Volume 8, Issue
October 2019

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Sequential Pattern Mining Analysis to Predict the Diseases

[1]Kanika Sharma, [2]Sanjeev Mahajan and [3]Ritika

[1] M.Tech Scholar, Beant College of Engineering and Technology, Gurdaspur
[2] Associate Professor Deptt. of Computer Science, Beant College of Engineering and Technology, Gurdaspur
[3] Assistant Professor Deptt. of Information Technology, Beant College of Engineering and Technology, Gurdaspur
E-mail: [1]kanikasharma030@gmail.com, [2]sanjeevmahajan@gmail.com, [3] ritikasood1987@gmail.com

ABSTRACT

The phenomenal advance in health and biotechnology produces a huge amount of data. Clinical data and high throughput information makes Electronic Health records (EHRs) expansive and complex. For the analysis of such huge amount of data, AI and data mining techniques have been utilized along with health services. Data mining is utilized to detect diseases using various informational datasets along with machine learning algorithms. There are many techniques available which are utilized for diagnosis of diabetes disease like FP growth, Apriori algorithm, etc. These techniques discover unknown patterns or relationships from a large amount of data and are utilized for making decisions for preventive and suggestive medicines. The main disadvantage of these techniques is that these discovers limited number of patterns. In this paper, we proposed a modified prefix span algorithm that discovers many patterns to detect diseases accurately. The results will help in predicting the diseases quicker and more accurately so that it leads to timely treatment of the patients.

Keywords: *Data mining, Prefix span, accuracy.*

1. INTRODUCTION

The critical approach of filtering of data set used to discover normal and abnormal patterns from the database is step by step analysis process. Filtering of data set is the process of extraction of useful information from the large database. The fetched information must be converted into user understandable form for future use. Mining approaches used at different places vary according to the size and complexity of the problem in hand. Mining approaches which are useful for detecting patterns from the database includes [1]web, text, sequential and temporal mining. Step by step analysis process is employed to discover patterns that are frequent within the database. The importance of pattern mining has grown because of its ability to discover the hidden patterns within the database [2] that are beneficial for the users and are almost impossible to extract manually. Patterns category discovery is vital for successful interpretation of the disease.

The step by step analysis process finds out frequent pattern from the sequence database. The well-known pattern mining methods[2] are utilized for web-log analysis, medical record analysis, and disease prediction. Achieving efficient patterns extraction from disease database in the least execution time is critical and for this purpose pattern mining algorithms are useful. Pattern mining algorithms are capable of extracting useful patterns that detect an abnormality and predict diseases. To detect patterns association rules are constructed. These association rules are capable of defining classes for abnormal pattern prediction. The sequential pattern is used to find a complete set of frequent subsequences. There are various methods for sequential pattern mining example Apriori-based and Pattern growth-based.

The Apriori-based algorithms are classified as given below:

- GSP (Generalized Sequential Pattern): It identifies the patterns that are common within the large dataset. GSP is a sequence pattern mining approach that is divided into multiple phases. In the first phase, all the single patterns discovered are counted. From the frequent patterns so discovered, candidate solutions are formed. Another pass is constructed to count the frequency of patterns in the previous phase. This process continues until all the distinct patterns from the database are extracted.

- SPADE(Sequential Pattern Discovery utilizing Equivalence classes): This method is used to achieve a vertical id list of the dataset [3] and then the intersection of IDs is obtained. This intersection is used for reducing the scan of the database and also decreasing overall execution time. It counts the sequence of each ID. Vertical representation is then converted into horizontal [4]. The algorithm stops when there is no sequence found. It utilizes breath first search and depth-first search to uncover the sequencing.

The pattern growth-based algorithms are classified as given below:

- PrefixSpan (Prefix-projected Sequential pattern mining): This method is used to explore the preprojection[4] in sequential pattern mining. It reduces the efforts of candidate subsequence generation by mining the complete set of patterns. Moreover, prefix-projection substantially reduces the size of the projected databases and leads to efficient processing.

