HEURISTIC APPROACH FOR HEALTH MINING USING GRAPH DATABASE

Amisha Wankhede[1], Dimpal Adate[2], Sneha Pawar[3] , Harshita[4],N.K. Patil[5]
[1]UG Student, Dept. of Computer Engineering, Savitribai Phule Pune University
[2]UG Student, Dept. of Computer Engineering, Savitribai Phule Pune University
[3]UG Student, Dept. of Computer Engineering, Savitribai Phule Pune University
[4]UG Student, Dept. of Computer Engineering, Savitribai Phule Pune University
[5]Assistant Professor, Dept. of Computer Engineering, Savitribai Phule Pune University

Abstract—This paper on health mining is all about the large amount of electronic health records and handling large amount of voluminous data. To overcome this problem we have to convert these voluminous electronic health records into equivalent graphs. The methodologies that we use in this data mining, KDD and Artificial neural networks . Many system related to health mining are having performance issues regarding he detection of health mining. So this paper proposes an idea on health mining. So here we are going to perform health mining and give subsequent graphs related to a patient's disease along with the risk prediction. The only disadvantage that we face is that sometimes our software Neo4j won't be able to convert voluminous data into equivalent semi-supervised heterogeneous graphs.

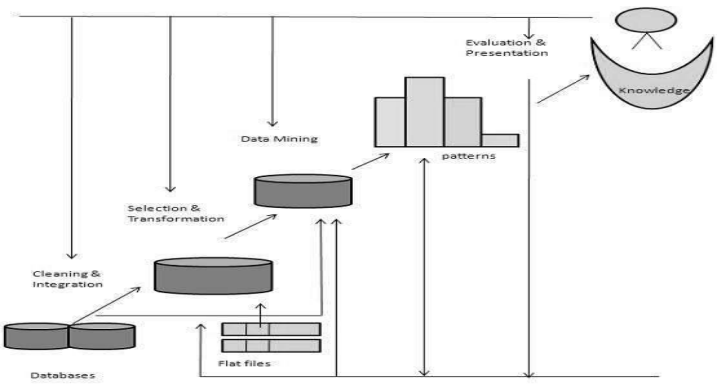
Keywords—Data mining, electronic health records, graph based approach

I. INTRODUCTION

Data mining is an assortment of algorithmic techniques to extract instructive patterns from raw data. Healthcare industry today produces large amounts of voluminous data about hospitals, resources, disease diagnosis, electronic patient records, etc. The large amount of data is important to be processed and scrutinized for knowledge extraction that empowers support for understanding the prevailing circumstances in healthcare industry. Data mining processes include outlining a hypothesis, gathering data, performing pre-processing, estimating the model, and understanding the model and draw the results. Before studying how data mining algorithms are being applied on medical data, let us understand what types of algorithms exists in data mining and how they are functioning. It came into existence somewhere in the middle of 1990's and appeared as a strong tool that extracts needful information from a bulk of raw data. In common, Knowledge Discovery (KDD) and Data Mining are related terms and are used interchangeably but several researchers assume that both terms are dissimilar as Data Mining is one of the most vital stages of the KDD process. The Knowledge Discovery in database is systematized in various stages whereas the first stage is selection of data in which data is gathered from different sources, the second stage is pre-processing the selected data, the third stage is transforming the data into suitable format so that it can be processed further, the fourth stage consist of Data Mining where suitable Data Mining technique is applied on the transformed data for extracting valuable information and evaluation is the last stage.

Knowledge Discovery in databases is the process of extracting high-level knowledge from low-level data. It is a process that comprises steps like Selection of Data, Pre-processing the selected data, Transformation of data into appropriate form, Data mining to extract necessary information and Interpretation/Evaluation of data.

Then we use another methodology that is ANN (Artificial



Neural Network).An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurons.

II. LITERATURE SURVEY

In this section of paper some important works are being analyzed to employ the feature of health mining as follows:

[1] Dr. K. Nachimuthu proposes Extracting Medical heath records in a graph based approach. They tend to recommend a graph-based, semi-supervised learning algorithmic rule mentioned to as SHG-Health (Semi-supervised Heterogeneous Graph on Health) for risk predictions to categorize an increasingly developing scenario with the bulk of the information unlabeled Wide-ranging experiments supported each real health examination datasets and artificial datasets are achieved to indicate the effectiveness and strength of our procedure. Limitations observed are it cannot handle voluminous data.

