rightarrow \neg Y$) ≥ miniconfi && mit($\neg X \Rightarrow \neg Y$) ≥ 1
Then output($\neg X \Rightarrow \neg Y$); NAR∪($\neg X \Rightarrow \neg Y$)
End for;
End for;
/*algorithm for generating PAR and NAR from inFIS
For itemset I in inFIS
Do begin

Step 3: For every itemset $X \cup Y = I$, $X \cap Y = \varnothing$
Suppo(X) ≥ minisuppo and Suppo(Y) ≥ minisuppo
Do begin

3.1: if confi($X \Rightarrow Y$) ≥ miniconfi && mit($X \Rightarrow Y$) ≥ 1
Then output($X \Rightarrow Y$); PAR∪($X \Rightarrow Y$)
Else

3.2: ifconfi($X \Rightarrow \neg Y$) ≥ miniconfi && mit($X \Rightarrow \neg Y$) ≥ 1
Then output($X \Rightarrow \neg Y$); NAR∪ ($X \Rightarrow \neg Y$)
ifconfi($\neg X \Rightarrow Y$) ≥ miniconfi && mit($\neg X \Rightarrow Y$) ≥ 1
Then output($\neg X \Rightarrow Y$); NAR ∪ ($\neg X \Rightarrow Y$)
ifconfi($\neg X \Rightarrow \neg Y$) ≥ miniconfi && mit($\neg X \Rightarrow \neg Y$) ≥ 1
Then output($\neg X \Rightarrow \neg Y$); NAR∪ ($\neg X \Rightarrow \neg Y$)
End for;

End for;
Step 4 Returns PAR and NAR.

Algorithm 2: Proposed algorithm for mining PAR and NAR

Modified Genetic algorithm (MGA) : The Modified genetic algorithm(MGA) uses genetic operators which carry out evolutionary process. It is selection process for identifying new population this new population is transformed to achieve new solution by undergoing genetic operations. It uses heuristic scan process which is used to engender answer for optimization and search problem for this it uses some primary process (*i*) Selection (*ii*) crossover (*iii*) Mutation (*iv*) heredity relation. The proposed method aprori_AMLMS-MGA in the earlier section the aprori_AMLMS are described. Here the MGA is applied to discover optimization association rule. The proposed algorithm is used in extraction of frequent itemset with (*i*) fitness value which is calculated for eeach individual itemset (*ii*) selection process is achieved from set of dataset the parent data's are identified and selected for reproduction (*iii*) reproduction it is achieved by the genetic operator crossover and mutation (*iv*) replacement replaces some individuals with another usually with parent data. The generation is said to be one complete transformation of the population.

2. EXPERIMENT RESULT

For the proposed method Apriori_AMLMS-MGA algorithm the analysis is done from the medical database for experimentation the number of information used is 2000 each information contains nearly 150 words the total number of words(attributes) with noisy data are 290346 and after the data cleaning process the

number of words (attributes) present 203640.the average words(attribute) used by this apriori_AMLMS-MGA algorithm is 82648. The experimental analysis first shows how the frequent and infrequent itemset are generated by the support(minsuppo) value as input for the dataset DS.

Table 1

Support value

Support	*Frequent itemset*	*Infrequent itemset*
0.10	135367	286839
0.15	113965	296890
0.20	104436	328655
0.25	83476	473191
0.3	68981	494308
0.35	48935	509370

