A Novel method for Frequent Pattern Mining

K.Rajeswari [#1], Dr.V.Vaithiyanathan[*2]
[#] Associate Professor, PCCOE & Ph.D Research Scholar
SASTRA University, Tanjore, India
[1] raji.pccoe@gmail.com
[*] Associate Dean Research, SASTRA University
Tanjore, India
[2] vvn@it.sastra.edu

***Abstract*— Data mining is a field which explores for exciting knowledge or information from existing substantial group of data. In particular, algorithms like Apriori aid a researcher to understand the potential knowledge, deep inside the database. However because of the huge time consumed by Apriori to find the frequent item sets and generate rules, several applications cannot use this algorithm. In this paper, the authors describe a novel method for frequent pattern mining, a variation of Apriori Algorithm, which will reduce the time taken for execution to a larger extent. Experiments were conducted with a number of benchmark and real time data sets and it is found that the new algorithm, proposed has better performance in terms of time taken and complexity**

Keyword- Data mining; Apriori; frequent item sets.

I. INTRODUCTION

In today's world, huge amount of data is collected & stored in databases. Data stored in internet, business applications like super market, medical domain in the form of text and images, organizational data, employee data, student data and much more are recorded. But identifying interesting correlation from data can help and improve decision making process [4]. Correlation include finding interesting associations between patterns like whether a particular page has a correlation with an another page always, if symptom i is present then always symptom j is present, if item I is purchased then item J is also purchased etc.. Decision based on these correlations can be made like whether a web page has to be repaired or not, to give a customer a loan or not, if a patient will get a disease or not. Frequent Item Mining is the study of Discovery of Associations and Correlations among items in large datasets. The standard algorithm which is traditionally used is the popular Apriori Algorithm.

Apriori is the basic algorithm for finding frequent item sets[1]. It Mines Frequent item sets for Boolean Association Rules. The algorithm uses iterative approach known as level-wise search. The Apriori Property says "All nonempty subsets of a frequent item set must also be frequent". Some of the key terms used in Apriori are as follows[2]:

- An Item set - A group of individual or other items E.g. {milk, bread, diaper}. A k-tem set is an item set that have k items.
- Support Count(σ) -It is the number of occurrences of an item set. E.g. σ({milk, bread, diaper}) = 2. Support is the fraction of transactions that hold an item set. E.g. s({milk, bread, diaper}) = 2/5. Support(s) is the fraction of transactions that hold both X and Y.
- Frequent Item set - An item set whose support is larger than or equal to a minsup threshold
- Association Rule - An implication expression of the form X⟶ Y, where X and Y are itemsets. E.g. {milk, diaper}⟶{beer}

The Pseudo Code of Apriori Algorithm[1] is summarized as given below.

1. Let k = 1
2. Generate frequent itemsets of length 1
3. Repeat until no new frequent itemsets are identified
4. Generate length k+1 candidate itemsets from length k-frequent itemsets
5. Prune candidate itemsets containing subsets of length k that are infrequent
6. Count the support of each candidate by scanning the databaseproduction
7. Eliminate candidates that are infrequent, leaving only those that are frequent

- Two Step Approach
    1. Frequent item set production - Produce all itemsets whose support ≥ minsup.
    2. Rule Generation. Generate high confidence rules from each frequent itemset.

II. LITERATURE REVIEW

A wide range of experiments on synthetic and real world data sets were conducted. Correlation based feature selection is used for Arrhythmia classification [5]. 22 attributes were selected giving good accuracy with different classifiers like Bayes classifier, Support vector machines, Neural Networks (MLP), C4.5 Decision tree classifiers. The findings of the study[6] have described uses concept hierarchies and fuzzy techniques. The summaries are produced at diverse levels of granularity, according to the concept hierarchies. Mining large datasets became a major issue. Hence research focus was diverted to solve this issue in all respect. It was the primary requirement to devise fast algorithms for finding frequent item sets as well as mining. The paper[1] has dealt this issue in depth and proposed a new approach that adopts subset lattice search space, using structural properties of frequent item sets to facilitate fast discovery. A cubic structure based algorithm[7] for fast discovery of frequent item sets is found which implements mining on large datasets where sheer volume of frequent patterns will be generated which are tough to be managed for further use. And also large frequent pattern set may degrade the performance of mining process[3]. Finally, the paper[8] has proven that the problem of counting maximal frequent item sets in a database of a given arbitrary support threshold is a polynomial time NP-hard problem.

III. . PROPOSED METHOD

The method proposed theoretically in paper[8] is tested experimentally in this work.

