Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IMPACT FACTOR: 6.199

IJCSMC, Vol. 8, Issue. 5, May 2019, pg.32 – 37

ANALYSIS OF TECHNIQUES USED TO DISCOVER PATTERNS FROM DATASET FOR DISEASE PREDICTION

Mr. Anurag Rana[1]; Mr. Ankur Sharma[2]; Disha Pathania[3]
Department of Computer Science and Engineering
Arni University Kathgarh, (Indora)-176401 Kangra Himachal Pardesh

Abstract: ***Disease detection by the use of technology becomes need of the hour. Lack of time and ignorance causes the problems to increase substantially. Historical medical records of persons can be used to analyse the patterns and discover the disease if any or the future outcomes in terms of disease to the person. This paper presents the comprehensive review of techniques under pattern mining used to discover distinct patterns from the given dataset. In addition sequential pattern mining is considered base to predict the diseases and techniques like pattern growth, incremental growth, prefix span etc. are comparatively analysed giving advantages and disadvantages of each. In other words Apriori based algorithms are analysed using proposed literature. Future enhancements are also suggested using the proposed literature.***

Keywords: Sequential pattern mining, pattern growth, prefix span

1. INTRODUCTION

The critical approach of data mining used to discover normal and abnormal patterns from the database is sequential pattern mining. Data mining is the process of extraction of useful information from large database. The fetched information must be converted into user understandable form for future use. [1]Mining approaches used at different places vary according to size and complexity of problem in hand. Mining approaches useful for detecting patterns from the database includes web, text, sequential and temporal mining. Sequential pattern mining is the process of discovering patterns that are frequent within database. [2]The interest in pattern mining grown due to its ability to discover the hidden patterns within the database, that are useful for the users and cannot be extracted manually. Patterns category discovery is vital for successful interpretation of the disease.

The sequential pattern mining finds out frequent pattern from the sequence database. [3] The well-known pattern mining methods are utilized for web-log analysis, medical record analysis and disease prediction. It identifies strong symptom/disease correlations which can be valuable information for the diagnosis and preventive medicine.

There are various types of classification of sequential pattern mining algorithm that are based on following criteria:

- It consider the sequence that are generated and stored and minimize the number of sequences for decreasing the overall cost.
- It also supports the sequence of frequency that are counted and tested. The maintenance of count support has to be done to eliminate database and data structure.

The sequential pattern mining is divided into two kinds based on the above criteria:

- Apriori Based
- Pattern growth based

Apriori Algorithm:[4] It is used to analysis the sequence of pattern and depend largely on the property repeated patterns. It utilized the repeated pattern that test the subset of the item set. The features of apriori based algorithm are as given below:

- It describes the breath first search algorithm that constructs the sequences in iteration of algorithm and traverse the search space.
- It also uses the sequential pattern mining that generates candidate sequence and then test each one for satisfying specific condition and it consume a lot of memory.
- It also scans the original database for generating long list of candidate sequence and it require lot of processing time.

The apriori based algorithms are classified as given below:

- GSP: It identifies the patterns that are common within the large dataset are discovered using the algorithm and then anomalies are highlighted. Hence noisy data can efficiently handle by this algorithm. The data would be scanned to count

- SPADE: [3], [4]It is an algorithm that discover faster the sequential pattern. It utilizes equivalence classes that make a vertical list of pattern in database format. Using these patterns each sequence is associated with objects where it occurs. After that by using intersection on the list of ids frequent sequences are found. This method reduces the execution time as number of scans of database is reduced. In this algorithm, first of all frequencies of 1-sequence is computed that are associated with only one item. Secondly 2-sequences are counted by using transformation of vertical representation to horizontal in memory, after this the count of pair of items in bi-dimensional matrix is found. So in this step only one scan is performed. This would generate n sequences by joining n-1 sequences utilizing their id list. The number of sequences that appeared in an item can give the size of the list. If minsup is less than the size than sequence is one frequent. The overall calculation stops when there no sequence found. It can utilizes search method like BFS(breath first search) or DFS(Depth first search) for generating the new sequences.
- Pre-fix span: [5], [6] it explores prefix projection that mines set of patterns . It is an efficient growth method that reduces the effort to generate candidate subsequences. It leads to efficient processing as prefix-projection reduces the overall size of database. In our literature it is shown that prefix span is better than the GSP and another algorithm as it is efficient in large databases.

