Int J Pharm Bio Sci 2015 April; 6(2): (B) 288 - 295

Research Article **Bioinformatics**

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

ANALYSIS OF IMPROVED TDTR ALGORITHM FOR MINING FREQUENT ITEMSETS USING DENGUE VIRUS TYPE 1 DATASET: A COMBINED APPROACH

D.KERANA HANIREX*[1], DR.K.P.KALIYAMURTHIE[2] AND DR.A.KUMARAVEL[3]

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

ABSTRACT

Association rule mining is the recent data mining research.We have presented an approach for mining frequent itemsets using Dengue virus type- 1 data set. This paper proposes an Improved Two Dimensional Transaction Reduction(ITDTR) algorithm which is a combined approach of transaction reduction and sampling in bio data mining. This system produces the same frequent item sets as produced from Apriori algorithm and FP-Growth Algorithm with the higher performance. This system reveals that Glycine (G), Leucine(L), Serine(S), Lysine(K) , Phenylalanine (F) are the dominating amino acids in Dengue Virus Type-1 data set with higher accuracy and efficiency.The efficiency of this algorithm is compared with Apriori algorithm,FP_Growth algorithm,Genetic algorithm, and TDTR[1,2,3,4] algorithm which we have implemented in our previous research work.

KEYWORDS: Data Mining,Bio Data Mining, Association Rule Mining,Apriori,FP-Growth,Genetic, Distributed ,TDTR and ITDTR

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

***Corresponding author**

INTRODUCTION

Association mining is a broadly used approach in data mining concepts. Data mining has emerged as a field of extensive research from a wide range of diverse groups of people. Data mining is defined as "The non-trivial extraction of inherent ,previously unknown and potentially valuable information from data"[5,6,7].In this paper we have presented an approach for mining frequent itemsets using Dengue virus type 1 data set.This paper proposes an improved TDTR(ITDTR) Algorithm a combined approach of transaction reduction and sampling.This system produces the same results as produced from Apriori algorithm and FP-Growth Algorithm but with high performance. This system reveals that Glycine (G), Leucine(L), , Serine(S), Lysine(K) , Phenylalanine (F) are the dominating amino acids which are strong association rules in Dengue Virus Type-1 with higher accuracy .The efficiency of this algorithm is compared with Apriori algorithm, FP_Growth algorithm, Genetic algorithm,Distributed and TDTR algorithm.

ASSOCIATION RULE MINING

Discovery of association rules is a subfield in data mining.The motivation for searching association rules is to analyze large amounts of super market basket data. Association rule specifies how often items are purchased together.The discovery of association rules can be divided into 2 phases: First discover all frequent itemsets and then association rules using the frequent itemsets[8].There are 2 interesting measures in association rule mining support and confidence.Support determines how often the rules occur in the database.Support of an association rule X =>Y is the ratio of the number of occurrences of {x,y} to the total number of transaction of D. Confidence measure of an association rule X =>Y is the ratio of the total occurrences for item X and Y } to the total number of occurance for item X.

RELATED WORK

Various algorithms have been proposed for association rule mining.Apriori is one of the famous basic algorithm for association rule mining.It uses breadth-first search of the pattern.In Apriori algorithm,candidate itemsets are generated iteratively.Various methods were introduced to improve the efficiency of the algorithm starting from Agarwal [9].The problems we faced in Apriori are multiple database scan and large number of candidate itemsets generation. Paper [10] proposes an algorithm of distributed mining association rules using the improved Apriori algorithm.It generates local frequent item set from different nodes then generates global frequent itemset. Researcher proposes various sampling techniques for association rule mining in order to reduce database size.There are different types of samplings such as random sampling,systematic sampling ,cluster sampling and stratified samplings are available. A.Mahafzah[11] developed parameterized sampling algorithm. V.Umarani proposes[12] progressive sampling based approach for association rule mining.Association rules can also be generated by applying classification techniques[16]. Transaction reduction is another method that helps in mining association rules . It is based on the fact that a transaction that does not contain any frequent k-itemset will not be required in the further database scans. AprioriTid algorithm [13] is another method for improving the performance of Association Rules. This algorithm is used to construct the frequent itemset. Thevar R.E; Krishnamoorthy proposed modified transaction reduction based frequent itemset mining algorithm(MTR-FMA)[14] which maintains its performance even at relatively low supports. Paper [15] proposes hash based approach for mining association rules. This paper proposes an improved TDTR algorithm (ITDTR) which is a combined approach of transaction reduction and sampling techniques.