[2] Hian Chye Koh and Gerald Tan introduced Data mining applications in healthcare. They have described that data mining applications can greatly benefit all parties involved in the healthcare industry. For example, data mining can help healthcare insurers detect fraud and abuse, healthcare organizations make customer relationship management decisions, physicians identify effective treatments and best practices, and patients receive better and more affordable healthcare services. Drawback is that their Health care transactions are too complex.

[3] Sheena Patel and Hardik Patel presented Survey of Data Mining Techniques used in Healthcare Domain they give us knowledge about Health care industry produces enormous quantity of data that clutches complex information relating to patients and their medical conditions. Data mining is gaining popularity in different research arenas due to its infinite applications and methodologies to mine the information in correct manner. Data mining

techniques have the capabilities to discover hidden patterns or relationships among the objects in the medical data. It is sometimes not predictable in terms of drawing accurate solutions.

[4] M.Durairaj and V. Ranjani familiarized us with Data Mining Applications in Healthcare sector: A Study. In this they focused to compare a variety of techniques, approaches and different tools and its impact on the healthcare sector.

The goal of data mining application is to turn that data are facts, numbers, or text which can be processed by a computer into knowledge or information. The main purpose of data mining application in healthcare systems is to develop an automated tool for identifying and disseminating relevant healthcare information. This paper aims to make a detailed study report of different types of data mining applications in the healthcare sector and to reduce the complexity of the study of the healthcare data transactions.

[5] R. Naveen Kumar and M. Anand Kumar suggested a paper on Medical Data Mining Techniques for Health care Systems Due to the sequence in the information technology, the prevalence of the healthcare organizations conserves their data electronically. Enormous progress in medical data leads to be scarce in the mining of well-informed in series from the mass data. There is a necessity for accomplished analysis tools to resolve covered relatives and desire in data. Data mining can represent new biomedical and healthcare details for clinical preference.

[6] Avneet Pannu came up with Artificial Intelligence and its Applications in Different Areas In the future, intelligent machines will replace or enhance human capabilities in many areas. Artificial intelligence is the intelligence exhibited by machines or software. It is the subfield of computer science. Artificial Intelligence is becoming a popular field in computer science as it has enhanced the human life in many areas. Artificial intelligence in the last two decades has greatly improved performance of the manufacturing and service systems. Study in the area of artificial intelligence has given rise to the rapidly growing technology known as expert system.

[7] Douglas F. Scott and R. Larry Grayson proposed Selected Health Issues in Mining Data on health-related illnesses and disease in the mining industry are scarce, and information on rates and costs is not readily available. Substantial amounts
of research are being directed to addressing these issues, including work at the National Institute for Occupational Safety and Health's (NIOSH) mining health and safety laboratories in Spokane and Pittsburgh. This paper briefly discusses the current status of some miner health-related issues, including those involving coal dust, silica dust, diesel particulate matter, asbestos, noise, lead, welding fumes, and skin disorders, as well as research and other activities aimed at protecting miners from occupational illnesses and diseases.

[8] Ling Chen and Xue Li introduced Personal health indexing and geriatric medical examination. The demerits to this methodology is to optimize problems that find optimal of labels as health score based on medical records that are infrequent, incomplete and sparse. Evolution of health care status of a person from cradle-to-grave is becoming possible.

[9] Bath, P.A. deals with Data Mining, Artificial nueral networks, machine learning, decision tree, rule based evolutionary, genetic algorithm. Results may not be accurate to this methodology. They will be widely recognized as complementary to traditional methods of analyses data in health and medicine.

[10] M.S. Viveros, J.P. Nearhos and M.J.Rothman introduced association rules and neural segmentation. By applying implementations on self organizing maps we found that there is no correct number of segments. Data mining can be used in large scales, real customer data with reasonable execution time.

III. PROPOSED METHODOLOGY

Below Presents Proposed System Overview of Health Mining Using Graph Analysis?

