Analysis of Improved TDTR Algorithm for Mining Frequent Itemsets using Dengue Virus Type 1 Dataset: A Combined Approach

D. Kerana Hanirex* and K. P. Kaliyamurthie
Department of Computer Science Engineering, Bharath University, Chennai – 600073, Tamil Nadu, India; keranahanirex.cse@bharathuniv.ac.in, kaliyamurthie.cse@bharathuniv.ac.in

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 algorithm which we have implemented in our previous research work.

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

1. 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”. In this paper we have presented an approach for mining frequent item sets 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.

2. Association Rule Mining
Discovery of association rules is an subfield in data mining1, 2. 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 item sets and then association rules using the frequent item sets3. 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

*Author for correspondence
Y is the ratio of the number of occurrences of \( \{x,y\} \) to the total number of transactions of \( D \). Confidence measure of an association rule \( X \Rightarrow Y \) is the ratio of the total occurrences for item \( X \) and \( Y \) to the total number of occurrence for item \( X \).

### 3. Related Work

Various algorithms have been proposed for association rule mining. Apriori is one of the famous basic algorithms for association rule mining\(^5\). It uses breadth-first search of the pattern. In Apriori algorithm, candidate item sets 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 item sets 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 item set. 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.

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-item set will not be required in the further database scans. AprioriTid algorithm\(^6\) is another method for improving the performance of association rules. This algorithm is used to construct the frequent item set. Thevar R. E; Krishnamoorthy proposed Modified Transaction Reduction based Frequent item set Mining Algorithm (MTR-FMA)\(^14\) which maintains its performance even at relative 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\(^5\).

### 4. 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 type 1 datasets**

![Figure 1. CSV File.](image)

The above figure 1 shows the CSV file. The following 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.
5. 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 \( \text{count1}[i] \)
   b. If the \( \text{count1}[i] \geq \text{min_sup} \) then put the transactions in to \( D_1 \)

2.a. FOR each \( I_i \in D_1 \) DO BEGIN
   a. count the number of transactions in \( \text{count2}[i] \)
   b. if \( \text{count2}[i] < \text{min_sup} \) then remove that \( I_i \) from \( D_1 \)

3. select sample \( S_i \) (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 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\(^8\).

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.

6. Dataset Description

This system uses dataset dengue virus type-1 data sets from GenBank: AAB27904.1 which consists of 777 amino acids\(^9\).

7. 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 type-1 dataset\(^10\). 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\(^11, 12\).

The accuracy of association rule is measured by the number of association rules generated\(^13\). The following Table 1 specifies the number of association rules generated for different confidence measures in various algorithms by taking the support value = 10%.

<table>
<thead>
<tr>
<th>Table 1.</th>
<th>Number of association rules generated for different confidence value in various algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conf.</td>
<td>Apriori</td>
</tr>
<tr>
<td>90</td>
<td>5</td>
</tr>
<tr>
<td>80</td>
<td>30</td>
</tr>
<tr>
<td>70</td>
<td>87</td>
</tr>
<tr>
<td>60</td>
<td>210</td>
</tr>
<tr>
<td>50</td>
<td>388</td>
</tr>
</tbody>
</table>

The following 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, TDTR\(^14, 15\).

![Figure 5. Accuracy Graph.](image)

The following Table 2 shows the efficiency of this Improved TDTR algorithm by taking the support value = 10%

<table>
<thead>
<tr>
<th>Table 2.</th>
<th>Time Taken for Different Confidence Value in Various Algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conf.</td>
<td>Apriori</td>
</tr>
<tr>
<td>90</td>
<td>.0020</td>
</tr>
<tr>
<td>80</td>
<td>.0012</td>
</tr>
<tr>
<td>70</td>
<td>.0010</td>
</tr>
<tr>
<td>60</td>
<td>.0009</td>
</tr>
<tr>
<td>50</td>
<td>.0008</td>
</tr>
</tbody>
</table>

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.

Some of the sample rules generated by ITDTR-Algorithm for confidence = 90 and support = 10 are shown in the following Figure 7.
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.

VIII. 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.

9. Future Enhancement

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.

10. References