Improving efficiency of graph based health data mining using language processing

RashmiTembhurne

M.E (CSE) Scholar
Department of CSE,
Truba Institute of Engineering & Information Technology
Bhopal, India
rashmi.dt@gmail.com

Prof. AmitSaxena

Department of CSE,
Truba Institute of Engineering & Information Technology
Bhopal, India
amitsaxena@trubainstitute.ac.in

Abstract: *- Nowadays people are vulnerable to multiple diseases due to environment changes which have multiplied the work of doctors around the world.The technology has helped the medical science to treat their patient more efficiently increasing the number of lives saved every day. But to treat the patient the technology is not limited , it has spread its roots now amongst detecting the diseases too. The medical database contains the total report of the patient, past illness, current problems and all the scan done till date.This data is very useful in calculating the future problems which the patient could have. To discover this knowledge from the database we need a technique which is known as mining technique, which is helpful enough to process such data. In this paper, we propose a technique which is a combination of the Natural Language Processing (NLP) and Graph mining.The technique will first filter the data so that only optimized data is sent for mining which is done with the help of graph-based technique providing efficient results.*

Keywords: *Medical Data, Graph Mining, Natural Language processing*

I. INTRODUCTION:

Data mining is defined as the non-trivial mining of structured or non-structured data for making efficient future decisions or for analysis purpose. Electronic Health Records (EHR's), is a medical history of patients which includes the patients past diseases, current diseases with the treatments given and every little detail about the person. These interpretations are very useful therefore, around the world, a lot of countries encourage HER that is health examination record which is a special type of EHR based on regular basis. For example, the governments of U.K. and Taiwan holds a periodic geriatric health examination.

This data helps the doctor to analyze risk, i.e. the data should tell whether the patient is at risk of particular diseases if yes then what is the current position and the treatment available. The system should be intelligent enough to disregard low risks and to calculate the near exactness of the high-risk diseases.

This analysis is done with the help off mining the data available with us. This data mining is termed as the KDD or knowledge discovery database. By using the mining process, the implicit relations and patterns are extracted to find an answers . In this system, the mining is done using the graph-based approach which extracts the features of the user data and compaires it with the data availed to find the results. By using the natural language process, the process of graph technique is optimized. Since the data can be in an implicit and non-trivial fashion which is very hard to process therefore, the NLP will first process the input data to optimize the system and converts it into the desired format and then the graphing technique is applied

In this paper, the accuracy of the technique with and without NLP is tested. The rest of the paper is structured as follows. The next section briefly explains the related work, following with the proposed technique, its simulation, and results. Lastly, the paper is concluded with the conclusion.

II. RELATED WORK:

The increase in the knowledge mining applications from medical databases is increasing rapidly. Mainly two mining techniques are applied to the medical data: Exploratory and explanatory. When data investigation is performed at the early stage of data analysis where the exact mining objective is not set is referred as exploratory mining and the techniques used for verification and descision making is referred as Explanatory mining.

A number of studies have done on explanatory mining in medical data over last few years. To figure out classification rules from medical data sets, genetic programming techniques has been applied. AdaBoost algorithm has been used to work on breast cancer survivability. For selected features medical data, fuzzy modeling has been developed. A system is proposed from health examination data to extract association rules and to support the continual disease analysis and management; a case-based reasoning model is used. Recently, a rule mining method having case-based reasoning is applied. Medical data warehouses are used as an extension to the

normal medical databases. There have been very fewer studies on exploratory mining techniques.
One of the studies that use exploratory mining technique is visualization of knowledge in the study of hepatitis patients. One more study is the improvement of visualization using OLAP functionality.

