Reliable Disease Prediction System Based on Association Rule Mining and Pyramid Data Structure

Y. Jeyasheela
Noorul Islam Centre for Higher Education
S. Vinila Jinny (vinijini@gmail.com)
Noorul Islam Centre for Higher Education https://orcid.org/0000-0003-2756-6736

Research Article

Keywords: Data mining, association rule mining, pyramid data structure.

Posted Date: July 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-666010/v1

Reliable Disease Prediction System based on Association Rule Mining and Pyramid Data Structure

Y. Jeyasheela[1], S. Vinila Jinny[2]

[1]Associate Professor, Department of Information Technology, Noorul Islam Centre for Higher Education, Kumaracoil, India.

[2]Associate Professor, Dept. of Computer Science and Engineering, Noorul Islam Centre for Higher Education, Kumaracoil, India.

[1]niceminime1est@gmail.com, [2]vinijini@gmail.com

Abstract – Data mining rules the world of data, as it is the base for analysing and relating the data with each other. Now-a-days, association rule mining concepts are applicable to almost all domains and the healthcare domain is in strong need of that. Taking this into account, this work attempts to propose a reliable disease prediction system with the help of association rule mining and pyramid data structure. This work processes a symptom dataset which is comprised of twelve health attributes. The health attributes with respect to four different diseases are clustered independently and are organised with the help of pyramid data structure. The process of clustering is carried out by Generalized Hierarchical Fuzzy C Means (GHFCM) and the strong association rules are built for predicting the disease. Finally, the effectiveness of disease prediction is evaluated with respect to the standard performance measures such as accuracy, precision, recall, F-measure and time consumption analysis. The performance of the proposed approach is observed to be satisfactory, which when compared to the existing techniques.

Keywords – Data mining, association rule mining, pyramid data structure.

1. Introduction

Due to the progression of technology, the medical world is modernized by employing numerous computer based techniques. The Computer Aided Diagnostic (CAD) systems are so popular these days, as it assists the healthcare professionals for better diagnosis of diseases. The CAD systems are imparted with sufficient knowledge, such that the abnormalities in the reports can easily be figured out. These suggestions are considered by the healthcare professional and the final diagnosis is made. This idea conserves time and increases the accuracy of the diagnosis. The main reason is that the CAD system is loaded with knowledge and it can relate the current instance with the stored instances.

As the healthcare professional makes final decision on the basis of the initial report of the CAD system, the accuracy of the diagnosis is very much improved [1,2]. These kinds of knowledge based medical systems have found an incredible place in the medical world. Most of the existing computer based medical systems aim to classify between the normal and the abnormalities being found in the medical data. However, these classification based systems do not consider the symptoms being encountered by the patient. This may reflect on the accuracy of the result, as the symptoms are disregarded.

Traditionally, the process of disease diagnosis is carried out by the healthcare professionals with the knowledge gained in their experience and analysis. However, it is certainly impossible to attain cent percent accuracy in diagnosis. At this juncture, computer based disease prediction system comes into picture, which supports the healthcare professional in attaining better diagnostic results. The medical disease prediction system deals with the human lives and hence, the precision and recall rates of the system are given utmost importance. The underlying point is that False Positive (FP) and False Negative (FN) rates must be as minimal as possible.

Now-a-days, all the medical records of the patients are stored in cloud to ensure better traceability. Hence, the disease prediction system considers the relationship between the data and the knowledge is gained from voluminous health records. In addition to the accuracy rates, it is necessary for the disease prediction system to provide results in a reasonable span of time. Recognising the importance of disease prediction system, this article presents a disease prediction system, which is based on association rule mining and classification.

This disease prediction system can be utilized as a preliminary scanning procedure for the users. The scope of this article is to predict the abnormality being found in the reading by means of association rule mining and the disease is classified. To achieve this, the proposed disease prediction system is divided into clustering phase and disease prediction phase.

In the disease prediction phase, the system can classify between the diseases such as obesity, diabetes, blood pressure, cardiovascular disease etc. Initially, a database with medical records and diagnostic results are passed on to the system. The numerical values of medical attributes are normalized, such that all the values lie in the range of 0 to 1. Each medical attribute is then clustered by means of Generalized Hierarchical Fuzzy C Means (GHFCM) algorithm and the frequent pattern is computed. The major contributions of this work are as follows.

- The proposed disease prediction system can predict about four different diseases such as obesity, diabetes, blood pressure and cardiovascular disease. On the other hand, most of the existing systems are meant for classifying between positive and negative cases.
- This approach performs clustering operation over disease and symptom based on the input dataset.
- The proposed approach consumes reasonable time, as the data is organised by pyramid data structure and the similar data are clustered together.
- The proposed work shows better precision, recall, F-measure rates, while consuming reasonable time.

The rest of the paper is organized as follows. The review of literature with respect to disease prediction system is presented in section 2. The proposed disease prediction system is described in section 3 and the performance of the proposed system is analysed in section 4. Finally, the conclusions of this paper are summarized in section 5.

2. Review of Literature

This section attempts to review the state-of-the-art literature with respect to disease prediction system.