TABLE I. A Simple Example of Patient * Symptom Matrix

Rec.No	Symptom A	Symptom B	Symptom C	Symptom D
P1	1	0	1	0
P2	1	1	1	0
P3	0	0	0	1
P4	0	1	1	0
P5	1	1	0	1
Support	3	3	3	2

The Table II is used to record the number of times a symptom occurs in the data base as in Table I. The different data sets used for experimenting is listed in Table III.

TABLE II. List with frequent patterns

Item	Increment in count	No_of_items
A	**1+1+1**	**1**
C	1+1+1	1
A,C	**1+1**	**2**
B	1+1+1	1
A,B	**1+1**	**2**
B,C	1+1	2
A,B,C	**1**	**3**
D	**1+1**	**1**
A,D	1	2
A,B,D	**1**	**3**

TABLE III List of data sets used[11][12]

Database	Total
Stage 1 Diabetes	499
Stage 2 Diabetes	499
Stage 3 Diabetes	499
Stage 4 Diabetes	499

TABLE IV Sample of Discretization

Age	
S.No	**Discretizied values**
1	0_19
2	20_39
3	40_49
4	50_59
5	60_max
Sex	
1	0 (Female)
2	1 (Male)

The data sets are pre-processed by converting the numerical values to categorical terms. This process of discretization is done as shown in Table IV. Real time data sets of diabetes are collected from Diabetes Research Center, Tanjore[11][12].

IV. RESULTS AND DISCUSSION

Experiments were conducted using C- sharp in Visual Stdio .net framework in a laptop with 500 GB Hard Disk Drive, 2 GB DDR3 Memory,. The time taken is exponentially high. The graphs below in Fig. 1, Fig. 2, Fig. 3 and Fig. 4 shows the comparison of the popular Apriori and our new method proposed.

