An Analysis on Dengue Infection Using Association Rule Mining

Tanvi Upadhyay[1], Prof. Sushil Chaturvedi[2]
[1, 2]Dept of CSE
[1, 2] SRCEM Collages

Abstract- *Data mining is used to find out the meaningful information from the large dataset. FP-growth (frequent pattern growth) uses an extended prefix-tree (FP-tree) structure to store the database in compressed form. FP-growth adopts a divide-and-conquer approach to decompose both the mining tasks and the databases. In this paper, we use FP-growth algorithm to analyze the association rule of library circulation records the paper describes about the Association rule mining, FP-Growth and a Dengue Fever. Also the paper discuss about the reviews of research work done in this filed by diverse researchers, intellectual, association etc. This study is intended towards an association rule generation using in healthcare especially for the viral infective diseases.*

Keywords- Data Mining,KDD,Association rule Mining, Dengue Infection,FP-Growth.

I. INTRODUCTION

Data Mining is an important domain of computer science field for many reasons. Knowledge Discovery from Data (KDD) is a process in which selected target data is mined or discovered from the big data. Data mining is used to mine data patterns which are previously undiscovered, novel, valid and potential. Predictions and descriptions are important and focused goals of this field of interest. Data Mining confronting presence in the concentration of 1990's and accomplish an effective apparatus that is sufficient for getting prior obscure example and valuable data from gigantic dataset. Different reviews highlighted that Data Mining methods elevate the information holder to study and catch unsuspected relationship among their divulgence which thus well-off for declaration Making. [1]The word "Knowledge" in KDD alludes to the revelation of examples which are extricated from the handled information. An example is an articulation depicting certainties in a subset of the information. In this manner, the distinction amongst KDD and data mining is that "KDD alludes to the general procedure of finding learning from data while data mining alludes to utilization of calculations for extricating designs from data without the extra steps of the KDD procedure." DM avoid the disclosure of energizing skill, comprising of examples, establishments, changes, irregularities and considerable frameworks from expansive amounts of certainties put away in databases and other data vaults[2]

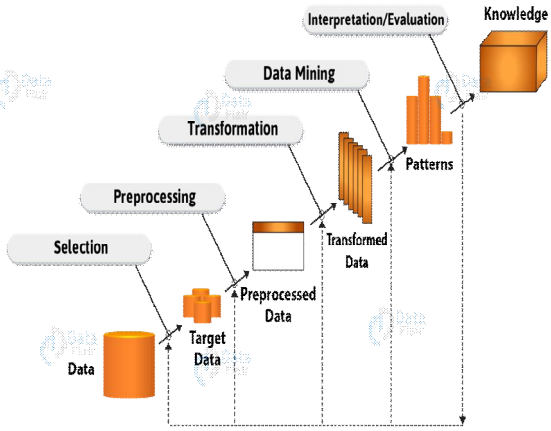

Figure: 1 KDD Process in DM [2]

Data Mining is a notable of the most prominent and persuading zones of research by the entire of the expectation of discovering moving data from gigantic front page new sets. [2] In Present time, Data Mining is well-suited mainstream in medicinal services exchange on the grounds that there is a favor of down to earth expository system for recognizing long shot and soak data in wellbeing information. In Medical market, Data Mining gives some portion of advantages one as identification of the rupture of trust in wellbeing safeguard, accessibility of restorative union to the patients at corrupt cost, recognition of reasons for illnesses and compassion of therapeutic help strategies. Data Mining confronting presence in the concentration of 1990's and accomplish an effective apparatus that is sufficient for getting prior obscure example and valuable data from gigantic dataset. Different reviews highlighted that Data Mining methods elevate the information holder to study and catch unsuspected relationship among their divulgence which thus well-off for declaration making. [3] Data mining refers to the extraction of predictive information that is in the hidden form from huge databases. Association rule mining is intended for the purpose of identifying the important rules that will be discovered inside the databases by performing some techniques of data mining and after using measures of interestingness. [4]

II. LITERATURE SURVEY

M. Sinthuja et al [2018] In this paper, the benchmark databases considered for comparison are Chess, Connect and Mushroom. It was found out that the IFP-Growth algorithm outperforms FP-growth algorithms for all databases in the criteria of runtime and memory usage.[5]