Protein phosphorylation has close association with the diseases and many researchers are focussing to analyse the phosphorylation for disease prediction. In [3], the phosphorylation sites are processed by means of combined feature selection technique based on Support Vector Machine (SVM). This technique involves minimal redundancy with maximal relevance. However, this system is so complex and it requires some background knowledge about phosphorylation.

In [4], an ultra low power with a secure Internet of Things (IoT) platform is presented for predicting cardiovascular diseases. The disease prediction is carried out with the help of Electro CardioGram (ECG) signals. This work requires prior knowledge about ECG signal processing and involves computational complexity.

A liver fibrosis prediction system for hepatitis patients is proposed in [5]. This work utilizes two soft computing approaches and compares the performance of the two techniques. The techniques employed are Fuzzy Analytical Hierarchy Process (FAHP) and Adaptive Neuro-Fuzzy Inference System (ANFIS). Both the soft computing approaches are trained and tested for predicting the disease. Finally, it is concluded that the performance of both the approaches are better.

In [6], a collective prediction of disease associated miRNAs is presented based on transductive learning. This work considers the similarities between the diseases and the association between miRNAs and diseases for constructing the network. The relevance score is computed by means of transductive learning and the network is upgraded in an iterative fashion. The performance of the proposed approach is tested in terms of precision, recall and F-measure rates. Yet, this work is based on miRNA and hence, prior knowledge is required to deal with the data.

A classification approach for predicting the chronic disease hospitalization over the stored electronic health records is presented in [7]. This work focuses on predicting two different heart diseases such as heart disease and diabetes. The classification problem is fixed as binary and utilized different machine learning algorithms to classify between the diseases. However, this work does not organize the data for performing classification.

An optimized disease ranking system is proposed in [8], which identifies the rare diseases being occurred in humans. Initially, this work passes a symptom dataset and the Orphanet is utilized to extract the information about rare diseases. With the aid of this database, a website is designed and connected to this database. This work focuses only on the rare diseases and it consumes more time to deal with various symptoms.

In [9], a dynamic detection system that deals with parkinson's disease is proposed on the basis of deep brain simulation. This work employs the dynamic feature extraction and classification processes for detecting the parkinson's disease. The dynamic feature extraction and classification is made possible by processing feature extraction, dimensionality reduction and classification algorithms. This work extracts the features by means of Discrete Wavelet

Transform (DWT) for extracting the features, the dimensionality is reduced by Maximum Ratio Method (MRM) and the dynamic k-Nearest Neighbour (k-NN) classifier is employed for classification.

In [10], a telehealth environment is proposed by coupling the Fast Fourier Transformation (FFT) with the ensemble machine learning technique for recommending the patients with heart disease. The input is given to the system by means of time series data and the input is divided by applying FFT for extracting the frequency information. The ensemble classifier based on bagging technique is utilized for predicting the patient's condition.

In [11], several machine learning approaches are compared against each other for the sake of predicting advanced liver fibrosis in chronic hepatitis C patients. The classification models are constructed by clubbing the serum biomarkers and clinical information. However, this work is meant for a specific disease and does not work as a suggestion system for the patients. In [12], an intelligent hybrid framework is proposed for predicting parkinson's disease. This work predicts the parkinson's disease by means of Support Vector Machine (SVM) and Bacterial Foraging Optimization (BFO). However, this work is meant for predicting Parkinson's disease alone.

In [13], the performance of neural networks is compared against decision trees for predicting the diabetes mellitus. A prediction model for diagnosing diabetes based on artificial metaplasticity with the help of multilayer perceptron is presented in [14]. In [15], a hybrid disease prediction model is proposed to predict diabetes.

Motivated by these works, this article aims to propose a disease prediction system that can predict four different diseases with better accuracy, precision and recall rates. The following section describes the proposed approach in detail.

3. Proposed Disease Prediction System

The proposed disease prediction system is explained in this section in addition to the overview of the work.

3.1 Overview of the Work

The disease prediction system assists the healthcare professional for attaining better diagnosis. Most of the existing disease-predicting systems consider one or two diseases and involves more computational complexity. In addition to this, the data organization is not performed in most of the existing disease prediction systems. The proposed approach follows a simple approach that employs data organization and classification for better prediction of diseases with the common symptoms.

Several disease prediction systems utilize miRNAs for predicting diseases, which strongly requires prior knowledge and the system involves more computational complexity. Considering this fact, the proposed disease prediction system performs analysis on the symptoms being experienced by the patient and suggests the disease, which helps the healthcare professional to attain better diagnosis. The proposed approach attains the goal by two key steps, which are clustering and disease prediction. The overall architecture of the proposed work is depicted in figure 1.