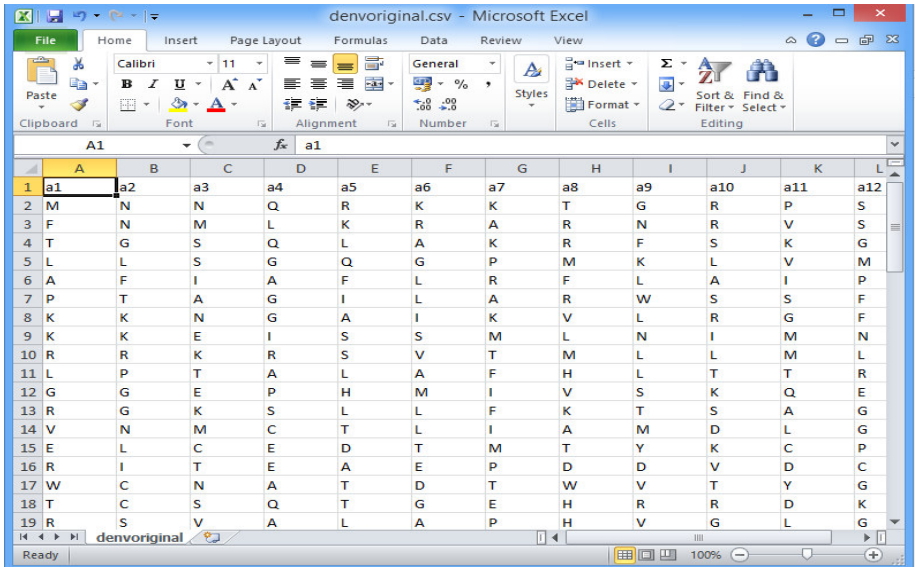
DATA PREPROCESSING

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.