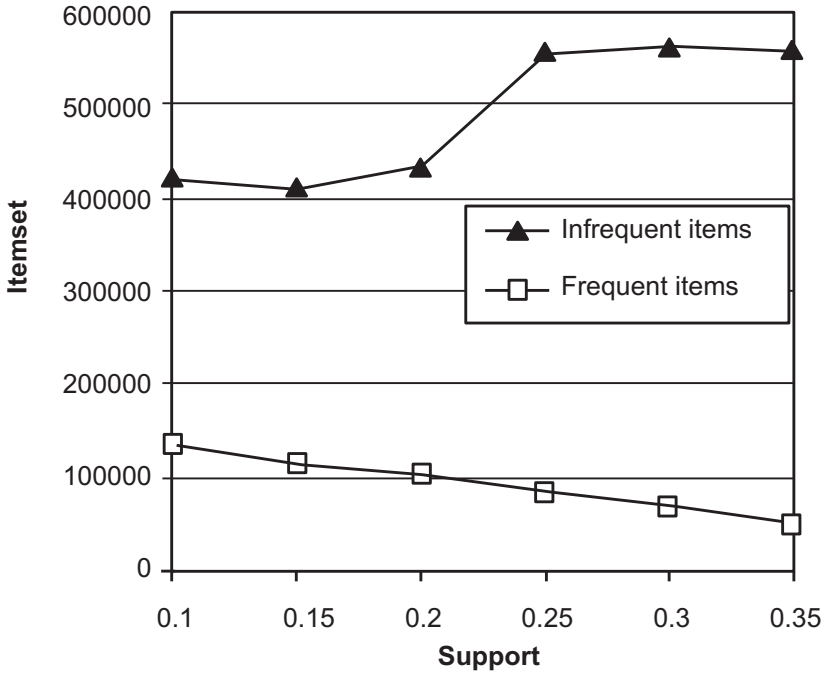

Figure 2: frequent and infrequent item From suppot value

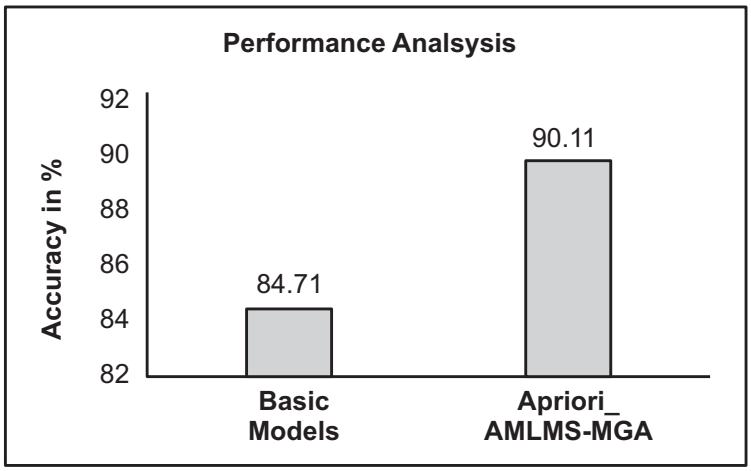

Figure 3: Performance Analysis of the proposed method

The proposed method is then compared with the existing method for performance and the relevance measures. The patterns that are extracted from the proposed system are evaluated for performance, accuracy and efficiency and got a very good performance result. The proposed method was compared with the performance of the existing method MLMS, IMLMS, AMLMS_GA these are considered as basic methods. When compared with these methods the performance of apriori_AMLMS-MGA algorithms works best. The performance analysis is shown in Figure 3 it is noticed that the performance and accuracy has increased compared with the basic model.

Relevance Measure

The relevance efficiency has two important measures the Precision and the Recall. The fraction of recapture document is known as the Recall and the fraction of pattern that is retrieved is known as the Precision. From the recall and precision value the effectiveness of the proposed method is measured.

$$E_{(\alpha - value)} = \frac{(\alpha^2 + 1) * Prec * Rec}{\alpha^2 * Prec * Rec}$$

Here E_α is the estimated value from the precision and recall. In the proposed system the α value is given as 1. Hence the efficiency measure is give as

$$E_{1.1} = \frac{2 * Prec * Rec}{Prec * Rec}$$

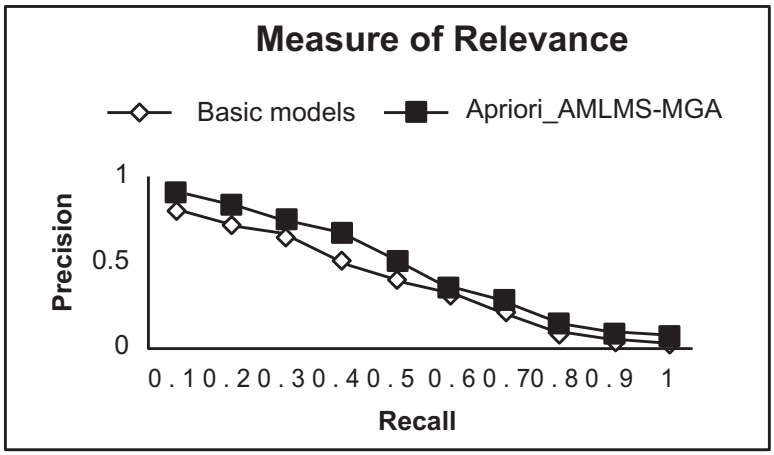

Figure 4: Relevance Measure

The Figure 4 clearly shows how the rules are relatively reduced by the proposed method and the proposed method shows hoe the negative association rule and positive association rule is accurately and more efficiently generated from the frequent and infrequent itemset by this proposed apriori_AMLMS-MGA.