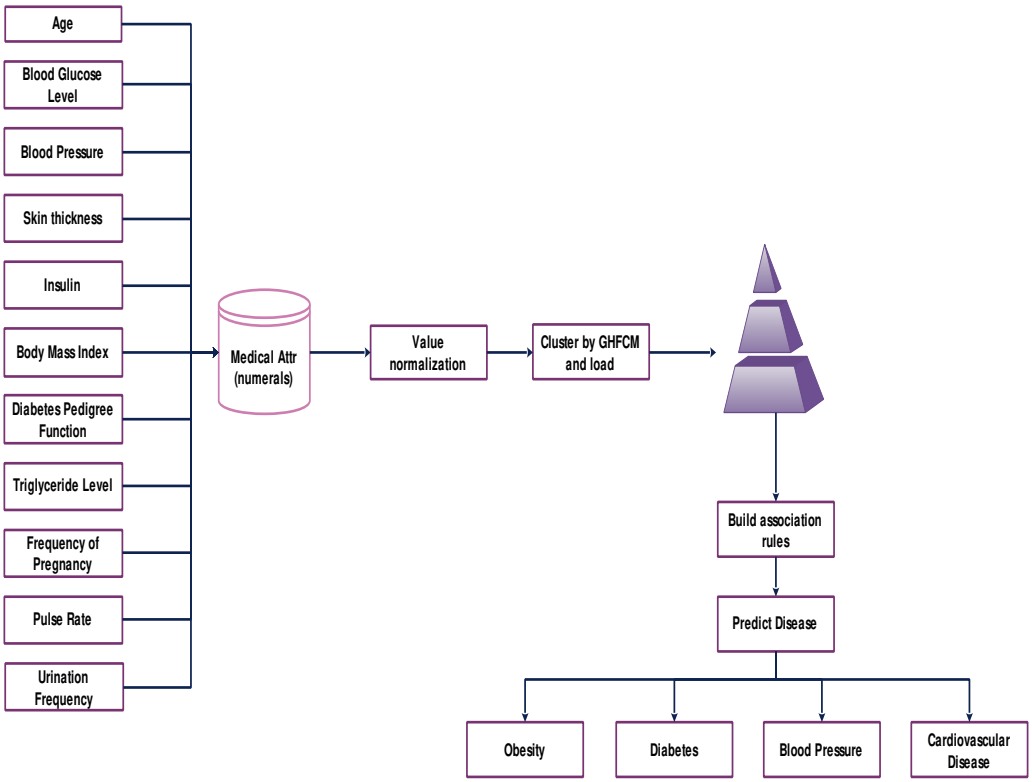

Fig.1. Overview of the proposed approach

As soon as the symptoms are passed to the system, it pre-processes the medical data and the values are normalized. The normalized values are then clustered by means of GHFCM algorithm. The association rules are then formed and the disease prediction is made. Finally, the disease prediction accuracy, precision, recall and time consumption rates are measured and the performance of the proposed approach is satisfactory. The proposed approach is presented as follows.

3.2 Disease prediction system

The core functionality of this work is attained in two significant steps, which are clustering and disease prediction. The association rules are formed for effective prediction of the disease. The symptoms encountered by the patient is loaded onto the disease prediction system and based on the symptoms, the diseases are predicted and the outcomes are provided to the healthcare professional. Initially, the dataset is pre-processed such that the dataset is fit for the forthcoming operations.

3.2.1 Data Pre-processing

The proposed disease prediction system predicts between four different diseases such as obesity, diabetes, blood pressure and cardiovascular disease. Hence, this work tailors the existing Puma Indian Diabetes dataset [16], which contains eight attributes such as Frequency of pregnancy, glucose level, blood pressure level, skin thickness, insulin, BMI, diabetes pedigree function and age. Apart from this, the dataset for the proposed work is added with certain other attributes such as triglyceride level, frequency of urination and pulse rate.

Initially, the dataset is analysed with respect to the range of values of each and every health attribute. The available range of all the health attributes is tabulated in table 1. From the table, the maximum and the minimal values of the attribute are known. Additionally, the missing values and the redundant records are removed from the dataset for better execution ability of the disease prediction system.

Table 1: Range of Health Attributes Present in the Dataset

Sl.No	Health Attributes	Data type	Min	Max
1	Frequency of Pregnancy	Number	0	17
2	Glucose Level	mg/dL	0	199
3	Blood Pressure Level	Sys/dias	0	122
4	Skin Thickness	mm	0	99
5	Insulin	mIU/L	0	846
6	BMI	Number	0	67.1
7	Diabetes Pedigree Function	Number	0.078	2.42
8	Age	Number	21	81
9	Class	Number	0	1
10	Triglyceride Level	mg/dL	150	499
11	Pulse rate	Number	60	100
12	Frequency of urination	(ml/day)	800	2000

The above presented table indicates the maximal and minimal values of the utilized health attributes. The standard Puma Indian dataset specifies the class by considering the eight health attributes. In addition to this, the proposed work considers three more attributes such as triglyceride level, pulse rate and frequency of urination. These three additional attributes are supplementary and are utilized to predict diseases other than diabetes. After the completion of initial level analysis of the data, the proposed approach attempts to normalize the data by applying the below given formula.

$$Norm_x = \frac{(x-ac_l)\times(n_h-n_l)}{ac_h-ac_l} \quad (1)$$

In the above equation, x is the value of the health attribute which is to be normalized. ac_l and ac_h are the least and highest values of the attribute, as given in table 1. n_h and n_l are the preferable range of values for performing normalization. The reason for normalizing the values is that all the health attributes fall under same range, which helps in easy data processing. In our case, the values of n_h and n_l are 1 and 0. This step is followed by the process of clustering operation, as presented in the following section.