Data preprocessing

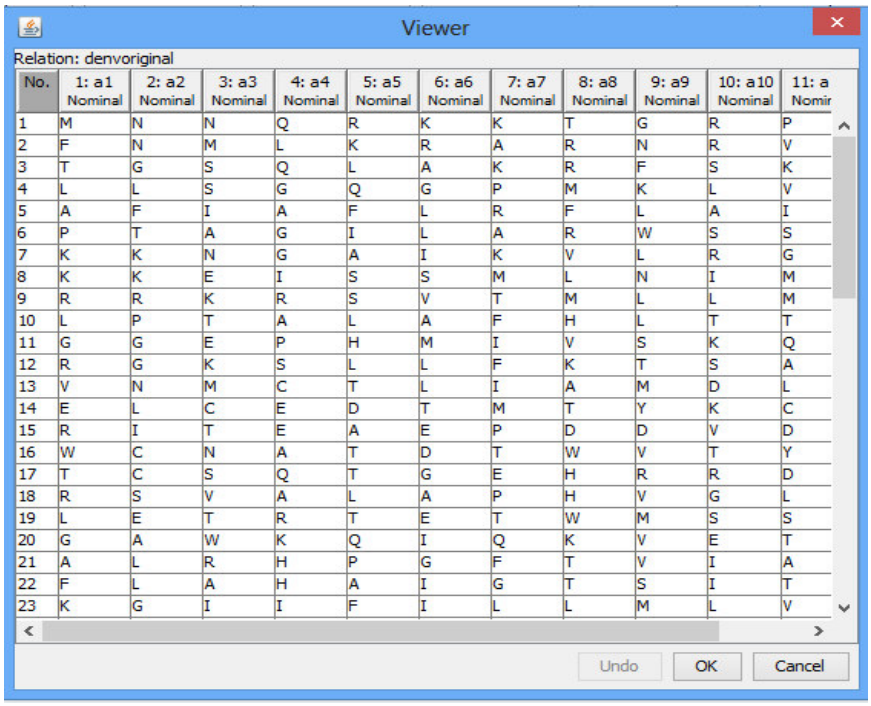
Polyprotein Dengue virus type1 datasets



	A	B	C	D	E	F	G	H	I	J	K	L
1	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10	a11	a12
2	M	N	N	Q	R	K	K	T	G	R	P	S
3	F	N	M	L	K	R	A	R	N	R	V	S
4	T	G	S	Q	L	A	K	R	F	S	K	G
5	L	L	S	G	Q	G	P	M	K	L	V	M
6	A	F	I	A	F	L	R	F	L	A	I	P
7	P	T	A	G	I	L	A	R	W	S	S	F
8	K	K	N	G	A	I	K	V	L	R	G	F
9	K	K	E	I	S	S	M	L	N	I	M	N
10	R	R	K	R	S	V	T	M	L	L	M	L
11	L	P	T	A	L	A	F	H	L	T	T	R
12	G	G	E	P	H	M	I	V	S	K	Q	E
13	R	G	K	S	L	L	F	K	T	S	A	G
14	V	N	M	C	T	L	I	A	M	D	L	G
15	E	L	C	E	D	T	M	T	Y	K	C	P
16	R	I	T	E	A	E	P	D	D	V	D	C
17	W	C	N	A	T	D	T	W	V	T	Y	G
18	T	C	S	Q	T	G	E	H	R	R	D	K
19	R	S	V	A	L	A	P	H	V	G	L	G

Figure 1
CSV file

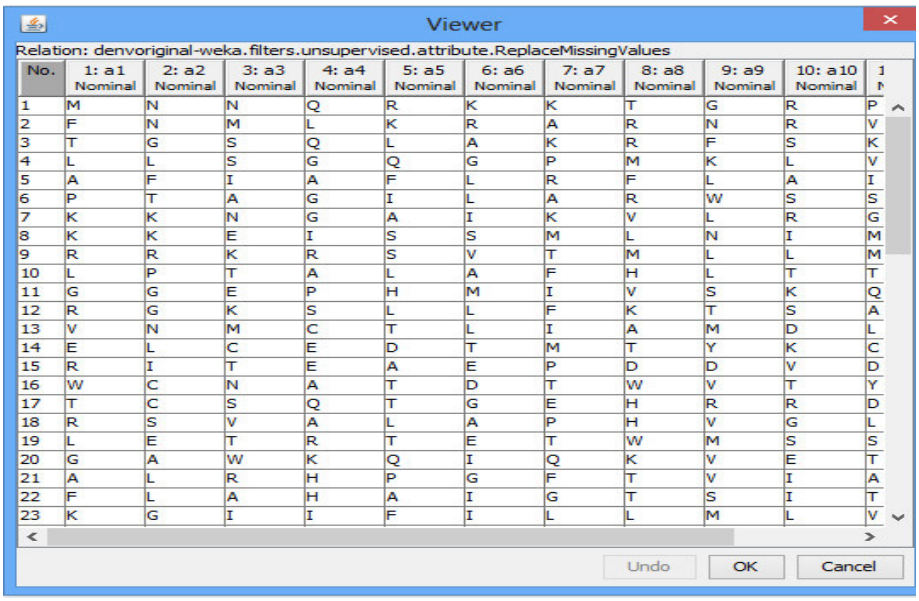
The above figure Figure 1 shows the CSV file.The following figure Figure 2 shows the arff file of polyprotein dengue virus data sets in Weka3.6.4 tool.Figure 3 depicts the dataset after applying the Replacemissingvalues filter using weka tool.