Yunlong Song et al [2011] In this paper, we use FP-growth algorithm to analyze the association rule of library circulation records. The results can make great sense to help to improve the quality of library collections. [6]

Lior Shabtay et al [2018] In this paper we present the GFP-growth (Guided FP-growth) algorithm, a novel method for finding the count of a given list of item-sets in large data. Unlike FP-growth, our algorithm is designed to focus on the specific multiple item-sets of interest and hence its time and memory costs are better. We prove that the GFP-growth algorithm yields the exact frequency-counts for the required item-sets. [7]

D. Kerana Hanirex et al [2015] 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. [8]

M.Bhavani et al The main objective of this study is to calculate the performance of various classification Techniques and compare their performance. The classification techniques used in this study are REP Tree, J48, SMO, ZeroR and Random Tree. The performance of classification techniques were compared by plotting graphs and table. Weka the data mining tool is used for the classification. [9]

Dr.Arun Kumar P.M et al [2017] The infection rates of Aedes Aegypti mosquitoes increase morbidity rate hence the decision tree is generated with the Aegypti rate as the root node and prevent further occurrences. The prediction of dengue infection carried out using Weka data mining tool and data mining techniques such as Decision tree and Support Vector Machine. Thus the model helps to predict the dengue cases earlier and reduce mortality rate. [10]

Ashwini Rajendra Kulkarni et al [2017] The paper describes about the Association rule mining and an Apriori Algorithm. Also the paper discuss about the reviews of research work done in this filed by diverse researchers, scholars, organizations etc. This paper is intended towards an association rule generation using in healthcare especially for the viral infective diseases. [11]

Ajinkya Kunjir et al [2016] This paper outlines the idea of predicting a particular disease by performing operations on the digital data generated in the medical diagnosis. In this project an efficient genetic algorithm hybrid with the techniques like back propagation and Naive Bayes approach for disease prediction is proposed. Bad clinical decisions would cause death of a patient which cannot be afforded by any hospital. To achieve a correct and cost effective treatment, computer technology Systems can be developed to make good decision. There is a lot of medical information unexplored, which gives rise to an important query of how to make useful information out of the data. [12]

K. Suguna et al [2017] In this paper we defined a framework for data preprocessing and pattern analysis using Apriori and FP-Growth algorithms. The Apriori algorithm preprocesses the data from the web log files. The FP-Growth algorithm extracts the frequent data from cleaned data. The appropriate analysis of a web server log proves that the websites works efficiently. [13]

III. DENGUE INFECTION

Dengue infection is vital disease caused by dengue germ, which extent in body of human by female mosquito. In late decades the danger of dengue disease has expanded drastically in tropical, as well as in sub-tropical areas. There are in the vicinity of 50 and 10 crores dengue diseases consistently, and more than 5 lakh cases are hospitalized. Dengue transmission is affected by an unpredictable arrangement of variables including nature, atmosphere and climate, human conduct and dengue infection serotype-particular crowd invulnerability among the human population. Dengue fever (DF) is one of the normal most life threading illnesses around the world particularly in India, which is prompted by the dengue infections (DENV) group offlavi-virus, principally transmitted from Aedes aegypti mosquito. As of late, transmission of viral invasion detailed in more than 100 nations with the simultaneous development of endemic territories transcendently in urban and country Regions has turned into a noteworthy general well being Concern. [14]

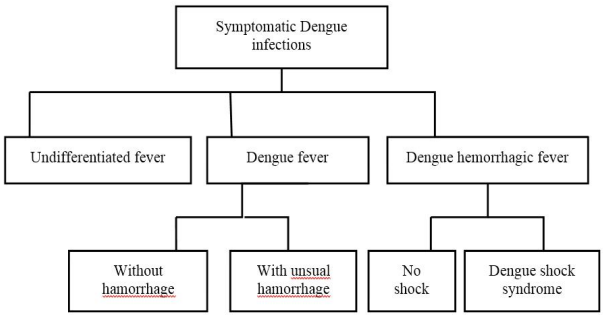

Figure: 2 Symptomatic Dengue infections [14]