3.2.2 Clustering by GHFCM

It is easy to perform clustering operation, when the values are normalized. Hence, the each and every health attribute are grouped by means of kFCM. The traditional FCM algorithm intends to cluster the data items based on the degree of similarity. kFCM algorithm can perform well with data with greater dimension. The objective function of kFCM is presented in equation 2.

The clustering approach of the proposed disease prediction system utilizes the algorithm GHFCM proposed in [17], which combines the Hierarchical FCM (HFCM) [18] and Generalized FCM (GFCM) [19] together. The advantages of HFCM and GFCM are automatically inherited to GHFCM. The standard objective function of GHFCM is presented below.

$$Ob_f = \sum_{i=1}^{N}\sum_{x=1}^{X}\sum_{y=1}^{Y} a_{ix}^{g} b_{ixy}^{h} \sum_{d\in NH_i} l_d m_{dxy} \quad (2)$$

This equation can be represented as in equation 3.

$$Ob_f = \sum_{i=1}^{N}\sum_{x=1}^{X}\sum_{y=1}^{Y} a_{ix}^{g} b_{ixy}^{h}\, l_d \left|\left|k_c - \mu_{xy}\right|\right|^2 \quad (3)$$

Where $i = \{1,2, \ldots N\}$ is the dataset, which contains N data items. X is the count of clusters, Y is the count of subclasses. The degree of membership of k_i in x^{th} cluster is represented by a_{ix} and g is the weight exponent of the fuzzy membership function a_{ix}. b_{ixy} is the sub-membership that satisfies the conditions $\sum_{x=1}^{X} a_{ix} = 1$ and $\sum_{y=1}^{Y} b_{ixy} = 1$. l_d is the weighing factor which controls the impact over the distance between the corresponding and the centre pixel. NH_i is the neighbourhood of the i^{th} pixel. k_c is the image intensity and μ_{xy} is the cluster centre. m_{dxy} is the sub-distance function, which is Euclidean distance. The following equations present the computation of a_{ix}, b_{ixy} and μ_{xy}.

$$a_{ix} = \frac{\left(\sum_{y=1}^{Y} \sum_{d \in NH_i} l_d b_{ixy}^{h} m_{dxy}\right)^{\frac{1}{(1-g)}}}{\sum_{p=1}^{X}\left(\sum_{y=1}^{Y} \sum_{d \in NH_i} l_d b_{ipy}^{h} m_{dxy}\right)^{\frac{1}{(1-g)}}} \quad (4)$$

$$b_{ixy} = \frac{\left(\sum_{d \in NH_i} l_d a_{ix}^{g} m_{dxy}\right)^{\frac{1}{(1-h)}}}{\sum_{p=1}^{Y}\left(\sum_{d \in NH_i} l_d a_{ix}^{g} m_{dxy}\right)^{\frac{1}{(1-h)}}} \quad (5)$$

The centroid of the cluster μ_{xy} is computed as follows.

$$\mu_{xy} = \frac{\sum_{i=1}^{N} \sum_{d \in NH_i} a_{ix}^{g} b_{ixy}^{h} K_C}{\sum_{i=1}^{N} a_{ix}^{g} b_{ixy}^{h}} \quad (6)$$

Thus, the local weighted generalized mean that takes the spatial and cluster information into account, is computed. The altered membership and the sub-membership is given by

$$a_{ix} = \frac{\sum_{d \in NH_i} l_d a_{dx}}{\sum_{p=1}^{X} \sum_{d \in NH_i} l_d a_{dp}} \quad (7)$$

$$b_{ixy} = \frac{\sum_{d \in NH_i} l_d b_{dxy}}{\sum_{p=1}^{Y} \sum_{d \in NH_i} l_d b_{dxp}} \quad (8)$$

By this way, the symptoms of the disease are clustered by GHFCM, which employs the fuzzy objective function that takes the hierarchical distance function and spatial constraints into account. This improves the robustness and efficiency of the clustering algorithm. As soon as the disease symptoms are clustered, the clusters are organized in the pyramid data structure.

As this work distinguishes between four diseases, each and every disease is represented by a single tier. This means that a tier involves two clusters, which are positive and negative. This idea makes it easy to perform search and retrieve operation. In addition to this, easy localization and perfect data organization is possible with this model.

The first tier is loaded with the common symptoms of diabetes that includes the first nine health attributes of table 1. The second tier is meant for obesity, which is based on the attributes BMI and triglyceride level. The blood pressure rate along with the pulse rate is considered as the basic symptom for diagnosing cardiovascular disease. As each tier is dedicated for a single disease, the clusters with positive and negative range of health attributes are easily differentiated. Besides this, it is easy for the system to construct the association rules and is presented as follows.