No.	1: a1 Nominal	2: a2 Nominal	3: a3 Nominal	4: a4 Nominal	5: a5 Nominal	6: a6 Nominal	7: a7 Nominal	8: a8 Nominal	9: a9 Nominal	10: a10 Nominal	11: a Nomir
1	M	N	N	Q	R	K	K	T	G	R	P
2	F	N	M	L	K	R	A	R	N	R	V
3	T	G	S	Q	L	A	K	R	F	S	K
4	L	L	S	G	Q	G	P	M	K	L	V
5	A	F	I	A	F	L	R	F	L	A	I
6	P	T	A	G	I	L	A	R	W	S	S
7	K	K	N	G	A	I	K	V	L	R	G
8	K	K	E	I	S	S	M	L	N	I	M
9	R	R	K	R	S	V	T	M	L	L	M
10	L	P	T	A	L	A	F	H	L	T	T
11	G	G	E	P	H	M	I	V	S	K	Q
12	R	G	K	S	L	L	F	K	T	S	A
13	V	N	M	C	T	L	I	A	M	D	L
14	E	L	C	E	D	T	M	T	Y	K	C
15	R	I	T	E	A	E	P	D	D	V	D
16	W	C	N	A	T	D	T	W	V	T	Y
17	T	C	S	Q	T	G	E	H	R	R	D
18	R	S	V	A	L	A	P	H	V	G	L
19	L	E	T	R	T	E	T	W	M	S	S
20	G	A	W	K	Q	I	Q	K	V	E	T
21	A	L	R	H	P	G	F	T	V	I	A
22	F	L	A	H	A	I	G	T	S	I	T
23	K	G	I	I	F	I	L	L	M	L	V

Figure 2
ARFF file

Replace missing values filter



No.	1: a1 Nominal	2: a2 Nominal	3: a3 Nominal	4: a4 Nominal	5: a5 Nominal	6: a6 Nominal	7: a7 Nominal	8: a8 Nominal	9: a9 Nominal	10: a10 Nominal	1
1	M	N	N	Q	R	K	K	T	G	R	P
2	F	N	M	L	K	R	A	R	N	R	V
3	T	G	S	Q	L	A	K	R	F	S	K
4	L	L	S	G	Q	G	P	M	K	L	V
5	A	F	I	A	F	L	R	F	L	A	I
6	P	T	A	G	I	L	A	R	W	S	S
7	K	K	N	G	A	I	K	V	L	R	G
8	K	K	E	I	S	S	M	L	N	I	M
9	R	R	K	R	S	V	T	M	L	L	M
10	L	P	T	A	L	A	F	H	L	T	T
11	G	G	E	P	H	M	I	V	S	K	Q
12	R	G	K	S	L	L	F	K	T	S	A
13	V	N	M	C	T	L	I	A	M	D	L
14	E	L	C	E	D	T	M	T	Y	K	C
15	R	I	T	E	A	E	P	D	D	V	D
16	W	C	N	A	T	D	T	W	V	T	Y
17	T	C	S	Q	T	G	E	H	R	R	D
18	R	S	V	A	L	A	P	H	V	G	L
19	L	E	T	R	T	E	T	W	M	S	S
20	G	A	W	K	Q	I	Q	K	V	E	T
21	A	L	R	H	P	G	F	T	V	I	A
22	F	L	A	H	A	I	G	T	S	I	T
23	K	G	I	I	F	I	L	L	M	L	V

Figure 3
After applying Replace Missing values filter

ALGORITHM

ITDTR 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 in to D_1
2.FOR each $I_i \in D_1$ 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.select sample Si (systematic sampling) from D_1
4.use negative border to select the optimal sample
5. Find all frequent itemsets from D_1

The following figure Figure 4 describes the flow of Improved TDTR (ITDTR) Algorithm.This proposed algorithm first implements Two Dimensional Transaction Reduction (TDTR) Algorithm[4] which consists of row and column reduction from the original database D and produce the reduced database D_1.In the Improved TDTR Algorithm it uses systematic sampling which selects the transaction from the database based on the regular interval.It uses negative border approach to select the optimal sample. The execution time and the number of rules generated by ITDTR are estimated and the results are compared with Apriori,FP-Growth Algorithm,Distributed and Genetic Algorithm

