Int J Pharm Bio Sci 2015 April; 6(2): (B) 336 - 340

Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

AN ADAPTIVE TRANSACTION REDUCTION APPROACH FOR MINING FREQUENT ITEMSETS: A COMPARATIVE STUDY ON DENGUE VIRUS TYPE1

D.KERANA HANIREX*[1] AND DR.K.P.KALIYAMURTHIE[2]

[1] Assistant Professor,Department of CSE, Bharath University Chennai-73.
[2] HOD,Department of CSE, Bharath University Chennai-73.

ABSTRACT

Frequent itemset mining plays an essential role in mining various patterns and in real time applications. The dataset utilised in our experimental analysis are real world data set for Dengue Virus Type 1 (DEN1) which is obtained from GenBank:AAB27904.1 which consists of 777 amino acids.In this paper, an adaptive TDTR(Two Dimensional Transactions Reduction) approach which we have proposed earlier is tested against this real Dengue virus type1 dataset and finally compared with standard Apriori algorithm and FP-Growth algorithm. The theoretical analysis and experiments prove its efficiency and accuracy for Dengue Virus Type1 dataset . This system reveals that Leucine(L), Phenylalanine (F),Lysine(K),Serine(S) and Glycine(G) are the dominating amino acids in Dengue Virus Type1 which is the same results produced from Apriori algorithm and FP-Growth Algorithm with high performance.

KEYWORDS: **Data Mining, Association Rule Mining(ARM),Apriori,FP-Growth,TDTR**

D.KERANA HANIREX
Assistant Professor,Department of CSE, Bharath University Chennai-73.

***Corresponding author**

INTRODUCTION

Frequent Itemset Mining is a KDD technique which is a basics of many other techniques, such as association rule mining , sequence pattern mining ,classification , clustering[1,12] and so on. An itemset is frequent when its weight is not under the minimum threshold value. Large volume of data are collected and stored for various business operations and for different applications. This process is known as data mining [2,3] or knowledge discovery [4] in databases. The association rule problem is divided into 2 subproblem. First identify all frequent itemset from the database and then construct the association rule using frequent itemset [5]. This paper deals with an efficient way of finding the frequent itemsets which leads to find the dominating amino acids among dengue virus type1 (DEN1). The following section describes about the association rule mining ,the results and discussions and finally concludes the paper.

ASSOCIATION RULE MINING

Mining frequent itemset is an important field in data mining. Frequent itemsets are subsets often occurring in a collection of items .Generating association rule from the frequent itemset is a simple task.But researchers focuses on optimizing the association rule discovery.It was originally proposed by Agarwal et al with Apriori algorithm[6].This algorithm uses generate-and-test approach for generating frequent itemsets. At each level, all candidate itemsets are tested by scanning the original database.The drawback of this algorithm is that if there is a large frequent itemset with size i,then 2^i subsets might be generated and tested which leads to multiple database scans and large number of candidate itemset generation.

RELATED WORK

Researchers suggests various techniques to improve this efficiency of the Apriori algorithm.Various improved Apriori Algorithm[7] were proposed. One of the fastest algorithm for frequent itemset mining is the FP-Growth (Frequent Pattern –Growth Algorithm) algorithm. FP-Tree algorithm is based on a prefix tree representation of the given database transactions which can save considerable amounts for memory for storing the transactions[8]. Several researchers proposed different techniques for association rule mining.The author focuses on mining the recent frequent itemset using regression[9] parameter. Hash technique[10] is also used to improve mining FIs.A tree based algorithm FP-Streaming [11] is proposed for mining streams of precise data. In our earlier research ,we have presented an TDTR approach for mining frequent itemsets. This TDTR approach reduces the number of transactions from the original database based on the minimum threshold value thus improving the performance.In this paper , the TDTR[14] approach is applied with Dengue Virus Type 1 data set and its efficiency can be compared with the standard algorithm like Apriori and FP-Growth.

DATASET DESCRIPTION

Research on biological science is one of the emerging task to solve the crucial problems particularly in finding the drugs. This system finds the dominating amino acids among the dengue virus which is used to discover the drugs for the dengue fever.We have tested our approach with sequential datasets for Dengue virus 1(DEN1): GenBank:AAB27904.1 which consists of 777 amino acids.