3.3 Association Rule Formation

Let x_i and y_i be the health attributes and the support of the association rule in the association rule set $RS(HA, DP)$ is presented in equation (9). Here, HA is the health attribute and DP is the disease prediction. The support of association rule $x_i \Longrightarrow y_i$ is computed by

$$S(x_i \Longrightarrow y_i) = \frac{|x_i \cup y_i \subset RS(HA,DP)|}{RS(HA,DP)} \quad (9)$$

The support of the association rule indicates the ratio of the count of disease prediction records with both health attributes (symptoms) to the total count of records.

The confidence of the association rule is computed by the following equation and is represented as $C(x_i \Longrightarrow y_i)$.

$$C(x_i \Longrightarrow y_i) = \frac{S(x_i \Rightarrow y_i)}{S(x_i)} \tag{10}$$

Equation 10 can be written as follows.

$$C(x_i \Longrightarrow y_i) = \frac{|x_i \cup y_i \subset RS(HA,DP)|}{|x_i \subset RS(HA,DP)|} \tag{11}$$

The confidence of the association rule is computed by dividing the total count of disease prediction records that contain both the health attributes (symptoms) by the number of association rules with the health attribute x_i.

Suppose, if any association rule encounters the support and confidence values greater than the support and confidence thresholds, then that rule is added to the frequent itemset. The rule must satisfy both support and confidence, such that it is included in the frequent itemset. This work sets the support and confidence thresholds as 2 and 50% respectively. Based on this idea, the strong association rules are constructed and the disease prediction is performed. The overall algorithm of this work is presented as follows.

Disease Prediction Algorithm

Input : Symptom dataset, S_{thr} *and* C_{thr}*;*
Output : Disease prediction
Begin
Pre-process the data;
Cluster the data;
Load into pyramid data structure;
Compute $S(x_i \Longrightarrow y_i)$ *by eqn. 9;*
Compute $C(x_i \Longrightarrow y_i)$ *by eqn.10;*
If $S(x_i \Longrightarrow y_i) > S_{thr}$ && $C(x_i \Longrightarrow y_i) > C_{thr}$
Add the rule in the frequent itemset;
End if;
End;

This disease prediction algorithm can be utilized to detect any number of diseases based on which the input dataset has to be given. When a patient enrols and enters the symptom with corresponding values, the proposed disease prediction system checks the pyramid tier by tier. The first tier is meant for obesity, hence the symptoms triglyceride level and BMI are considered to check for the presence of the disease.

The second tier is meant for diabetes that considers eight different symptoms of the dataset and frequency of urination in addition to it. The third tier is meant for Blood Pressure (BP), which is predicted by a single value itself as low and high BP. The final tier is meant for cardiovascular disease, which considers the pulse rate, blood pressure and obesity for

predicting the cardiovascular disease. By this way, the diseases are predicted and the outcome of the proposed work is analysed in the following section. The performance of the proposed approach is compared with the existing approaches and the results are presented as follows.

4. Results and Discussion

The performance of this work is analysed on a stand-alone computer with 8 GB RAM and intel i7 processor. The capability of the proposed approach is tested by simulating the work using Java in Netbeans platform. The performance of the proposed approach is tested in terms of accuracy, precision, recall, time consumption and F-measure rates. The outcomes of the proposed disease prediction system are compared with the existing techniques proposed in [13-15].

The performance of the proposed approach is tested over the Puma Indian Diabetes dataset, which contains 769 entities with eight attributes. In addition to the existing attributes, three more attributes are added to the dataset. Initially, the accuracy, precision, recall and misclassification rate of the proposed approach is tested and compared with the existing approaches. The disease-wise prediction accuracy, precision, recall and misclassification rate are also tested. The following equations present the formulae for computing the performance metrics.

$$Acc_{rate} = \frac{TP+TN}{TP+TN+FP+FN} \tag{12}$$

$$Pr_{rate} = \frac{TP}{TP+FP} \tag{13}$$

$$R_{rate} = \frac{TP}{TP+FN} \tag{14}$$

$$F_m = \frac{2*Pr_{rate}*R_{rate}}{Pr_{rate}+R_{rate}} \tag{15}$$

In equations (12-15), TP is the true positive, TN is the true negative, FP is the false positive and FN is the false negative rates. In this case, TP is the correctly predicted entities and TN is the correctly rejected entities. FP is the incorrectly predicted disease, such that it shows the positive to the unaffected disease and FN is the incorrectly predicted disease that disease prediction system claims negative to the affected disease. Hence, both false positive and false negative rates are quite dangerous, as this system is involved with human lives. The performance of the proposed approach is presented as follows.