III. PROPOSED SYSTEM:

The proposed system is divided into three major modules. First is the Natural Language Processing (NLP) following with Graph-Based Mining module and the last is the database tested. The system is explained as follows:

A. Natural Language Processing:

The medical information contains a lot of information which is not useful while calculating the decisions. If the system is provided with such redundant data the output will also be redundant therefore, NLP is proposed to optimize the data input to the system.. It will take the input and extract the important features and will only provide the required data to the system for analysis. It is a field that works with the human and machine interactions. The main process of the system is to converts human language and it helps the machine to understand users' language which he/she speaks or writes. An intelligent system is created which is capable of keyword extraction and searching keywords from the files.
The NLP follows a set of steps, firstly the dataset which is given as the input is tested for example if we are searching for the brain diseases but the patient record contains details about heart, lungs, bones and other organs and diseases he/she faced till the time. Such a set is given to the NLP for processing. From the set of data, the keyword is to be extracted. The NLP finds keywords of all the files. For example in the ECG of the patient the NLP will mark it as its main keyword as the Heart, ECG and the values fetched as the result all the other information on the report is useless when making decisions. The third step of the system is the keyword association it is the matching phase, here all the other datasets are tested and are processed to fetch their keywords.
Then these keywords are matched to find the output result of the system. The keywords in NLP are calculated using tagging and chunking. Part of Speech the process of Tagging is ,where the sentences are tagged with their POS which is based on its definition and the context from which it is used. The POS tag will highlight the nouns, pronouns, verbs and other parts of speech in the data inputted.
For example.
Harry saw the Axe Here harry is a pronoun, the saw is the verb and axe are the nouns. Similarly, the chat is read and the POS tag is done on it. Whereas , chunking is nothing but partial parsing. It highlights those regions which are not overlapped. Every and each chunk has a head, which is possibly the keyword. For example:
[Walk] [Straight past] [The lake] Chunks are non-recursive, i.e. they do not contain another chunk of the same category.

B. Graph-Based Mining:

A standout amongst the most attractive data structure in software engineering and discrete arithmetic are graphs. It can be induced that the diagram based information mining has been generally known in the course of recent years. One of the real issues in the continuous item set mining is time intricacy and the issue stands settled in our proposed new approach.
The graph-based approach checks the whole database just once and this unmistakable element is the most extreme required after paradigm in the space of recognizing regular sets that result in creating an enormous measure of applicant sets. By the processing, it makes a graph which is basically a coordinated diagram. The weights of the chart are stored in the memory and are meant as a contiguousness network.
Here the items from the dataset are exhibited through the vertex of the coordinated diagram and the heaviness of the vertex speaks to the bolster support count of one item set. The vertices having weights are associated with an edge. If there is an occurrence of mining huge k-itemsets (k>=3), the connections must be characterized. It might happen that various exchanges in a database may contain the comparable arrangement of items regardless of the possibility that two exchanges are totally not same as each other and the case might also be that the two exchanges may contain indistinguishable itemsets along these lines their subsets might be basic.
To understand the technique we need to understand three basic terms first is the Graph (G) which is linear G = (V, E) where V = {v1, v2} are the objects called vertices and E = {e1, e2} which are termed as the edges. Vertices and edges are associated with each other. The second term is the Directed Graph an edge which is associated with an ordered pair of V * V is called a directed edge of G.
A graph in which every edge is directed is called a directed graph. Last is the adjacency matrix, where graph G with n vertices and no parallel edges is an n by n symmetric binary matrix X. To evaluate the system two semantics are performed. First is the fully qualified adjacency matrix and second is the reduced matrix close to frequent itemsets. One of the transactional datasets is shown in the following figure.

TID	a	b	c	d	e
1	0	0	0	0	1
2	0	6	0	1	1
3	2	0	2	0	1
4	0	0	0	0	0
5	1	0	6	0	1
6	1	1	1	0	0

Figure 1: Data Table of transactional database D

The database D is scanned and a directional graph is created as shown in the following figure.