With indications of headache, retro orbital pain, joint-pain, muscular pain and rush evidence. It is also known as bone breaking illness. Dengue infection has endangered 2.5 billion populations all around the world. Every year there are 50 million people who suffer from it globally. Dengue is divided into two types, i.e., type 1 and type 2, according to world health organization. First one is classical dengue called dengue fever and the other is dengue hemorrhagic fever. DHF1, DHF2, DHF3 and DHF4 are further four types of dengue hemorrhagic fever. DHF is revealed by start of fever which continues for 2 to 7 days with number of signs like leakage of plasma, shock and weak pulse. In earliest cases it's hard to differentiate dengue fever from dengue hemorrhagic fever. [15]

IV. MINING TECHNIQUES

A. Association Rule Mining

Mining association rule is the recent research area. Data mining is used to find the hidden pattern and useful information from the data base. Association rule mining is one of the major tasks of the Data mining. The other data mining tasks include clustering, classification. Frequent item set mining is not only used in market basket analysis but also used in bioinformatics such as gene expression data and protein analysis. The algorithm which is developed for market basket problem can also be applied to solve various bioinformatics problem such as analyzing frequent patterns in amino acids. Here each transaction is identified by the sequence of amino acids. Various algorithms have been proposed to find the frequent item sets but it differs in its computational efficiency. Frequent item sets can be converted into association rules that can be used in further applications. [16] Association rule learning is a predominant and thoroughly studied method for examining rules. Association rule mining. Discovery of frequent item sets is the most pivotal aspect of data mining. Here frequent item sets are sets of items that are purchased together frequently in a transaction. Mining frequent item sets are improved by numerous new methods and an agile data structure is also introduced. Numerous algorithms for frequent item set mining are discussed elaborately and the performance can be seen in the literature survey of frequent item set mining. [17]

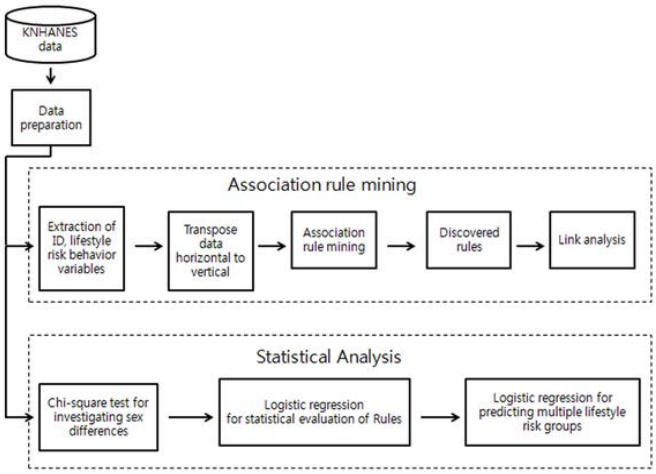

Figure: 3 Association Rule Mining[17]

Association rule mining is the process meant to discover the interesting relations that exist between the variables that form a part of large databases. This type of information is primarily used for the process of decision making in the marketing sectors for the promotional activities and placement of the products. These decisions also help to increase the sales of the consumer products in supermarkets etc. Now to note here the association rule mining does not take into account the order of the items either within the transaction or across the transaction unlike the sequence mining. [18]

Parameters of Association Rule Mining: Association rule learning finds out whether there exists any relation between the varied products or any frequent patterns. Popularly-known two of the constraints are minimum value of thresholds on two parameters are support and confidence to select interesting rules form dataset. Support and confidence are the parameters that are essentially used to conduct the association rule mining. Now here the term support refers to how frequent a particular item has occurred in the database while the confidence represents the count of the times the rule appeared to be true. Now for instance to let there be an item W and number of transactions denoted by T. So here the support of W will be the frequency of occurrence of item W in T as given in (1). It will be the proportion of the transactions present in our database that contain the item-set W.