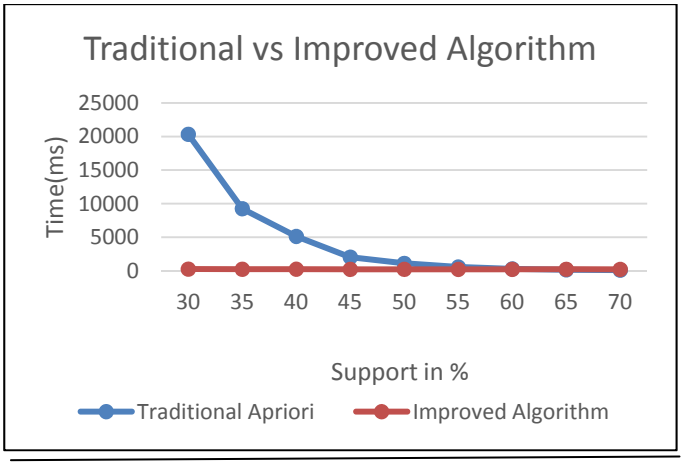

Fig 1. Diabetes Stage 1

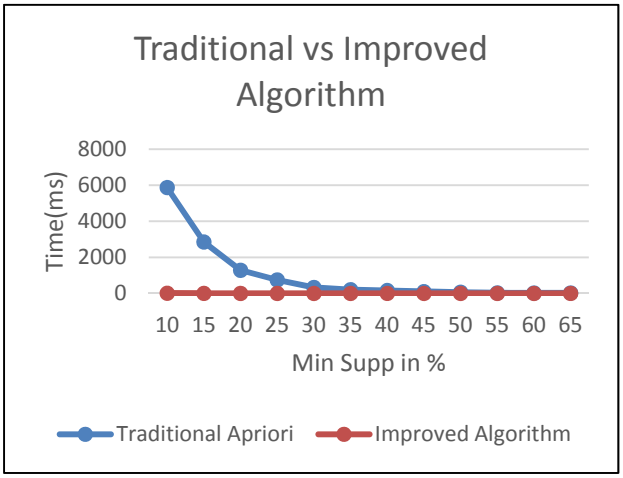

Fig 2. Diabetes Stage 2

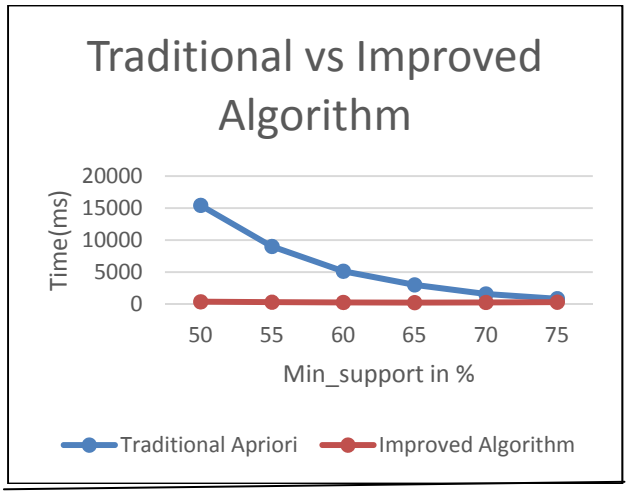

Fig 3. Diabetes Stage 3

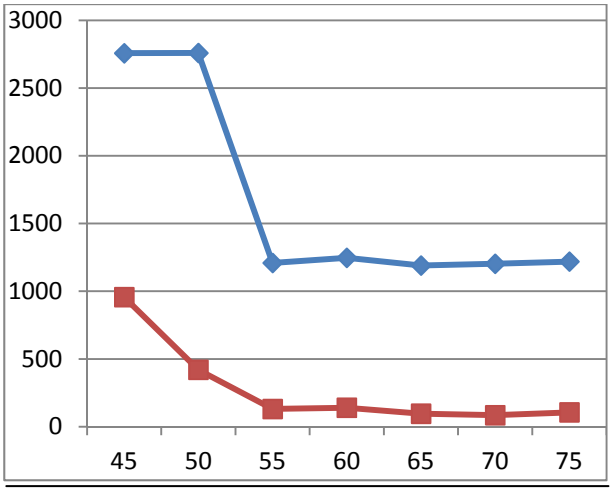

Fig 4. Diabetes Stage 4

The Table V below shows an comparison of Traditional Popular Apriori algorithm and the proposed new approach. The authors have made a comparison with different parameters.

TABLE V Comparative analysis

Sr.No	Parameter	Traditional Apriori algorithm	Novice algorithm proposed
1	No. of combinations generated	t*2**n	t*2**(n/2)
2	No. of comparisons	t*2**n*k	t*k
3	No. of passes	2**n	1

V. CONCLUSION

- The As more scans are required for traditional Apriori algorithm to generate k-frequent item set, the proposed algorithm is designed in a way that it requires only one scan to find k-frequent item set.
- Use of hashing technique improves retrieval of item sets, thereby improving overall efficiency of proposed algorithm.
- Future research will focus on reducing the obtained data and improve the quality of retained patterns.
- Application of frequent pattern mining varies from bioinformatics, web mining, software bug detection and analysis and improves the performance of XML management systems.

REFERENCES

[1] Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." Proc. 20th Int. Conf. Very Large Data Bases, VLDB. Vol. 1215. 1994.
[2] Han, J., & Kamber, M. (2001). Data mining: Concepts and techniques. San Francisco, CA: Morgan Kaufmann Publishers.
[3] Goethals, Bart. "Survey on frequent pattern mining." Ph.D. thesis, HIIT Basic Research Unit, Department of Computer Science, Univ. of Helsinki, Finland (2003)
[4] A.Hall, Thesis on 'Correlation based feature selection for Machine Learning', Apr 1999.
[5] Arıkan, Umut, and Fikret Gürgen. "Discrimination Ability of Time-Domain Features and Rules for Arrhythmia Classification." Mathematical and Computational Applications 17.2 (2012): 111-120.
[6] Raschia, G., L. Ughetto, and N. Mouaddib. "Data summarization using extended concept hierarchies." IFSA World Congress and 20th NAFIPS International Conference, 2001. Joint 9th. Vol. 4. IEEE, 2001.
[7] Ivancsy, Renata, and Istvan Vajk. "Fast Discovery of Frequent Itemsets: a Cubic Structure-Based Approach." Informatica 29 (2005): 71-78.
[8] Yang, Guizhen. "The complexity of mining maximal frequent itemsets and maximal frequent patterns." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.
[9] K.Rajeswari, Dr.V.Vaithiyanathan, 'Mining Association Rules Using Hash Table', Nov 2012 by International Journal of Computer Applications [Impact factor :0.814].DOI/ISBN: 10.5120/9132-3320
[10] K. Rajeswari, Dr. V. Vaithiyanathan(2011), 'Heart Disease Diagnosis: An Efficient Decision Support System Based on Fuzzy Logic and Genetic Algorithm', International Journal of Decision Sciences, Risk and Management by Inderscience Publications. Impact factor 0.253. ISSN : 1753- 7169(print) 1753- 7177 (Online).PP 81-97. DOI : 10.1504/IJDSRM.2011.040749
[11] K. Rajeswari, Dr. V. Vaithiyanathan, Dr.T.Gurumoorthy, 'Modeling Effective Diagnosis of Risk Complications in Type 2 Diabetes – A Predictive model for Indian Situation', EUROPEAN Journal of Scientific Research, Vol.54 No.1 (2011), pp.147-158, ISSN 1450-216X/ 1450-202X.IJCA
[12] K. Rajeswari, Dr. V. Vaithiyanathan ,'Fuzzy based modeling for diabetic diagnostic decision support using Artificial Neural Network', International Journal of Computer Science and Network Security, Vol.11 No.4, April (2011), pp.126-130.