DATA PREPROCESSING

Data need to be processed in order to improve the quality of the data .The various tasks of data mining are data transformation, data integration, data discretization ,data cleaning and data reduction. Data cleaning involves removing noisy data, incomplete data, inconsistent data. Data integration combines data from multiple sources. Data transformation task contains data aggregation, generalization, data smoothing, and normalization. Data reduction includes data aggregation, high dimensionality reduction, data compression and discretization. In this paper we have applied preprocessing using Weka3.6.4 tool for this Dengue virus data set.

ALGORITHM

TDTR ALGORITHM

Input: Database D,min_support s

Output : frequent itemset

//Algorithm to find frequent itemset
1.FOR each $t_i \in D$ DO BEGIN
a.count the number of items in count1[i]
b. If the count1[i] ≥ min_sup then put the tranactions into D_1
2.FOR each Ii $\in D_1$ DO BEGIN
a. count the number of transactions in count2[i]
b. if count2[i] < min_sup then remove that Ii from D_1
3*. Find all frequent itemsets from D_1*

RESULTS AND DISCUSSIONS

EVALUATION METRICS

For evaluating the performance our TDTR approach ,we have chosen the well known standard algorithm such as Apriori and FP-Growth tree algorithm. We have chosen the evaluation parameter such as efficiency and accuracy. Accuracy of the system is measured by considering the number of rules generated by varying the confidence. The Efficiency of the system is measured by considering the time taken in secs.

PERFORMANCE ANALYSIS ON THE ALGORITHMS (Apriori,FP-Growth,TDTR approach)

If we take the minimum support as 10% by varying the confidence from 90% to 50% we are getting the number of association rules which is given in the following table

Table 1
No. of rules generated for different confidence value

Confidence	Apriori[13]	FP-Growth	TDTR
90	5	5	60
80	30	20	186
70	87	63	375
60	210	145	611
50	388	272	897
Total number of rules generated	1246	896	2708

From the table we can understand clearly for the increasing confidence value the number of rules get decreased. By varying the support value for different confidence the following number of rules are generated and the accuracy can be measured for different algorithm such as Apriori, FP-Growth and TDTR

.

Table 2
Apriori algorithm

No. Of Rules generated				
Confidence	Support			
	40	30	20	10
80	1	2	2	30
70	2	3	13	87
60	6	13	39	210
50	6	22	74	388

Table 3
FP-Growth Algorithm

No. Of Rules generated				
Confidence	Support			
	40	30	20	10
80	1	1	1	20
70	2	3	9	63
60	4	9	27	145
50	6	18	56	272

Table 4
TDTR Algorithm

Confidence	Support			
	40	30	20	10
80	2	3	18	186
70	5	11	52	375
60	14	31	111	611
50	16	41	162	897

From table 2,3 and 4 we can understand the TDTR approach generates large number of association rules as compared to Apriori algorithm and FP-Growth Algorithm.The efficiency of the system is measured by considering the time it takes to find the association rule. The following table shows the time taken in seconds for different confidence measures.

Table 5
Time taken for different confidence value

Time taken for different confidence(in secs)			
confidence	Apriori	FP-Growth	TDTR
90	.0020	.0018	.0014
80	.0012	.0010	.0009
70	.0010	.0008	.0007
60	.0009	.0007	.0005
50	.0008	.0006	.0004

From the above table we can understand clearly for the increasing confidence value the execution time gets increased. The execution time of our TDTR approach is considerably lesser than the Apriori algorithm and FP-Growth Algorithm which increases the efficiency of the system. The efficiency of the frequent itemset mining for dengue virus type1 data set can be given in the following figure

.