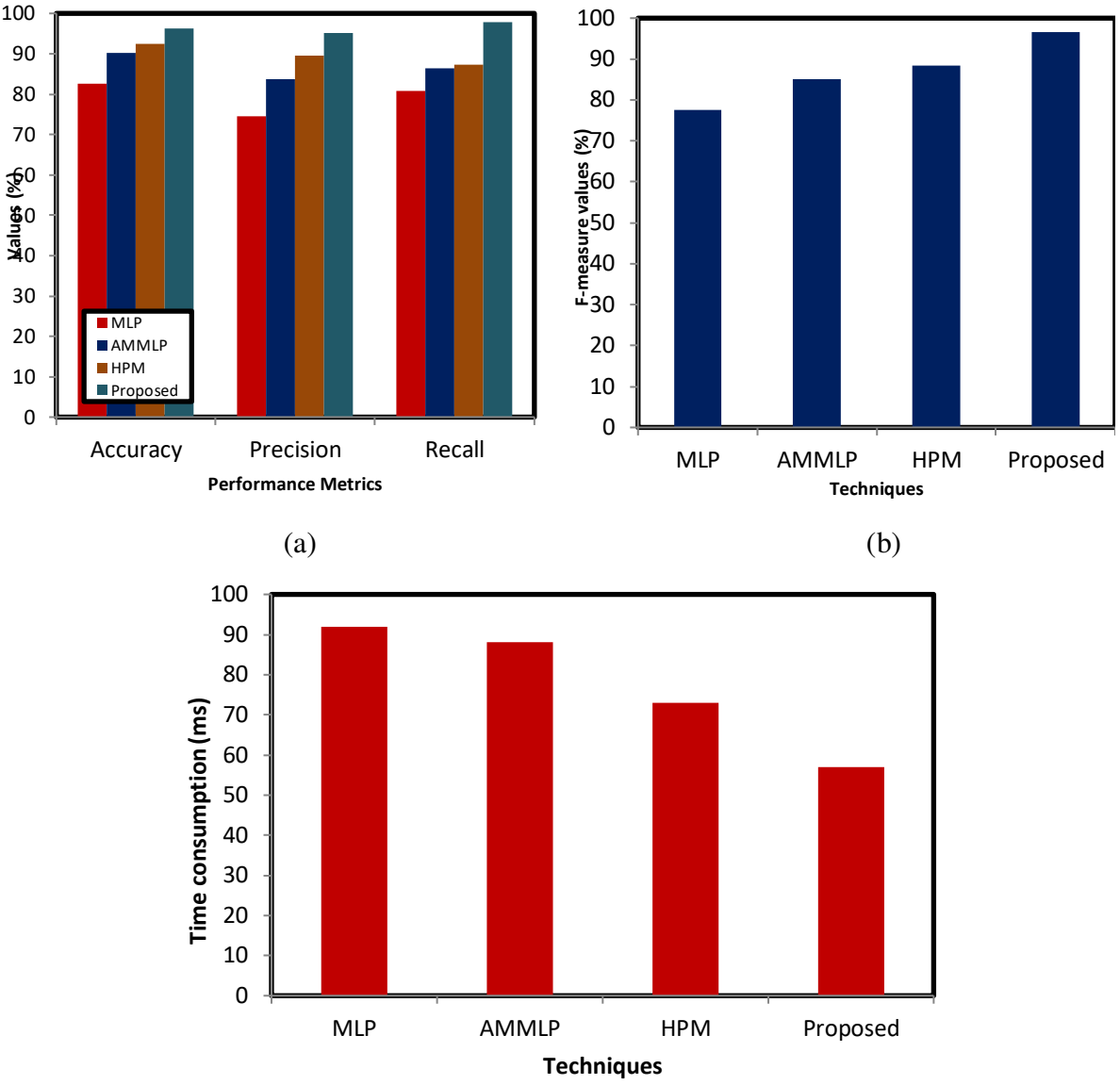

Fig.2. Comparative analysis w.r.t (a) precision, recall, accuracy (b) F-measure, (c) Time consumption analysis

The performance of the proposed approach is tested, compared with the existing approaches and the results are presented in figure 2. From the experimental analysis, it is evident that the proposed approach proves better accuracy, precision, recall, F-measure rates, when compared to the existing techniques. The main objective of this work is to attain greater precision and recall rates, which in turn increases the F-measure rates.

The main reason for the better performance rates is the inclusion of strong association rules for predicting diseases. On the other hand, the reason for minimal time consumption is the utilization of pyramid data structure, which makes it easy for the system to analyse the clusters. The disease prediction accuracy of the proposed system is tested with respect to diseases and the results are presented in figure 3.

