Disease pattern recognition using modified prefix span algorithm

RITU VERMA
Research scholar
SSCET, BADHANI.

ISHA AWASTHI
ASSISTANT PROFFESOR
SSCET , BADHANI.

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

Keywords: data mining, spade, accuracy

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 large database. The fetched information must be converted into user understandable form for future use. Mining approaches used at different places vary according to size and complexity of problem in hand.[1] Mining approaches useful for detecting patterns from the database includes web, text, sequential and temporal mining. Step by step analysis process 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 step by step analysis process 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 step by step analysis process 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 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] in this method vertical id list of dataset is achieved and then intersection of ids has been obtained. This intersection is used for reducing scan of database and also decreasing overall execution time. It count the sequence of

each id and vertical representation are then converted into horizontal. The algorithm stops when there is no sequence found. It utilizes breath first search and depth first search to uncovering the sequencing.

- Pre-fix span: [5], [6]in this projection method is used to discover patterns and it uses sub sequences generation. It mine complete dataset into pattern using candidate sequence and then gives efficient processing of dataset.

In section 2 literature survey is given , section 3 describes about the proposed methodology , section 4 describes the results obtained and section 5 gives conclusion and future scope.

RELATED WORK

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.

CHENG et al. (2017), proposed a sequential mining approach for early assessment of chronic disease. 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[8]. 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.

Kunjir,et al.(2017), enhancement which can be improved proposed multiclass Naïve Bayes algorithm is used for prediction of particular disease but training it on set of data before implementation. This is downloaded from UCI repository work. The proposed system can help doctors to take clinical decisions where traditional decision support system fails, J47 algorithm is also used for proving the worth of study of accuracy in diabetes disease, breast cancer and diabetes approaches 83% by using this approach[9]. This accuracy requires in future.

Alamanda, et al.(2017), proposed sequence pattern mining in order to detect the time duration used for promotion .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[10]. Time interval based pattern is used in this case. In preprocessing missing values are not considered.

Ghosh,et al.(2015) Proposed a technique that extract sequential patterns from hypotensive patient groups. These patterns are further utilized to inform medical decisions and randomized clinical trials[11]. 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.

Zhang,et al.(2016) Proposed a technique named ConSgen that are used to identify the contiguous sequential generator and also minimize the redundant patterns, It utilizes the divide conquer technique to find the sequential generator with contiguous constraints[12]. But it does not considered the gapped alignments and also not discovered the binding sites.

M. Zihayat et al.(2016), it identified a problem of top –k utility based regulation pattern which is used to find out meaning in biology. Firstly proposed a utility model called TU-SEQ which is used to find top –K high utility gene regulation sequential patterns[13]. It is considering the relation between the various patterns and interactions in biological studies.

Sarac et al.(2017), Proposed a mining technique that are used to reduce the complexity and cost of the data storage. It divide chunks into separate parts and regression analysis are to be done to analysis the trial variable and samples datase. But it does not considered separate chunks for feature analysis and separate storage reservoir also not utilized.

Ahmed (2017), Proposed an application that utilizes the data mining technique to predict the diabetes disease. 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),In this paper the 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[14]. We have generated small search space which will include earlier candidate sequence pruning then database is analyzed and compression technique is used to analyze.

Proposed system

METHODOLOGY

The proposed algorithm uses the prefix span algorithm for determining patterns which can be grouped together to form clusters. 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($Person_{id_i} = dataset(person_{id_i})$))

Result comparison in terms of accuracy, sensitivity and specificity are given as under

- Repeat while all the missing values are tackled
  If ($Missing_i$)
  $Missing_i$=MPV
  End of if
  End of loop
- Apply Pre-fix span algorithm for pattern growth determination
- Form clusters
  Repeat until values in dataset are examined
  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.

RESULTS

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

Accuracy is obtained by subtracting the actual result from the approximate result. In terms of predictions accuracy is obtained as

$$Accuracy = \frac{Correct_{Pre}}{Total_{Pred}}$$

Equation 5: Accuracy in terms of prediction

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

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

Equation 6: Sensitivity evaluation formula

Specificity is another parameter used to evaluate correctness of the proposed system. It is given as under

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

Equation 7: Specificity obtaining formula

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

Image set name	Parameters	Existing (%)	Proposed (%)

Level 1 Diabetic(Mild)	Accuracy Specificity Sensitivity	85 84 84	95 94 92
Level 2 Diabetic(Moderate)	Accuracy Specificity Sensitivity	85 86 87	95 96 97
Level 3 Diabetic(Severe)	Accuracy Specificity Sensitivity	86 87 87	91 94 96