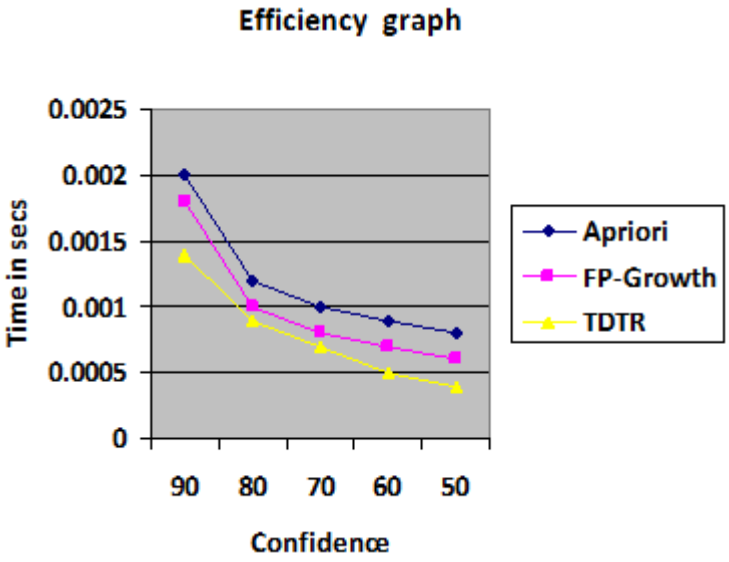

Figure 1
Efficiency graph

CONCLUSION

In this paper we have investigated our adaptive TDTR approach for mining the frequent itemset using the real Dengue Virus Type1 Dataset and the results were compared with standard Apriori algorithm and FP-Growth Algorithm. This TDTR approach finds the dominating amino acids by finding the frequent itemsets among the dengue virus data set. This system reveals that Leucine(L), Phenylalanine (F),Lysine(K),Serine(S) and Glycine(G) are the dominating amino acids in Dengue Virus Type-1 which is the same results produced from Apriori algorithm and FP-Growth Algorithm.But this TDTR algorithm exhibits good performance and stable behaviour with execution time. This simple algorithm can be easily extensible. It is theoretically and experimentally proved that this proposed approach solves the problem efficiently and accurately.

REFERENCES

1. Wei Hou,Bingru Yang ,Zhun Zhou,Chensheng Wu.An adaptive Frequent Itemset mining Algorithm for Data Stream with Concept Drifts,382-385,IEEE International Conference on CSSE,(2008).
2. R.Agarwal,S.Ghosh, A.Swami.Database Mining: A performance Perspective, IEEE transactions on Knowledge and Data Engineering,914-925,(1993).
3. S.Tsur ,Data Dredging,IEEE Database Engineering Bulletin,13(4):58-63,(1990).
4. G.Piatetsky- hapiro,Editor,Knowledge Discovery in databases ,AAAI/MIT Press(1991).
5. Han.J,Efficient mining of partial periodic patterns in time series database,15th International Conference on data engineering,IEEE computer society,Sydney,Australia:106-115,(1999).
6. R.Agarwal,R.Srikant . Fast algorithms for mining generalized association rules,20th International conference on very large databases (VLDB'94), 1994).
7. Lijuan Zhou,Shuang Li,Mingsheng Xu. Research on Algorithm of Association Rules in Distributed Database System, IEEE 2nd Int. Conference on Informatics in Control,Automation and Robotics,3:216-219(2010).
8. J.Han,.Pei and Y.Yin, Mining Frequent Patterns without candidate generation ,In Proceedings of the .Conference on the Management of Data(SIGMOD'00), (2000).
9. J.Chang W.Lee, A Sliding Window Method for Finding Recently Frequent Itemsets over Online Data Streams,Int. journal of Information Science and Engg.,20,753-762,(2004).
10. C.Jin,W.Qian,C.Sha,J.X.Yu, A.Zhou, Dynamically Maintaining Frequent Items Over a Data Stream,In CIKM the International Conference on Information and Knowledge Management , (2003).
11. C.Giannella,J.Han,J.Pei,X.Yan P.S.Yu, Mining Frequent patterns in data streams at multiple time granularities in Data Mining: Next Generation Challenges.
12. Kerana Hanirex.D,M.A.Dorai Rengaswamy,Efficient Algorithm for Mining Frequent Itemsets using Clustering techniques,IJCSE ,3(3):1028-1032;(2011).
13. Kerana Hanirex.D, K.P.Kaliyamurthie ,Finding the Dominating Amino Acids in Dengue Virus Type1 Study on mining frequent itemsets, 4(3):(B);880 – 889, Int. Journal of Pharma and Bio Sciences, July, (2013).
14. Kerana Hanirex.D, An Efficient TDTR Algorithm for Mining Frequent Itemsets, International Journal of Electronics and Computer Science Engineering,2(1):251-256;(2012).