2. RELATED WORK

CHENG et al. (2017), proposed a sequential mining approach [5]for early assessment of chronic disease. The clinical database is considered. A dataset of patients is

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

derived from Taiwan that has the richest of risk patterns. Data preprocessing is 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 results. The problem with this approach is that no precautions have been suggested. The classification accuracy is 80% and a further improvement in classification is needed.

Chi-jane Chen, et al. (2017), analyzed the Prefix span algorithm [6] that is suitable to identify various trajectory patterns in medical data. This medical database was used to detect chronic diseases among patients. This algorithm helped the doctors to easily identify the problem of the patient so that possible measures could be taken as soon as possible, before the next chronic disease could develop in the patient's body. The result of various chronic diseases is represented as statistical order.

Kunjir, et al. (2017), proposed multiclass Naive Bayes algorithm[7] that is used for prediction of a particular disease. The dataset used for operation is taken from the UCI machine learning website. The discussed approach deals with prediction accuracy corresponding to a particular disease. The result in terms of the confusion matrix is also presented.

Alamanda, et al. (2017), proposed sequence pattern mining[8] in order to detect the time duration used for promotion. The sequence or pattern is checked 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.

Alzahrani(2016), proposed data mining method for disease prediction. Sequential data mining [9]is used in order to accomplish the data preprocessing mechanism. After applying the preprocessing mechanism, attributes are analyzed 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 is identified and a counter is maintained to count the occurrence of each disease within the dataset. Next phase determines the second sequence of diseases presents within the dataset. The overall process yields 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.

Zhang, et al. (2016), proposed a technique named ConSgen[10] that is used to identify the contiguous sequential generator and also minimizes the redundant patterns, It utilizes the divide and conquers technique to find the sequential generator with contiguous constraints. But it does not consider the gapped alignments and also not discover the binding sites.

Patel et al. (2016), analyzed various sequential pattern mining algorithm. It discovers various challenges in these algorithms and improved the performance by proposing constraints in patterns. It enhances the existing CAI prefix span algorithm[11] by introducing time constraints. The comparative results showed that it is better and they can also further enhance it by applying efficient constraints.

Ghosh, et al. (2015), proposed a technique that extracts sequential patterns from hypotensive patient groups. These patterns are further utilized to inform medical decisions and randomized clinical trials[12]. It further extended by including various clinical features and also includes some sequential patterns. It also does not consider missing value during the preprocessing phase.

Abbasghorbani et al. (2015), analyzed various pattern mining techniques and the features of all the algorithms. It introduced various minimizing support counting [13]which is used for minimizing search space. They have generated small search space which includes earlier candidate sequence pruning. Then the database is analyzed with compression technique.

Manika et al. (2014), discussed three distinct algorithms and presented a comprehensive comparison between the three. The comparison is in terms of a number of iterations required to extract patterns out of the available dataset. Execution time[14] is the prime parameter on the basis of which this work is discussed. Pre-fix span algorithm performance is better as suggested through this literature.

3. PROPOSED SYSTEM

The proposed algorithm uses the prefix span algorithm for determining patterns which can be grouped together to form clusters. The pre-processing mechanism includes most probable value replacement with the missing value.

Algorithm

Input: Dataset
Output: Classification Accuracy, Disease Prediction
Input Dataset
Data=$Dataset_i$
Where 'i' are the number of rows within the dataset
Apply Pre-processing mechanism to resolve the missing values
MPV=mean(Values$Person_{id_i} = dataset(person_{id_i})$)
Repeat while all the missing values are tackled
If ($Missing_i$)
$Missing_i$=MPV
End of if
End of loop
Apply the Pre-fix span algorithm for pattern growth determination
Form clusters
Repeat until values in the dataset are examined

Volume 8, Issue
October 2019

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

If($Dataset_{iValue}$==$Dataset_{i+1value}$)
$Cluster_i$=$Dataset_{iValue}$
End of if
i=i+1
End of loop

- Predict disease looking at the pattern clusters
- Result: Accuracy, Disease.