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

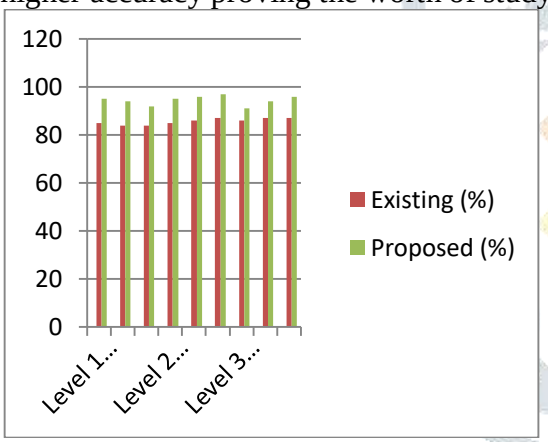

Figure 1 Confusion matrix

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

Proposed system

- **Spade algorithm**

Load Dataset

First of all data set is loaded from offline sources and dataset is synthetically prepared. After loading the data set the no. of patients are displayed in the box.

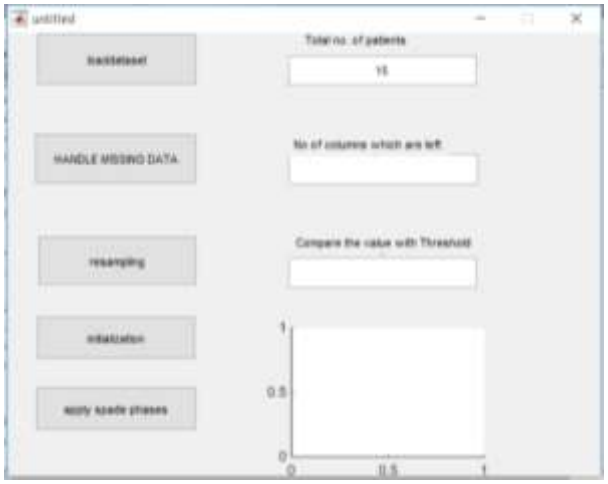

Figure 2 Load Dataset

Handle missing data:

For handling missing data MPV algorithm is used . It will eliminate noisy data from the dataset and display the remaining data.

Figure 3 Handle Missing data

Resample data

In this phase the repeated samples of original data has been create and statistical interference is used. It will generate approximate values. After that formatting of dataset is done using initialization phase.

Figure 4 Resampling of data

On applying spade:

The spade algorithm is applied to generate patterns and classification accuracy is improved. It will predict next temporal values.

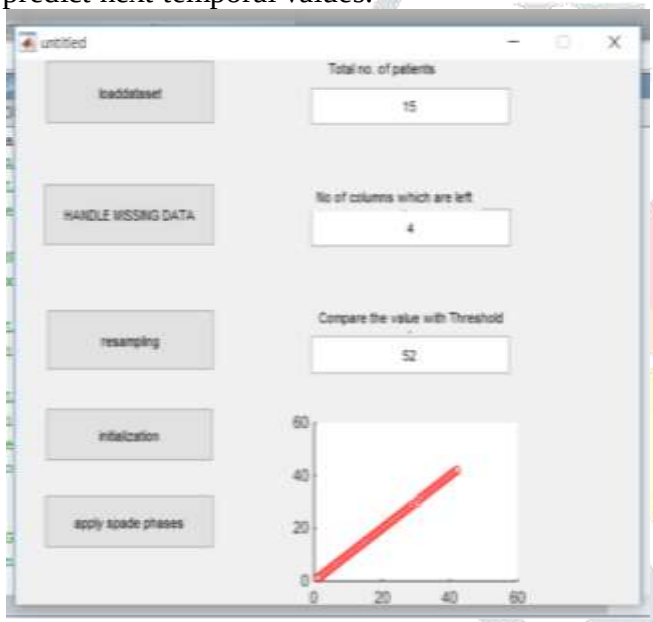

Figure 5 On applying Spade Algorithm

Parameter Result

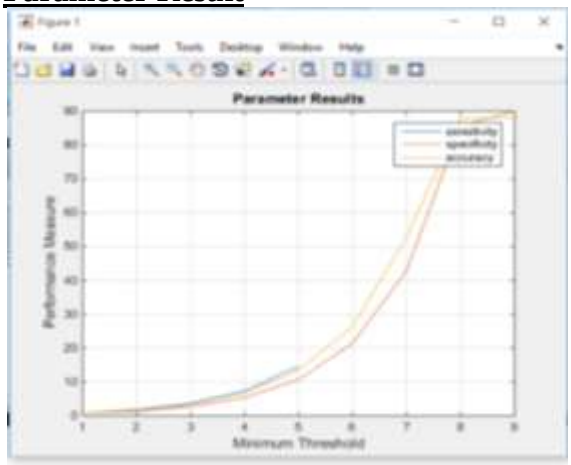

Figure 6 Comparison of specificity , sensitivity and accuracy.