Support (W) = Frequency of item sets (W) /Total number of transactions T……. (1)

And support for two items V and W occurring together is given by …….. (2)

Support (VW) = Frequency of item sets (V) and (W) occurring together /Total no. of Transactions…….. (2)

The parameter confidence is used as an indication of how often the rule was found to be true in our transaction database. Now here the confidence value associated with a rule, V) W, with respect to the set of transactions T will be the proportion of the transactions in our dataset that include V and W both simultaneously. Confidence is given by (3)

Confidence (V⟹W) = Support (VW)/Support (V)…… (3)

a) FP-Growth: We have designed an algorithm for FP Growth technique by taking into the consideration of the repeating patterns (items) in the frequent item-sets. In many cases, Apriori frequent set generate-check method significantly reduces the size of frequent sets with good performance. The algorithm just needs to examine the database once again. At the primary output of the database, the frequent l-item set will be created. At the second scan, the l-item sets are utilized to produce the FP-tree by filtering through infrequent items. The FP-tree contains all the frequent item sets, so the higher request frequent item sets can be mined from it. The mining of the FP-tree begins with each frequent pattern of length 1. The conditional pattern base is developed and after that, the (conditional) FP-tree is developed. At that point, the algorithm mines the tree iteratively to create the frequent patterns. But there are still some inherent weaknesses:

- It may produce a large number of frequent sets.
- It may need to scan the database repeatedly and Check a lot of frequent sets through pattern matching.

Using frequent-pattern growth, the FP-growth, can effectively address the issue. The main idea of the algorithm is to take the following divide and conquer strategy: compress the database which provide frequent sets to a frequent-pattern tree (or FP-tree), but still retain relational information of item sets; then divide this compressed database into a set of conditional database (a special type of project database), each associated with a frequent set, and do data mining on each database. Related concepts are as follows: FP-Tree: Sort data items in the transaction data table by support, then insert the data items in each Transaction into a tree with NULL as its root by Descending turn and record the support of each node occurs.

- Conditional pattern base: Contains the set of prefix Path which appears together with the suffix pattern set in the FP-Tree.
- Condition tree: Construct the conditional pattern base into a new FP-Tree according to the principles of the Formation of FP-Tree. [19]

b) Apriori Algorithm: Apriori is an association rule mining technique which when given the input of transactional databases it mines all frequently occurring items in the transaction. Here when given the Electronic Medical Record as the input to Apriori it then generates a set of risk factors that occur frequently and indicates those to be factors for developing diabetes. Apriori algorithm is, the most classical and important algorithm for mining frequent item-sets, proposed by R.Agrawal and R.Srikant in 1994. Apriori is used to find all frequent item-sets in a given database DB. The key idea of Apriori algorithm is to make multiple passes over the database. It employs an iterative approach known as a breadth-first search (level-wise search) through the search space, where k-item-sets are used to explore (k+1)-item-sets. The working of Apriori algorithm is fairly depends upon the Apriori property which states that" All nonempty subsets of frequent item-sets must be frequent". It also described the anti monotonic property which says if the system cannot pass the minimum support test, all its supersets will fail to pass the test. Therefore if the one set is infrequent then all its supersets are also frequent and vice versa. This property is used to prune the infrequent candidate elements. In the beginning, the set of frequent 1-itemsets is found. The set of that contains one item, which satisfy the support threshold, is denoted by L. In each subsequent pass, we begin with a seed set of item-sets found to be large in the previous pass. This seed set is used for generating new potentially large item-sets, called candidate item-sets, and count the actual support for these candidate item-sets during the pass over the data. At the end of the pass, we determine which of the candidate item-sets are actually large (frequent), and they become the seed for the next pass. Therefore, L is used to find L!, the set of frequent 2-itemsets, which is used to find L , and so on, until no more frequent k-item-sets can be found.[20]

c) OPUS: Opus is an efficient technique that functions recursively with respect to the parameters in the Left hand side and the right hand side. The algorithm considers the current left hand side, compares with

available left hand side then updates the current left hand side. The parameters can be constrained for left hand side and the right hand side. Unlike other association algorithms it monotonic doesn't require any parameter like support or confidence for Association rule mining. User can specify the maximum number of associations to be generated.