4. PERFORMANCE PARAMETERS AND RESULT ANALYSIS

The performance of the system is analyzed by the use of parameters such as accuracy, specificity, and sensitivity.

Accuracy is obtained by dividing the number of true assessment by the number of all positive and negative assessment in terms of predictions, accuracy is obtained as:

$$Accuracy = \frac{TN+TP}{TN+TP+FN+FP}$$

Sensitivity is obtained by dividing the number of positive predictions to the total true positive rate.

$$\text{Sensitivity}=\frac{Correct_{Positive_{predictions}}}{Total_{Positives}}$$

Specificity is obtained by dividing the true negatives to the sum of true positive and false positive rate. This is used to evaluate the correctness of the proposed system. It is given as under:

$$\text{Specificity}=\frac{True_{Negitives}}{TP+FN}$$

RESULT COMPARISON IN TERMS OF ACCURACY, SENSITIVITY, AND SPECIFICITY ARE AS UNDER:

Image set name	*Parameters*	*Existing (%)*	*Proposed (%)*
Disease(Mild)	Accuracy	85	**95**
	Specificity	84	**94**
	Sensitivity	84	**92**
Disease (Moderate)	Accuracy	85	**95**
	Specificity	86	**96**
	Sensitivity	87	**97**
Disease (Severe)	Accuracy	86	**91**
	Specificity	87	**94**
	Sensitivity	87	**96**

Classification accuracy of the proposed system is more as compared to existing techniques. Multiple class prediction mechanism showing higher accuracy proving the worth of study.

Results and performance analysis as indicated through the plot shows that the proposed prefix span algorithm along with MPV algorithm yields better result.

The disease detection and prediction is given through accurate classification, result in terms of plots is given as under:

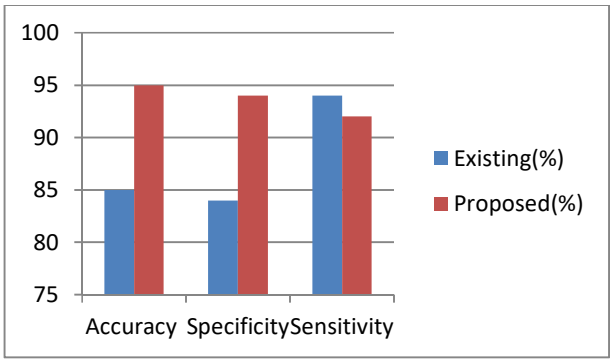

Figure 1: Mild Disease Detection

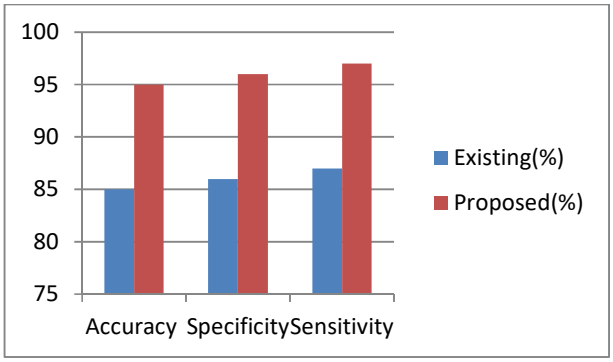

Figure 2: Moderate Disease Detection

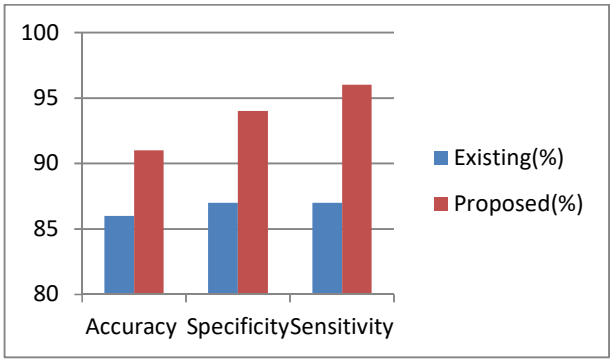

Figure 3: Severe Disease Detection

Following figures 5 and 6 shows comparison of Prefix Span Algorithm with Spade Algorithm:-