The proposed system works to reveals best possible mechanism to expose the patterns from the large dataset. The dataset can be derived from UCI machine learning website related to chronic diseases. Rest of the paper is organised as under: section 2 gives the literature survey of techniques used to discover patterns from dataset, section 3 gives the comparative analysis of techniques discussed in literature survey, section 4 gives the conclusion and future scope and last section presents references.

2. LITERATURE SURVEY

Alzahrani, (2016) proposed data mining method for disease prediction [7] for this purpose sequential data mining is used in order to accomplish this data preprocessing mechanism is applied. After applying preprocessing mechanism the attributes will be analyzed this will be done using passes on medical data. The first pass determines whether support for each disease is present or not at the end of this phase the frequent disease within the database will be identified, a counter will be maintained to count the occurrence of each disease within the dataset. Next phase determines the second sequence of diseases present within the dataset. The overall process yield the diseases which can cause the occurrence of other diseases. The disease resulting in another disease is termed as candidate generation. And for declaring that it is generated from the previous level Pruning is used.

Alamanda et.al. (2017), proposed sequence pattern mining in order to detect the time duration used for promotion [8] the sequence or pattern is checked from within the database. The weight of each sequence in each database is achieved from the interval of the successive element in the sequence and the mining is performed on the basis of weight considering time interval. Time interval based pattern is used in this case. In preprocessing missing values are not considered.

Ahmed (2017), proposed an application that utilizes the data mining technique to predict the heart disease[9]. Also it guide the patient to take treatment at early stage. But is completely dependent upon patient input and does not considered predefined dataset values. It also not utilizes the missing value that are essential to predict diseases.

Abbasghorbani et. al. (2015), conducted analysis of various pattern mining techniques are done and also the features of all the algorithms. It introduced various minimizing support counting which is used for minimizing search space[10]. We have generated small search space which will include earlier candidate sequence pruning then database is analyzed and compression technique is used to analyze.

Béchet et al. (2012), proposed paper presents the sequential pattern mining to discover the rare disease within human body where experiments are conducted using data mining tool WEKKA[11]. This show betterment in percentage for classification accuracy.

Chen et al. (2017), used a pattern growth method to analyze the medical database to specify the combination of chronic disease[6].It introduce prefix span algorithm that identify all possible patterns in the images but it constrained only specific disease and can further improved for efficient search, it shows the results in terms of HTN and DP diseases.

CHENG et al. (2017), proposed a sequential mining approach for early assessment of chronic disease[12]. The clinical database is considered .A dataset of patients derived from Taiwan, it derives richest of risk patterns. Data preprocessing as performed to rectify the problem if found but missing values are not considered .sequential pattern mining is used to observe the risk pattern and generate the result. The problem with this approach is that no precautions have been suggested. The classification accuracy is 80% further improvement in classification is needed. The chronic disease is analyzed in this paper built in over the existing problem.

Eenan (2009), proposed a non-homogeneous mark over model[13] which is used to identify the chronic disease in the patients. The algorithm uses global optimization that efficiently identify the number of frequent pathway required to analyses the patient. The result shows that the proposed methodology probability is better than existing ones but this approach can be extended using admission scheduling policy.

Ghosh et al. (2015), proposed a technique that extract sequential patterns from hypotensive patient groups[14]. These patterns are further utilized to inform medical decisions and randomized clinical trials. It further extended by including various clinical features and also include some sequential patterns. It also does not considered missing value during the preprocessing phase.

3. COMPARATIVE ANALYSIS OF TECHNQIUES USED FOR PATTERN DISCOVERY

In most of the existing literature problem with the pre-processing phase is discovered. Missing value handling mechanism is not optimized using the existing mechanism. This section presents comparative analysis of techniques to extract best possible mechanism for future enhancement.

AUTHORS	Technique	Advantage	Disadvantage	Future enhancement
Ahmed (2017)	Sequential Pattern mining for disease detection	Pre-processing mechanism is used to handle any problem with the extracted values from dataset	Missing values are not tackled using this approach	Missing values handling using clustering approach can be used along with this literature
Alzahrani, (2016)	Frequent patterns discovered from dataset using sequential pattern mining	Noisy data at pre-processing stage is tackled	Missing values causes the problem and classification accuracy is a problem	Missing values could be tackled at pre-processing stage using most probable clustering mechanism
Ghosh et al. (2015)	Sequential pattern mining for disease prediction	Useful patterns are extracted for predicting the disease at early stage	Missing data handling mechanism is missing	No clustering mechanism is employed that can be incorporated to accomplish greater classification accuracy

Eenan (2009)	Heterogeneous model for disease detection	Pre-processing mechanism is employed in order to form patterns with desired data only	Classification accuracy is compromised however execution speed is improved	Classification accuracy improvement using missing value handling
Béchet et al. (2012)	Rare disease detection using sequential pattern mining	Uncommon diseases are predicted with high accuracy	Missing values could cause classification accuracy to decay considerably	Classification accuracy improvement by using prefix-span algorithm

Table 1: Comparative analysis of techniques used for pattern discovery

From the comparative analysis it is concluded that pre-processing mechanism can be improved to overall improve the classification accuracy of disease prediction.