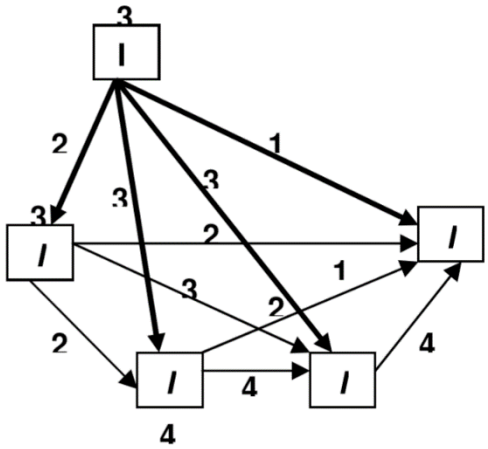

Figure 2: Directed Graph G of Database D

The graph is also stored in the memory in the form of adjacency matrix as follows:

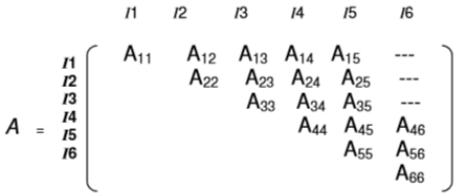

Figure 3: adjacency matrix of Database D

Now to reduce the matrix to frequent itemsets the verification of the value count of each element of the matrix A is done. For any diagonal element if the value of A_{ij} is less than min_sup the corresponding row and column are deleted that is given as 0. Following equation shows the frequent itemsets calculation list.

A_{12} .list ∩ A_{13}.list = {I1, I2, I3} = (T1, T2)
A_{12} .list ∩ A_{14}.list = {I1, I2, I4} = (T1, T2)
A_{13} .list ∩ A_{14}.list = {I1, I3, I4} = (T1, T2, T3)

Ultimately the reduced data table can be fetched as shown in the following figure:

TID	a	b	c	d
1	0	0	0	0
2	0	6	0	1
3	2	0	2	0
4	0	0	0	0
5	1	0	6	0

Figure 4: Reduced Data Table

Finally we have generated the following frequent itemsets:

level-1minsup=4

Large-1 itemset = (a1}:1
= {c2}:1

Large-2 itemset={c2,d1}:2
= {c2, a1}:2

Finally, the algorithm for the Graph-based Approach is as follows:
Initialize: Set of traction D, Total number of itemsets with the occurrence.

Step 1: Scan: Scan the database and create discrete graphs from the values,

Step 2: Identify: Update and fetch the values of each element in the matrix

Step 3: Construct: The reduced adjacency matrix processes,

Step 4: Mine: number of levels are to be mined using operators.

Output: Frequent Itemsets.

The simulation of the system is explained in the following section.

IV. IMPLEMENTATION:

This section of the paper deals with the implementation of the proposed technique. The flow chart below shows the system flow of the proposed technique:

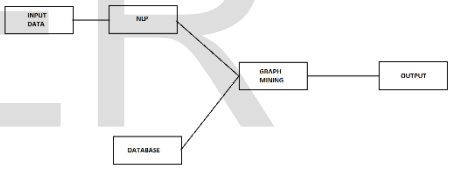

Figure 5: flow of the proposed technique

Database: The database is made up of images and reports of the human body and patient history till date which are provided by medical professionals. The dataset can be divided into multiple sub-parts as the human body can be divided into the brain, hands, and lungs. As shown in the figure below the data set is made of different images which are collected medical imaging, which is a study of body parts. The dataset used in this system is University of California medical dataset.

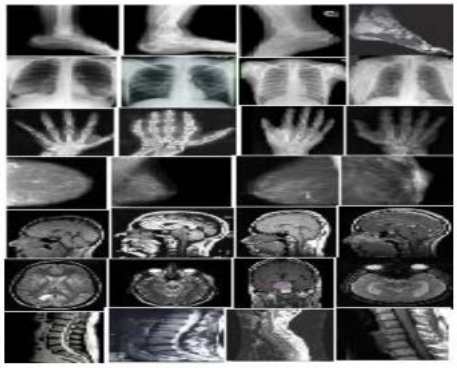

Figure 6: Typical BioMedical Database

To perform the analysis firstly the input is given.