3. CONCLUSION

Mining negative association rule has now got a great interest among the researchers. The proposed method generates both positive and negative associations rule for medical database which depicts association between diseases and symptoms which essentially helps the medical consultant in diagnosis. Here from the frequent itemset the negative association rules are generated and from the infrequent itemset the positive association rules are generated through this new apriori_AMLMS-MGA method apart from the tradition association rule mining. It is clearly demonstrated through the experimental analysis that the proposed method is very accurate and provides a very good performance and more efficient. In future the quality of the generated rules can be improved.

4. REFERENCES

1. NingZhong, Yuefeng Li, and Sheng-Tang Wu, “Effective Pattern Discovery for Text Mining” IEEE Transactions vol. 24, no. 1, January 2012.
2. KavithaMurugeshan, Neeraj RK,” Discovering Patterns to Produce Effective Output through Text Mining Using Naïve Bayesian Algorithm” International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-6, May 2013.
3. LuepolPipanmaekaporn and Yuefeng Li,” A Pattern Discovery Model for Effective Text Mining” Springer, 2012, pp-540.
4. XiangjunDong,ZhendongNiu, DonghuaZhu,ZhiyunZheng, QiutingJia, “Mining Interesting infrequent and frequent Itemset based on MLMs Model” Springer, 2008, pp 444-451.
5. Xiangjun Dong, ‘“Ming Interesting Infrequent and Frequent Itemset Based on Minimum Correlation Strength”, Springer, 2011, pp 437-443.
6. Mrs.K.Mythili, MrsK.Yasodha,” A Pattern Taxonomy Model with New Pattern Discovery Model for Text Mining”International Journal of Science and Applied Information Technology Volume 1, No.3, ISSN No. 2278-3083July – August 2012.
7. CharushilaKadu, Praveen Bhanodia, Pritesh Jain, “Hybrid Approach to Improve Pattern Discovery in Text mining”International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 6, June 2013.
8. 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, June 2012.
9. Spyros I. Zoumpoulis, Michail Vlachos, Nikolaos M. Freris, Claudio Lucchese, “Right-Protected Data Publishing with Provable Distance-based Mining “ IEEE Transactions on knowledge and data engineering, vol. 21, no. 19, november 2012.
10. K. Aas and L. Eikvil, “Text Categorisation: A Survey,” Technical Report Raport NR 941, Norwegian Computing Center, 1999.
11. R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” Proc. 20th Int’l Conf. Very Large Data Bases (VLDB ’94), 1994,pp. 478-499.
12. H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, “Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections,” Proc. IEEE Int’l Forum on Research and Technology Advances in Digital Libraries (ADL ’98),1998, pp. 2-11.
13. N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders, “Word- Sequence Kernels,” J. Machine Learning Research, vol. 3,2003, pp. 1059-1082.
14. Miss DiptiS.Charjan, Prof. MukeshA.Pund ,” Pattern Discovery For Text Mining Using Pattern Taxonomy”. International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 10- October 2013.
15. J. Han and K.C.-C. Chang, “Data Mining for Web Intelligence,”Computer, vol. 35, no.11, Nov. 2002,pp. 64-70.
16. J. Han, J. Pei, and Y. Yin, “Mining Frequent Patterns without Candidate Generation,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’00), 2000,pp. 1-12.
17. Rupesh Dewang, JitendraAgarwal, ”Anew Methos for Generating All Positive and Negative Association rules” International Journal on computer Science and Engineering Volume 4, No.4, ISSN No. 0975-3397 April 2012.
18. I.Berin Jeba Jingle and Dr.J.Jeya A.Celin, “Markov Model in Dicovering knowledge in Text mining”, Journal of Theoretical and Applied Information Technology”vol.70,no.3,Dec 2014,pp.459-463.
19. I.Berin Jeba Jingle and Dr .J. Jeya A.Celin” Discovering Useful Patterns in Text Mining using AMLMS-GA algorithm, ”International Journal of Applied Engineering Research” vol.10, no.18, 2015,pp.39763-39767.