Current LHS: currently considered LHS of the rule

Available LHS: set of risk factors that can be added to the LHS of the rule. [21]

V. CHALLENGING ISSUES IN DATA MINING

A. Issues in Medical Data Mining

Human medicinal information is immediately the most fulfilling and troublesome of every single natural data to separate and analyze. Separating helpful information and giving logical choice - making for the analysis and treatment of illness from the database progressively gets to be distinctly essential. Data mining in solution can manage this issue.

B. Data Related Problems

The enthusiasm for frameworks for self-governing basic leadership in therapeutic and building applications is developing, as information is turning out to be all the more effectively accessible. In spite of the fact that the two territories solution and designing give off an impression of being remote as far as the hidden procedures, both face numerous normal difficulties. One of the issues important to both zones is a self-sufficient forecast. Since the therapeutic data is normal for excess, multi-attribution, inadequacy, and firmly related with time, medicinal information mining differs from others one. The real zones of heterogeneity of medicinal Information is:

- Volume and many-sided quality of medicinal information:- The medicinal properties of plant species have made an outstanding contribution in the origin and evolution of many traditional herbal therapies. These traditional knowledge systems have started to disappear with the passage of time due to scarcity of written documents and relatively low income in these traditions. Over the past few years, however, the medicinal plants have regained a wide recognition due to an escalating faith in herbal medicine in view of its lesser side effects compared to allopathic medicine in addition the necessity of meeting the requirements of medicine for an increasing human population.

- Doctor's understanding:- The Doctor is the title character in the long-running BBC science fiction television programme Doctor Who. Since the show's inception in 1963, the character has been portrayed by twelve lead actors.[note 1] In the programme, "the Doctor" is the alias assumed by a centuries-old alien—a Time Lord from the planet Gallifrey—who travels through space and time in the TARDIS, frequently with companions

- Affectability and specificity investigation :- Hospital epidemiology frequently involves evaluating new diagnostic tests. This is particularly relevant in the example of active surveillance for multidrug-resistant organisms (MDROs) in high-risk populations. Determining the sensitivity and specificity, and therefore effectiveness, of potentially new diagnostic tests requires these surveillance methods to be epidemiologically and statistically assessed. Advice on study design and biostatistical methods for evaluating diagnostic tests has recently been published, but there are additional practical concerns for hospital epidemiology and other disciplines where the incidence of disease is low that are only addressed in this editorial.

- Poor scientific portrayal:- Scientists have an image problem. Just ask any fifth-grader. Chances are, they'll probably tell you that a scientist is Caucasian, male, can be found wearing a lab coat, and leads a lonely laboratory existence . Perhaps he has eccentric character traits or odd-looking hair . That's some fairly discouraging news, but hey, what do kids know? The perceptions of adults are what really matter, right? Sadly, it seems that this stereotype is also held by many high-school students, college students, adults, and even scientists themselves.
- Sanctioned shape:- Sanctions restrict or prohibit the export of certain goods to some countries, individuals and entities. They are administered by the Department of Foreign Affairs and Trade. [22]

VI. CONCLUSION

Data Mining is turning into the most rising field in the healthcare sector in light of the fact that there is a need of proficient and powerful expository technique for searching complicated and valuable data in wellbeing information In our review, we have talked about data mining in the medical field.

Data mining are likewise utilized as a part of dengue fever prediction. Dengue fever (DF) is a mosquito-endured irresistible infection accompanied around by the infections of the classTogaviridae, subgenus Flavivirus. Additionally, Association rule mining has examined with various – diverse algorithms like – Association rule mining and FP-Growth.

REFERENCES

[1] Ajinkya Kunjir, Harshal Sawant, Nuzhat F. Shaikh, "A Review on Prediction of Multiple Diseases and Performance Analysis using Data Mining and Visualization Techniques", International Journal of Computer Applications (0975 – 8887) Volume 155 – No 1, December 2016.

[2] Mohammadian, M., "Intelligent Agents for Data Mining and Information Retrieval," Hershey, PA Idea Group Publishing, 2004 .

[3] Divya Tomar and Sonali Agarwal, " A survey on Data Mining approaches for Healthcare", International Journal of Bio-Science and Bio-Technology Vol.5, No.5 , pp. 241-266, 2013.

[4] Kaur J. , Madan N. (2015). Association Rule Mining: A Survey. International Journal of Hybrid Information Technology ,8(7),pp.239-242 .