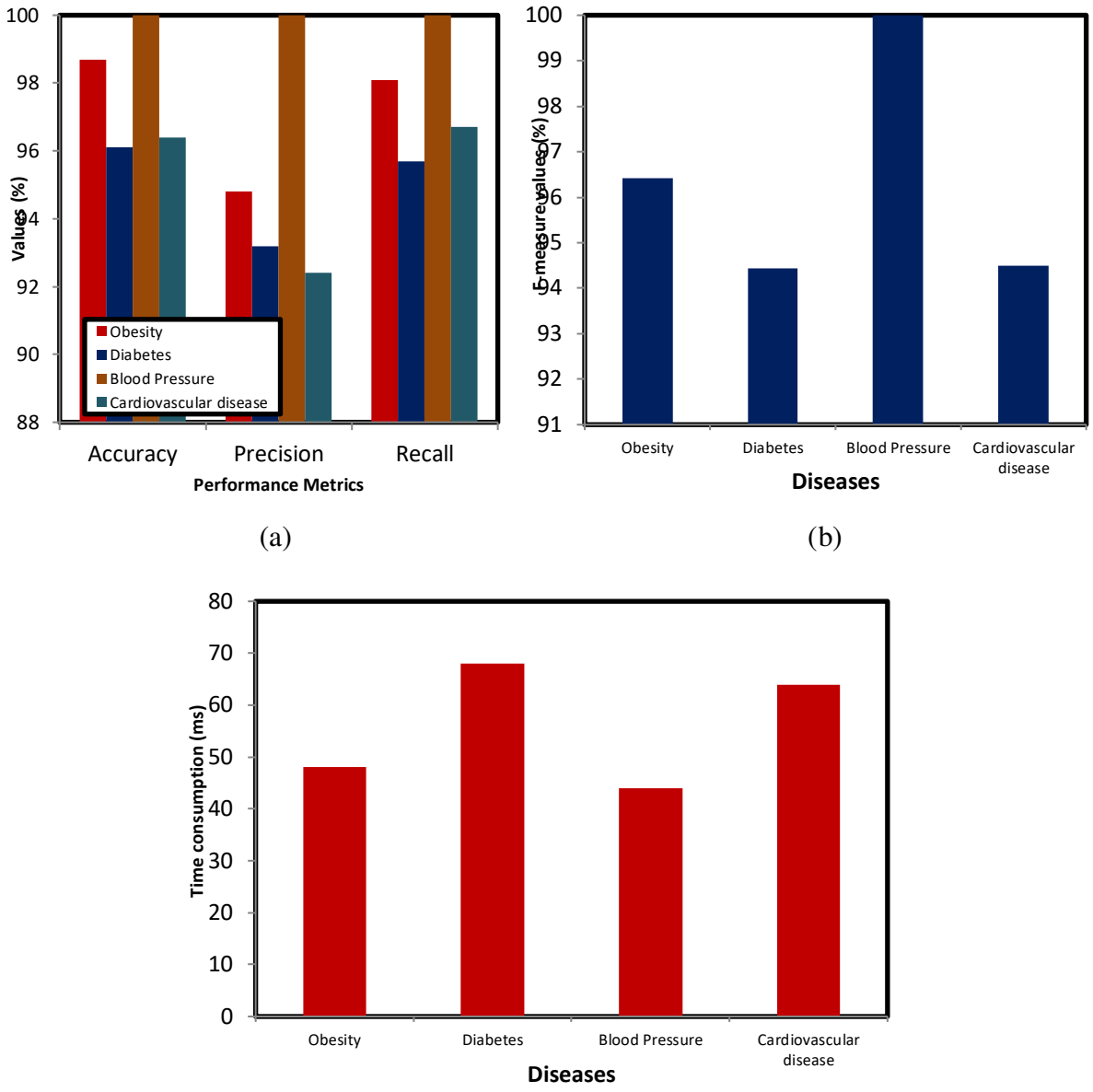

Fig.3. Performance analysis w.r.t disease types (a) Precision, recall, accuracy rate analysis (b) F-measure rate analysis, (c) Time consumption rate analysis

The experimental results show that the proposed approach works in a balanced fashion for all diseases by considering the symptoms passed as input. Hence, the proposed approach predicts between four different diseases with better accuracy rates in reasonable amount of time. The main reasons for reasonable performance rates are the inclusion of clustering concept and the strong association rules. The following section outlines the conclusions of the proposed approach.

5. Conclusion

This article presents a reliable disease prediction system that is based on association rule mining. Initially, the passed dataset is pre-processed and the clustering operation is performed by GHFCM. The reason for the choice of GHFCM is its ability to deal with data of any dimension. As soon as the data are clustered, the strong association rules are built and

the disease is predicted based on the association rules. This way of disease prediction is carried out to distinguish between four diseases such as obesity, blood pressure, diabetes and cardiovascular disease. The performance of the proposed work is analysed in terms of precision, recall, accuracy and time consumption rates. The performance of the work is satisfactory and it outperforms the existing approaches. In future, this work is planned to be extended such that it can predict any disease, which is possible by passing a voluminous symptom dataset.

Declarations

Funding Not Applicable

Conflicts of interest/Competing interests

We know of no conflicts of interest associated with this publication, and there has been no significant financial support for this work that could have influenced its outcome.

Availability of data and material The dataset generated during the current study are available from the corresponding author on reasonable request.

Code availability The code generated during the current study are available from the corresponding author on reasonable request.

Authors' contributions As Corresponding Author, I confirm that the manuscript has been read and approved for submission by all the named authors.

Ethics approval

The article suggest a novel data structure for storing information and perform efficient mining.

Consent to participate

We believe these findings will be of interest to the readers of your journal.

Consent for publication

We declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere

References

[1] A. Probandari , L. Lindholm , H. Stenlund , A. Utarini , A.K. Hurtig , "Missed opportunity for standardized diagnosis and treatment among adult tuberculosis patients in hospitals involved in public-private mix for directly observed treatment short-course strategy in indonesia: a cross-sectional study", BMC Health Serv. Res., 10 (1) (2010) 1–7 .