CONCLUSION

In this paper an automated system that utilizes MPV along with prefix span algorithm for detecting diabetes proposed. Pre-processing phase is critical and is well defined using noise handling and resizing operation. Obtained images are fed into the trained network for feature extraction using prefix span algorithm and classification is performed using MPV. Hybrid approach followed gives better results. An effective pattern development technique, PrefixSpan, is proposed and contemplated in this paper. The main objective of the proposed literature is creating optimized detection using prefix span for better accuracy. Higher accuracy is achieved by the use of said literature. In future, proposed strategy can be examined against the real time datasets for better evaluation of accuracy.

REFERENCES

[1] M. E. Student, C. T. Nadu, and C. T. Nadu, "Heart disease classification and its co-morbid condition detection using WPCA genetic algorithm," pp. 287–291, 2016.

[2] K. Adlakha, "Recapitulation of Ant Colony and Firefly Optimization Techniques," *AIS*, vol. 1, no. 2, pp. 175–180, 2015.

[3] C. Anusha, S. K. Vinay, H. J. Pooja Raj, and S. Ranganatha, "Medical data mining and analysis for heart disease dataset using classification techniques," *Natl. Conf. Challenges Res. Technol. Coming Decad. (CRT 2013)*, pp. 1.09–1.09, 2013.

[4] I. Ţăranu, "Data mining in healthcare: decision making and precision," *Database Syst. J.*, vol. 5, no. 4, pp. 33–40, 2015.

[5] S. Sharma, "Data Preprocessing Algorithm for Web Structure Mining," pp. 1–5, 2016.

[6] S. D. Thepade, S. Vasaikar, N. Bhavsar, R. More, and A. Bhatkhande, "VEHICLE TRAFFIC DENSITY ESTIMATION USING BAYES , RULE , TREE FAMILY DATA MINING CLASSIFIERS APPLIED ON BACKGROUND SUBTRACTED TRAFFIC IMAGES," *IEEE ACCESS*, pp. 87–92, 2016.

[7] A. Adamov, "Data Analytics and Web Insights in area of Data Mining and Analytics."

[8] E. Trunzer, I. Kirchen, J. Folmer, G. Koltun, and B. Vogel-Heuser, "A flexible architecture for data

mining from heterogeneous data sources in automated production systems," *2017 IEEE Int. Conf. Ind. Technol.*, vol. d, pp. 1106–1111, 2017.

[9] N. R. Kasat and S. D. Thepade, "Novel Content Based Image Classification Method Using LBG Vector Quantization Method with Bayes and Lazy Family Data Mining Classifiers," *Procedia Comput. Sci.*, vol. 79, pp. 483–489, 2016.

[10] C. Zhang, Q. Yuan, and J. Han, "Bringing Semantics to Spatiotemporal Data Mining : Challenges , Methods , and Applications," pp. 5–8, 2017.

[11] A. A. R. Mining, "Mining of Market-Basket Data using MapReduce on YARN Framework," 2016.

[12] A. Alsa, "An Integrated Customer Relationship Management and Data Mining Framework for Customer Classification and Risk Analysis in Health Sector," pp. 41–46, 2017.

[13] J. Chandrasekaran, H. Feng, and Y. Lei, "Applying Combinatorial Testing to Data Mining Algorithms," pp. 253–261, 2017.

[14] F. Tian, T. Lan, K. M. Chao, N. Godwin, Q. Zheng, N. Shah, and F. Zhang, "Mining suspicious tax evasion groups in big data," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 10, pp. 2651–2664, 2016.

[15] B. V Kiranmayee, "A Novel Data Mining Approach for Brain Tumour Detection."

[16] D. Sumeet and P. Chowriappa, "Feature Selection and Extraction Strategies in Data Mining," *Data Min. Bioinforma.*, pp. 113–144, 2012.

[17] N. Gandhi and L. J. Armstrong, "A review of the application of data mining techniques for decision making in agriculture," pp. 1–6, 2016.

[18] E. Pinheiro, W. Weber, and L. Barroso, "Failure trends in a large disk drive population," *Proc. 5th USENIX Conf. File Storage Technol. (FAST 2007)*, no. February, pp. 17–29, 2007.

[19] A. Sharma and V. Mansotra, "Emerging applications of data mining for healthcare management - A critical review," *2014 Int. Conf. Comput. Sustain. Glob. Dev.*, pp. 377–382, 2014.

[20] K. Yan, X. You, X. Ji, G. Yin, and F. Yang, "A Hybrid Outlier Detection Method for Health Care Big Data," *2016 IEEE Int. Conf. Big Data Cloud Comput. (BDCloud), Soc. Comput. Netw. (SocialCom), Sustain. Comput. Commun.*, pp. 157–162, 2016.