[5] M.Sinthuja, Dr. N. Puviarasan, Dr. P.Aruna, "Research of Improved FP-Growth (IFP) Algorithm in Association Rules Mining", International Journal of Engineering Science Invention (IJESI),2018.

[6] Yunlong Song, Ran Wei, "Research on Application of Data Mining Based on FP-Growth Algorithm for Digital Library", 2011 IEEE.

[7] Lior Shabtay, Rami Yaari and Itai Dattner, "A Guided FP-growth algorithm for fast mining of frequent itemsets from big data", March 20, 2018.

[8] D. Kerana Hanirex and K. P. Kaliyamurthie, "Analysis of Improved TDTR Algorithm for Mining Frequent Itemsets using Dengue Virus Type 1 Dataset: A Combined Approach", Indian Journal of Science and Technology, Vol 8(31), DOI: 10.17485/ijst/2015/v8i32/87280, November 2015.

[9] M.Bhavani and S.Vinod kumar, "A DATA MINING APPROACH FOR PRECISE DIAGNOSIS OF DENGUE FEVER" Vol.(7)Issue(4), pp.352-359.

[10] Dr.Arun Kumar.P.M, Associate Professor, Chitra Devi.B, Karthick.P, Ganesan.M and 3.Madhan.A.S, "Dengue Disease Prediction Using Decision Tree and Support Vector Machine", SSRG International Journal of Computer Science and Engineering- (ICET'17) - Special Issue - March 2017.

[11] Ashwini Rajendra Kulkarni , Dr. Shivaji D. Mundhe, "Data Mining Technique: An Implementation of Association Rule Mining in Healthcare", International Advanced Research Journal in Science, Engineering and Technology, Vol. 4, Issue 7, July 2017.

[12] Ajinkya Kunjir, Harshal Sawant, Nuzhat F. Shaikh, "A Review on Prediction of Multiple Diseases and Performance Analysis using Data Mining and Visualization Techniques", International Journal of Computer Applications (0975 – 8887) Volume 155 – No 1, December 2016.

[13] K. Suguna, K. Nandhini, PhD, "Frequent Pattern Mining of Web Log Files Working Principles", International Journal of Computer Applications (0975 – 8887) Volume 157 – No 3, January 2017.

[14] Yoon Ling Cheong et al. , "Assessment of land usefactors associated with dengue cases in Malaysia usingBoosted Regression Trees", Spatial and Spatio-temporal Epidemiology 10, Published by Elsevier, pp. 75–84,2014.

[15] Kamran Shaukat DAar, "Dengue Fever Prediction: A Data Mining Problems ", Data Mining in Genomics & Proteomics,2016.

[16] S.N. Sinha et al. , "Adefovir dipivoxil—A possibleregimen for the treatment of dengue virus (DENV)infection",Chemometrics and Intelligent Laboratory Systems 155, Elsevier, pp. 120–127, 2016.

[17] D. Kerana Hanirex, K.P.Thooyamani and Khanaa, "performance of association rules for dengue virus type 1 amino acids using an integration of transaction reduction and random sampling algorithm", IJPSR,2017.

[18] M. Inbava lli(2015).Efficient Analysis of Frequent item set Association Rule Mining Methods. International Journal of Scientific & Engineering Research, 6(4).

[19] Manu Goel and Kaun Goel, "FP-Growth Implementation Using Tries for Association Rule Mining",Springer Nature Singapore Pte. Ltd. 2017.

[20] Charanjeet Kaur" Association Rule Mining using Apriori Algorithm: A Survey", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 2, Issue 6, June 2013.

[21] R. S. Chen, R. C. Wu and J. Y. Chen, "Data Mining Application in Customer Relationship Management Of Credit Card Business", In Proceedings of 29th Annual International Computer Software and Applications Conference (COMPSAC'05), Volume 2.

[22] J. Jain, S.K. Dubey, J. Shrinet, S. Sunil, "Dengue Chikungunya co-infection:A live-in relationship", Biochemical and Biophysical Research Communications,doi: 10.1016/j.bbrc.2017.02.008, 2017.