For example, the doctor wants to know the status of patient cancer for which all the records are given as the input. The steps to implement are as follows:

Step 1: Input: In this stage, the doctor will upload the patient records into the system, for example, the doctor is having copies of the Scan the blood and various other reports. These reports are uploaded to the system.

Step 2: Filter; once the system is filled with the required data the system starts to sort the data. This sorting is done to enhance the efficiency of the system. Since the brain is to be treated, the system will only keep the files with a brain as their main keyword. For example, there are two reports present one is the bone fracture report and other is brain swelling. Since the fractured one is irrelevant it will be discarded and the swelling report will be sent for further analysis.

Step 3: Once NLP filters the report type now is the time to optimize the filtered input data. Now taking the same swelling report, the data is pictorial as well as in written like the swollen percentage around the ages. So the NLP will remove the joiners the images and all the non-related stuff and will only send important data.

Step 4: once the data is filtered the system gives the data to graph mining where the system take takes the user input and compares it with the data in the database. The patient swollen brain data will be compared with swollen brain data of other patients.

Step 5: the best results are given out like how much the brain is swollen at that point what type of treatment can be given and so on.

V. RESULTS:

In this section, we will test the accuracy of the proposed technique. To test the accuracy the system is first implemented only using Graph-Based mining whereas in the next phase NLP is also applied. The following tables are fetched as a result.

Items to test	Time (milli sec.)	Accuracy (%)
10	10	75.35
20	12	76.12
50	13	76.91
100	15	77.88
200	18	78.01
500	20	78.55
1000	23	79.29

Table 1: Accuracy without NLP

As observed from the above table as the items increase the accuracy also increases and the time required to process the data also. The highest achievable accuracy without NLP is approximately 80%.

Items to test	Time (milli sec.)	Accuracy (%)
10	7	78.35
20	8	79.12
50	9	79.91
100	10	80.88
200	12	81.01
500	13	81.55
1000	18	82.29

Table 2: Accuracy with NLP

As observed from the above table as the items increase the accuracy also increases and the time required to process the data also. But in this case, the highest achievable accuracy with NLP is approximately 80% and also the time required to process the data is much less than without NLP.

From the above two tables, we can say that when the system is applied with NLP the accuracy is increased and the delay decreased making it a more efficient system.

VI. CONCLUSION:

We concluded that when the system is applied with NLP the accuracy is increased and the delay decrease, making it a more efficient system.

VII. REFERENCES:

[1]. AnuragChoubey , Dr. Ravindra Patel and Dr. J.L. Rana, “GRAPH BASED NEW APPROACH FOR FREQUENT PATTERN MINING”, International Journal of Computer Science & Information Technology (IJCSIT) Vol 4, No 1, Feb 2012.

[2]. Wael Ahmad AlZoubi, “Mining Medical Databases Using Graph based Association Rules”, International Journal of Machine Learning and Computing, Vol. 3, No. 3, June 2013.

[3]. Ling Chen, Xue Li, Member, IEEE, Quan Z. Sheng, Member, IEEE, Wen-ChihPeng, Member, IEEE, John Bennett, Hsiao-Yun Hu, and Nicole Huang, “Mining Health Examination Records — A Graph-based Approach”, 1041-4347 (c) 2016 IEEE.

[4]. Mohammed Abdul Khaleel, Sateesh Kumar Pradham, G.N. Dash, “A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases”, 2013, IJARCSSE.

[5]. CharuC.Aggarwal, “GRAPH DATA MANAGEMENT AND MINING: A SURVEYOFALGORITHMSANDAPPLICATIONS”, IBM T. J. Watson Research Center.
[6]. Lars Bungum, Bj¨ornGamb¨ack, “Evolutionary Algorithms in Natural Language Processing”, Norwegian Artificial Intelligence Symposium, Gjøvik, 22 November 2010.