4. CONCLUSION AND FUTURE SCOPE

The disease prediction at early stage is the need of the hour. Database for disease detection could be of varying size. Discovering patterns out of the available database can be accomplished using pattern mining algorithms. There are number of algorithms which are discussed however each algorithm discussed suffers from missing value handling anomaly. Missing value handling can be accommodated using most probable value replacement mechanism. This mechanism uses the value repeated most number of times as most probable value which can be replaced with the missing value. By doing so classification accuracy can be improved during disease detection and prediction.

REFERENCES

[1] K. Uragaki, T. Hosaka, Y. Arahori, M. Kushima, T. Yamazaki, K. Araki, and H. Yokota, “Sequential pattern mining on electronic medical records with handling time intervals and the efficacy of medicines,” *Proc. - IEEE Symp. Comput. Commun.*, vol. 2016-Augus, pp. 20–25, 2016.

[2] J. Lee, U. Yun, and G. Lee, “Analyzing of incremental high utility pattern mining based on tree structures,” *Human-centric Comput. Inf. Sci.*, 2017.

[3] J. W. Huang, C. Y. Tseng, J. C. Ou, and M. S. Chen, “A general model for sequential pattern mining with a progressive database,” *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 9, pp. 1153–1167, 2008.

[4] J. D. Ren, Y. Dong, and H. T. He, “A parallel algorithm based on prefix tree for sequence pattern mining,” *Proc. - 2010 1st ACIS Int. Symp. Cryptogr. Netw. Secur. Data Min. Knowl. Discov. E-Commerce Its Appl. Embed. Syst. CDEE 2010*, pp. 6–11, 2011.

[5] M. Pater and D. E. Popescu, “The benefits of using prefix tree data structure in multi-level frequent pattern mining,” *SOFA 2007 - 2nd IEEE Int. Work. Soft Comput. Appl. Proc.*, no. 1, pp. 179–182, 2007.

[6] C. J. Chen, T. W. Pai, S. S. Lin, C. C. Yeh, M. H. Liu, and C. H. Wang, “Application of PrefixSpan Algorithms for Disease Pattern Analysis,” *Proc. - 2016 Int. Comput. Symp. ICS 2016*, pp. 274–278, 2017.

[7] M. Y. Alzahrani, “Discovering Sequential Patterns from Medical Datasets,” 2016.

[8] S. Alamanda, S. Pabboju, and N. Gugulothu, “An Approach to Mine Time Interval Based

Weighted Sequential Patterns in Sequence Databases," *2017 13th Int. Conf. Signal-Image Technol. Internet-Based Syst.*, pp. 29–34, 2017.

[9] F. Ahmed, "A Simple Acute Myocardial Infarction (Heart Attack) Prediction System Using Clinical Data and Data Mining Techniques," pp. 22–24, 2017.

[10] S. Abbasghorbani and R. Tavoli, "Survey on Sequential Pattern Mining Algorithms," *2015 2nd Int. Conf. Knowledge-Based Eng. Innov.*, pp. 1153–1164, 2015.

[11] N. Béchet, P. Cellier, T. Charnois, B. Cremilleux, and M. C. Jaulent, "Sequential pattern mining to discover relations between genes and rare diseases," *Proc. - IEEE Symp. Comput. Med. Syst.*, 2012.

[12] Y. CHENG, Y.-F. Lin, K.-H. Chiang, and V. Tseng, "Mining Sequential Risk Patterns from Large-Scale Clinical Databases for Early Assessment of Chronic Diseases: A Case Study on Chronic Obstructive Pulmonary Disease," *IEEE J. Biomed. Heal. Informatics*, pp. 1–1, 2017.

[13] B. R. M. Eenan, "Non-homogeneous Markov models for sequential pattern mining of healthcare data," pp. 327–344, 2009.

[14] S. Ghosh, M. Feng, H. Nguyen, and J. Li, "Hypotension Risk Prediction via Sequential Contrast Patterns of ICU Blood Pressure," *IEEE J. Biomed. Heal. Informatics*, vol. 20, no. 5, pp. 1416–1426, 2015.