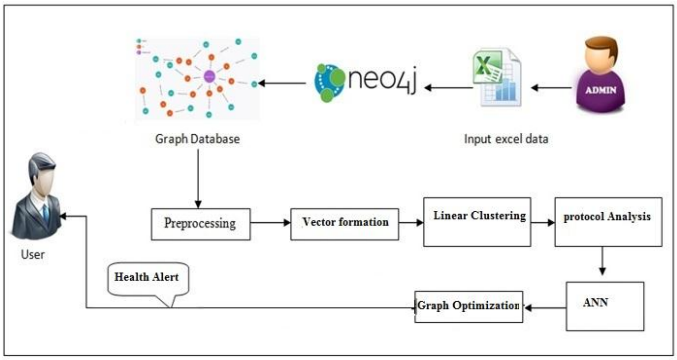


Figure 1: Proposed System Overview

Proposed System is been developed for Health Mining using graph based Approach. The Propose Framework implements Seven Phase Approach in Health Mining Prediction.

Phase 1: Graph Database Creation

In graph database creation Dieses to Symptoms are been related. Association is been created between nodes and edges represent parametric values. Neo4j Database is been used to create graph. Node names are been extracted from graph database based on this conodes are been extracted. Master Graph database is been used to form weighted relation between nodes.

Phase 2: Preprocessing

In this Stage Extracted nodes are been used to read from dataset. Unique Attributes are been selected for preprocessing. Filtration of dataset is been done for future processing

Phase 3: Vector Identification

Based on nodes from graph database and preprocessed data Vector are been identified and vectors to process are been generated. This vectors represent diseases to symptom relates

Phase 4: Linear Clustering

In this phase data is been sequentially clustered and diseases having similar symptoms are been connected commonly. for every common symptom clusters are been generated . This Process is been termed as linear cluster as it only classifies data in parts.

Phase 5: Protocol Analysis

This Phase Data mining is been done based on Health mining protocols . Traditional formal approaches such as theorem proving and model checking have been widely used to analyze security protocols. Ideally, they assume that the data communication is reliable and require the user to predetermine authentication goals mentioned

Phase 6: ANN Formation

Neural network Model is been used for Decision making. Based on No of Input parameters and Symptoms Three neurons are been Generated. Secondly cluster filtration process is been started where cluster ranges are been computed and based on this cluster ranges fine grained clusters generation process is been initiated. Finally Fine Cluster Formation process is been initiated where only fine Grained clusters are been generated from input cluster set.

Phase 7: Graph Optimization

This phase with ever increasing data of heal mining graph optimization is been required. Objective function adopted here is Diseases minimization prediction. As Risk found for Given user is been computed and likely diseases are been only predicted.

Finally user is been Alerted for Predicted likely Diseases based on above Evaluation.

Algorithmic procedure Adopted is as shown below.

ALGORITHM 1: graph Creation

//Input : Data collection Set S ={ s_i,h_i,s_m}

//Output: Hyper Graph **G**(s_i,h_i,s_m)

Where

S_i – Disease names(node)

h_i – Symptoms(node)

s_m- Parameter

Step 0:Start
Step 1: Get the Set **S**
Step 2: FOR **i=0** to Size of **S**
Step 3: Separate s_i, $h_i$into List L_s,L_h
Step 4: **END FOR**
Step 5:Get unique elements form L_s andL_h
Step 6: N_s=Size of L_s(Number of nodes for Disease names)
Step 7: N_h=Size of L_h(Number of nodes for Symptoms)
Step 8: FOR **i=0** to Size of N_s
Step 9: FOR **j=0** to Size of N_h
Step 10: Identify the relational Edges **E** using s_m
Step 11: Form Graph **G**
Step 12: **END FOR**
Step 13: **END FOR**
Step 14: return **G**
Step 15: Stop

IV RESULTS AND DISCUSSIONS

Health Mining Framework is been developed in Neatbeans with Neo4J Graph Database fro relation making. Common Java IDE Netabeans has been adopted in development .Framework is put under testing to find performance evaluation of proposed work Numerous test have been done to find best performance of system.

Performance is evaluated based on the precision and recall parameters. Precision is defined as the ratio of number of relevant diseases identified are detected to the total number of relevant and irrelevant diseases detected. Relative effectiveness of the system is well expressed by using precision parameters.