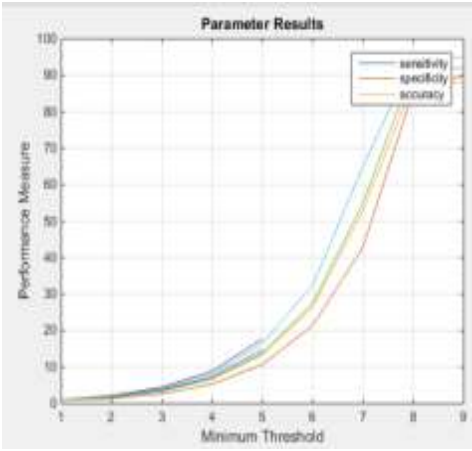

Figure 5: Prefix Span Algorithm

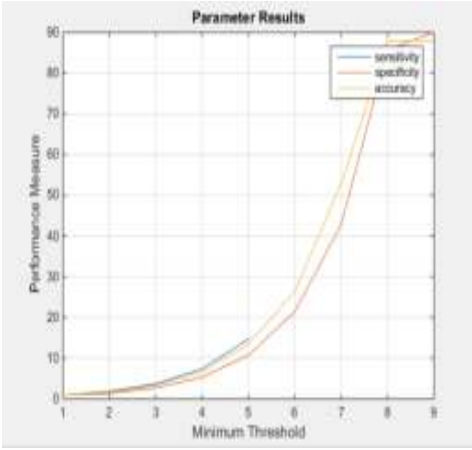

Figure 6: Spade Algorithm

5. CONCLUSION

The main objective of the proposed work is to do optimized detection using a prefix span for better accuracy. Obtained images are fed into the trained network for feature extraction using the prefix span algorithm and classification is performed using MPV along with the prefix span. The hybrid approach followed has given better results in terms of accuracy, specificity, and sensitivity. In future, the proposed strategy can be examined against the real-time datasets for better evaluation of accuracy.

REFERENCES

[1] E.Merlin.MercySDhivya, "Heart Disease Classification and Its Co- Morbid Condition Detection Using WPCA Weighted Principal Component Analysis and Genetic Algorithm," in *International Journal of Innovative Research in Computer and Communication Engineering (An ISO Certified Organization)*, 2007, vol. 3297, no. 6, pp. 7562–7568.

[2] 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.

[3] M. J. Zaki, "SPADE: An efficient algorithm for mining frequent sequences," *Mach. Learn.*, vol. 42, no. 1–2, pp. 31–60, 2001.

[4] J. Pei *et al.*, "PrefixSpan: Mining Sequential Patterns by Prefix-Projected Growth," *Proc. 17th Int. Conf. Data Eng.*, pp. 215–224, 2001.

[5] 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.

[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] A. Kunjir, H. Sawant, and N. F. Shaikh, "Data mining and visualization for prediction of multiple diseases in healthcare," *Proc. 2017 Int. Conf. Big Data Anal. Comput. Intell. ICBDACI 2017*, pp. 329–334, 2017.

[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] M. Y. Alzahrani, "Discovering Sequential Patterns from Medical Datasets," 2016.

[10] J. Zhang, Y. Wang, C. Zhang, and Y. Shi, "Mining contiguous sequential generators in biological sequences," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 13, no. 5, pp. 855–867, 2016.

[11] R. Patel and T. Chaudhari, "A Review on Sequential Pattern Mining using Pattern Growth Approach," pp. 4–7, 2016.

[12] 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.

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

[14] M. Verma and D. Mehta, "Sequential Pattern Mining: A Comparison between GSP, SPADE and Prefix SPAN," *Int. J. Eng. Dev. Res.*, vol. 2, no. 3, pp. 2321–9939, 2014.