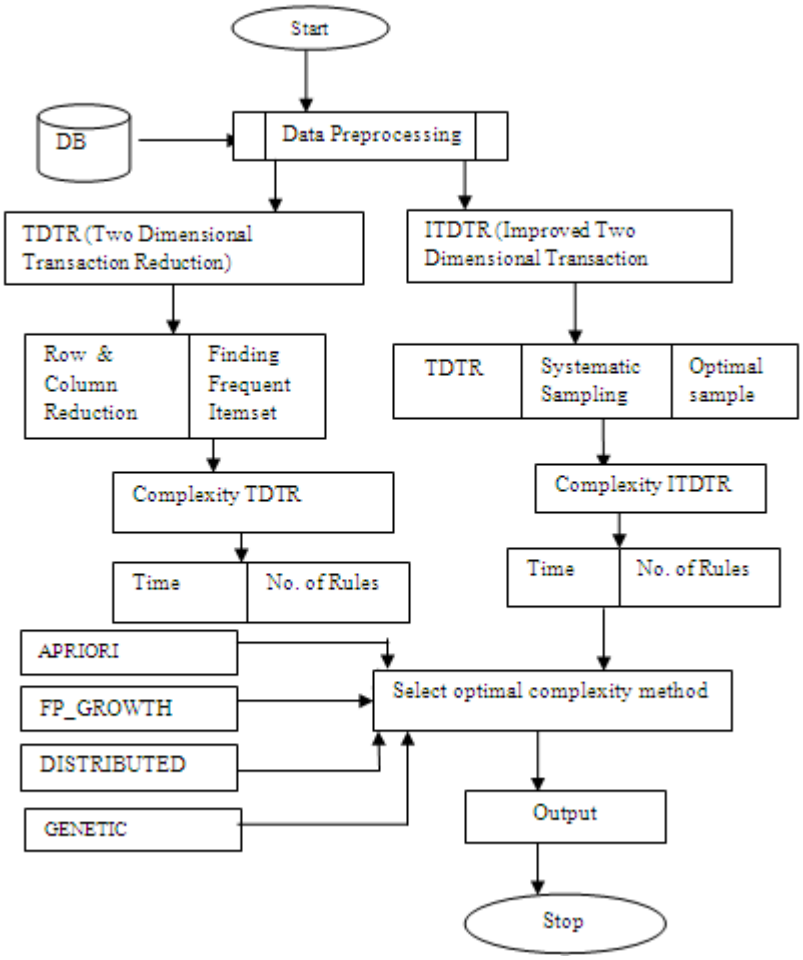


Figure 4
Flow Diagram of Improved TDTR (ITDTR) Algorithm

DATASET DESCRIPTION

This system uses dataset Dengue Virus Type-1 data sets from Gen Bank: AAB27904.1 which consists of 777 amino acids.

EXPERIMENTAL RESULTS

This section describes the results obtained from our improved TDTR (ITDTR) algorithm.The experiments are implemented in java(jdk1.6).The experiment was implemented through the Dengue virus type1 dataset.Here the accuracy is generated by finding the number of association rules generated for different threshold values. The effectiveness of the association rule mining is measured by considering the time taken to generate the association rules from databases.The accuracy of association rule is measured by the number of association rules generated. The following table table1 specifies the number of association rules generated for different confidence measures in various algorithms by taking the support value=10%.

Table 1
Number of association rues generated for different confidence value in various algorithms

Number of Association Rules						
Confidence	Apriori	FP-Growth	Distributed	Genetic	TDTR	ITDTR
90	5	5	58	55	60	70
80	30	20	178	170	186	196
70	87	63	380	390	375	395
60	210	145	625	620	611	630
50	388	272	885	892	897	997

The following figure Figure 5 shows that our proposed approach ITDTR has generated large number of association rules when compared with algorithms such as Apriori, FP-Growth, Distributed, Genetic and TDTR.