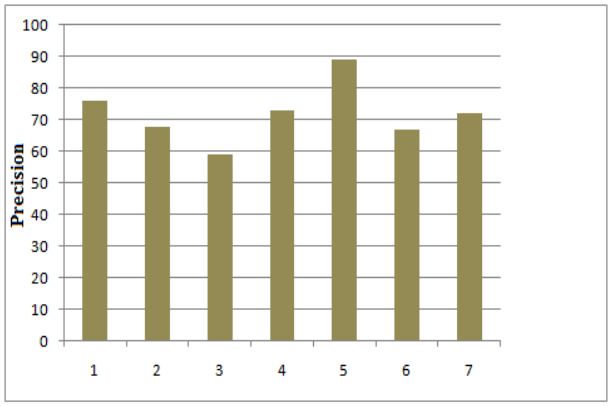


Fig.2. Average precision for Health Mining System

Whereas the recall can be defined as the ratio of number of relevant Diseases detected the total number of relevant Diseases not detected. Absolute accuracy of the system is well narrated by using recall parameters.

System can be evaluated using precision and recall parameters, and they can be more clearly elaborated as follows.

• F = The numbers of relevant traffic symbols are detected,
• G = the number of relevant traffic symbols are not detected, and
• H = the number of irrelevant traffic symbols are detected.

So, Precision = (F/ (F+ H))*100
And Recall = (F/ (F+G))*100

In Fig. 2, by observing it is clear that the average precision obtained for Health Mining System approximately 67%.

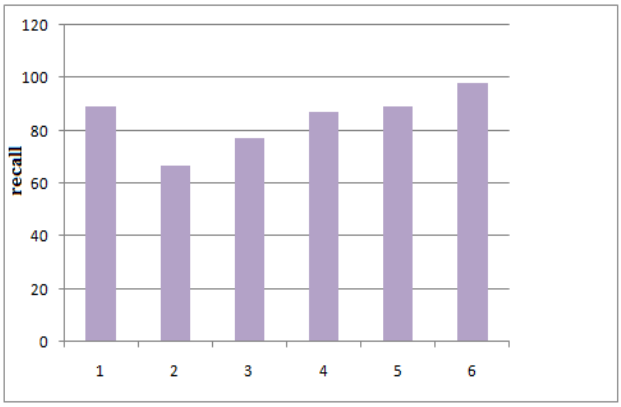


Fig.3. Average Recall for Health Mining System

Figure 3 shows that the system gives 78% recall for the Health Mining. By comparing these two graphs we can conclude that the health Mining method gives high recall value compare to the precision value.

V CONCLUSION AND FUTURE SCOPE

In Future System could be extended for Multiple diseases Symptoms values. Complex graph Processing Would be Required . As Design and Development of advanced graph Extraction framework is been required which would Extract better nodes and relation Mining. Better Evaluation of system and additional Decision support would better system.

REFERENCES

[1] Dr. K. Nachimuthu , Extracting Medical heath records in a graph based approach;2014, IEEE.
[2] Hian Chye Koh and Gerald Tan introduced Data mining applications in healthcare; 2012,IEEE.
[3]] Sheena Patel and Hardik Patel,Survey of Data Mining Techniques used in Healthcare Domain;2016,IJIST.
[4] M.Durairaj and V. Ranjani , Data Mining Applications in Healthcare sector: A Study;2013, IJSTR.
[5]] R. Naveen Kumar and M. Anand Kumar suggested a paper on Medical Data Mining Techniques for Health care Systems;2016,IJESC.
[6] Avneet Pannu , Artificial Intelligence and its Applications in Different Areas;2015,IJEIT.
[7]] Douglas F. Scott and R. Larry Grayson ,Selected Health Issues in Mining;2015,IEEE
[8]] Ling Chen and Xue Li introduced Personal health indexing and geriatric medical examination;2014,IEEE.
[9] Bath, P.A. deals with Data Mining health and medical information;2004,IEEE.
[10] M.S. Viveros, J.P. Nearhos and M.J.Rothman. Applying data mining technique to health insurance information system;2015,IRJET.