[2] M. Puppala, T. He, S. Chen, "Meteor: an enterprise health informatics environment to support evidence-based medicine", IEEE Trans. Biomed. Eng. 62 (12) (2015) 2776–2786.

[3] Xiaoyi Xu ; Ao Li ; Minghui Wang, "Prediction of human disease-associated phosphorylation sites with combined feature selection approach and support vector machine", IET Systems Biology, Vol.9, No.4, pp.155-163, 2015.

[4] Muhammad Yasin ; Temesghen Tekeste ; Hani Saleh ; Baker Mohammad ; Ozgur Sinanoglu ; Mohammed Ismail, "Ultra-Low Power, Secure IoT Platform for Predicting Cardiovascular Diseases", IEEE Transactions on Circuits and Systems I: Regular Papers, Vol.64, No.9, pp.2624-2637, 2017.

[5] Shaker El-Sappagh ; Farman Ali ; Amjad Ali ; Abdeltawab Hendawi ; Farid A. Badria ; Doug Young Suh, "Clinical Decision Support System for Liver Fibrosis Prediction in Hepatitis Patients: A Case Comparison of Two Soft Computing Techniques", IEEE Access, Vol.6, pp.52911-52929, 2018.

[6] Jiawei Luo ; Pingjian Ding ; Cheng Liang ; Buwen Cao ; Xiangtao Chen, "Collective Prediction of Disease-Associated miRNAs Based on Transduction Learning", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.14, No.6, pp. 1468-1475, 2017.

[7] Theodora S. Brisimi ; Tingting Xu ; Taiyao Wang ; Wuyang Dai ; William G. Adams ; Ioannis C, "Predicting Chronic Disease Hospitalizations from Electronic Health Records: An Interpretable Classification Approach", Proceedings of the IEEE, Vol.106, No.4, pp. 690-707, 2018.

[8] Marc Piñol ; Rui Alves ; Ivan Teixidó ; Jordi Mateo ; Francesc Solsona ; Ester Vilaprinyó, "Rare Disease Discovery: An Optimized Disease Ranking System", IEEE Transactions on Industrial Informatics, Vol.13, No.3, pp.1184-1192, 2017.

[9] Ameer Mohammed ; Majid Zamani ; Richard Bayford ; Andreas Demosthenous, "Toward On-Demand Deep Brain Stimulation Using Online Parkinson's Disease Prediction Driven by Dynamic Detection", IEEE Transactions on Neural Systems and Rehabilitation Engineering, Vol.25, No.12, pp. 2441-2452, 2017.

[10] Ji Zhang ; Raid Luaibi Lafta ; Xiaohui Tao ; Yan Li ; Fulong Chen ; Yonglong Luo ; Xiaodong Zhu, "Coupling a Fast Fourier Transformation With a Machine Learning Ensemble Model to Support Recommendations for Heart Disease Patients in a Telehealth Environment", IEEE Access, Vol.5, pp. 10674-10685, 2017.

[11] Somaya Hashem ; Gamal Esmat ; Wafaa Elakel ; Shahira Habashy ; Safaa Abdel Raouf ; Mohamed Elhefnawi ; Mohamed I. Eladawy ; Mahmoud ElHefnawi, "Comparison of

Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol.15, No.3, pp.861-868, 2018.

[12] Zhennao Cai ; Jianhua Gu ; Hui-Ling Chen, "A New Hybrid Intelligent Framework for Predicting Parkinson's Disease", IEEE Access, Vol.5, pp.17188-17200, 2017.

[13]Ahmad Aliza, MustaphaH Aida. Comparison between neural networks against decision tree in improving prediction accuracy for diabetes mellitus. ICDIPC 2011, Part I. CCIS 188; 2011. p. 537–45.

[14] Marcano-Cede~no Alexis, Torres Joaquín, Andina Diego. A prediction model to diabetes using artificial metaplasticity. IWINAC 2011, Part II. LNCS 6687; 2011. p. 418–25

[15] Patil BM. Hybrid prediction model for Type-2 diabetic patients. Expert Syst Appl 2010;37:8102–8.

[16] http://archive.ics.uci.edu/ml/datasets/PimaþIndiansþDiabetes.

[17] Yuhui Zheng, Byeungwoo Jeon, Danhua Xu, Q.M. Jonathan Wu and Hui Zhang, Image segmentation by generalized hierarchical fuzzy C-means algorithm, Journal of Intelligent & Fuzzy Systems 28, 961–973, 2015.

[18] Pedrycz, A. & Reformat, M. Hierarchical FCM in a stepwise discovery of structure in data, Soft Computing, vol.10, no.3, pp: 244-256, 2006.

[19] Karayiannis, Nicolaos B. "Generalized fuzzy c-means algorithms." Fuzzy Systems, 1996., Proceedings of the Fifth IEEE International Conference on. Vol. 2. IEEE, 1996.