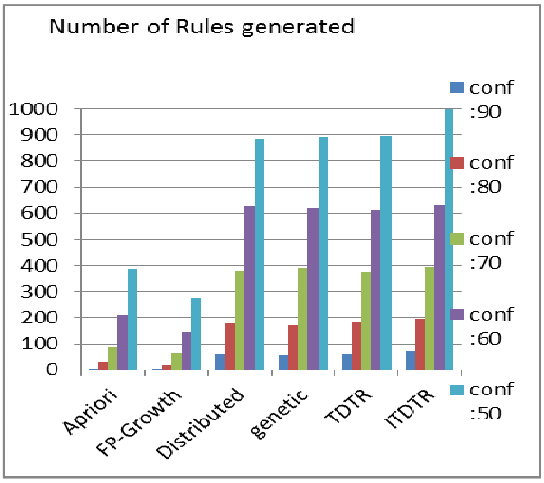


Figure 5
Accuracy graph

The following table table2 shows the efficiency of this Improved TDTR algorithm by taking the support value =10%

Table 2
Time taken for different confidence value in various algorithms

Time taken for different confidence(in secs)						
Confidence	Apriori	FP-Growth	Distributed	Genetic	TDTR	ITDTR
90	.0020	.0018	.0018	.0021	.0014	.0013
80	.0012	.0010	.0009	.0012	.0009	.0007
70	.0010	.0008	.0007	.0010	.0007	.0006
60	.0009	.0007	.0007	.0008	.0005	.0004
50	.0008	.0006	.0006	.0010	.0004	.0003

Table 2 can be represented as a graph. The following Figure 6. shows that our proposed approach ITDTR has taken less time for different confidence measure when compared with various algorithms.

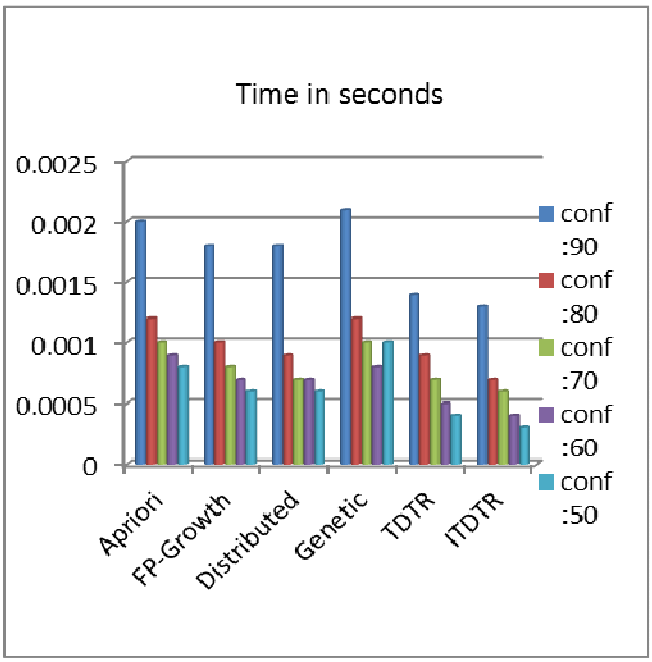


Figure 6
Efficiency graph

Some of the sample rules generated by ITDTR-Algorithm for confidence =90 and support=10 are shown in the following figure Figure7.

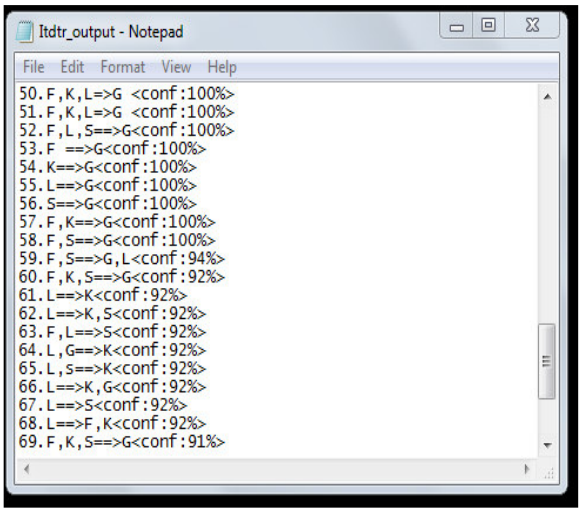


Figure7
Association Rules

From the above figure F,K,L,S,G are strongly associated with confidence 100%. 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 dataset.

CONCLUSION

In this paper we present an innovative ITDTR(Improved Two Dimensional Transaction Reduction) algorithm for mining frequent itemsets and to find association rules. 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 with higher accuracy .The efficiency of this algorithm is compared with Apriori algorithm, FP_Growth algorithm, Genetic algorithm , Distributed and TDTR algorithm.The results shows that our

proposed ITDTR algorithm provides higher efficiency and accuracy.In future, this work can be extended by combining various techniques like partition,distribution along with TDTR and the efficiency of this algorithm can be improved further.

REFERENCES

1. D.KeranaHanirex.,Dr.K.P.Kaliyamurthie. An Adaptive Approach For Mining Frequent Itemsets: A Comparative Study On Dengue Virus Type 1, IEEE International Conference on Human Computer Interaction , (2013)
2. D.KeranaHanirex. Dr.K.P.Kaliyamurthie . Mining Frequent Itemsets Using Genetic Algorithm, Middle-East Journal of Scientific Research ,19 (6): 807-810,(2014).
3. D.KeranaHanirex.,Dr.K.P.Kaliyamurthie. Finding the Dominating Amino Acids in Dengue Virus (Type-1) Study on mining frequent itemsets ,Int. Journal of Pharama and Bio Sciences, July; 4(3): (B) 880 – 889;(2013)
4. D.KeranaHanirex. An Efficient TDTR Algorithm for Mining Frequent Itemsets, International Journal of Electronics and Computer Science Engineering, V2(N1):251-256;(2012).
5. Frawley W.,Piatetsky Shapiro G., Matheus C. Knowledge Discovery in Databases An Overview, AI magazine, 213-228;(1992).
6. Jeffrey W., Seifert. Datamining:AnOver view , CRS report for Congress(2004).
7. Fayyad U.Datamining and Knowledge Discovery in Databases Implications from Scientific Databases,Proceedings of the 9th International Conference on Scientific and Statistical Database Management, ,2-11,(1997).
8. VenkatesanT.,Chakaravarthy.,Vinayaka Pandit and YogishSabharwai. Analysis of Sampling techniques for association rule mining,In proceedings of the 12thInternationalConference on Database Theory, 361:276-283;(2009).
9. R Agarwal, Imielinski ., A.Swami,Mining Association Rules between sets of Items in large Databases,ACM Sigmoid Conference on management of data,New York ,(1993).
10. Lijiuan Zhou.,Shuang Li.,Mingsheng Xu.Research on Algorithm of Association Rules in Distributed Database System,IEEE 2nd International Asia Conference on Informatics in Control,Automation and Robotics,216-219,(2010).
11. BA Mahafzah., AF Al-Badarneh., MZ Zakaria.A new sampling technique for Association Rule mining,Journal of Information Science Vol 35, 358-376,(2009).
12. V.Umarani., M.Punithavalli., An Emprical Analysis over the Four Different Methods of Progressive Sampling Based Association Rule Mining,European Journal of Scientific Research,66(4) :620-630,(2011).
13. Rakesh Agarwal., Ramakrishnan Srikant., Fast Algorithms for Mining Association Rules, 20th International Conference on VLDB, Santigo,(1994).
14. Thevar R.E., Krishnamoorthy R. A New Approach of Modified Transaction Reduction Algorithm for Mining Frequent Itemset, ICCIT ,11th conference on Computer and Information Technology,(2008).
15. Jong Soo Park,Ming-Syan Chen,Philip S.Yu,Using a Hash based method with Transaction trimming for Mining Association Rules,IEEE tranasctions on Knowledge and Data Engineering,9(5), 813-825;(1997).
16. D.KeranaHanirex.,Dr.K.P.Kaliyamurthie, Multi-Classification Approach for Detecting Thyroid Attacks,IJPBS, 4(3), (B) 1246 – 1251,July (2013).