Extraction of frequent patterns from medical datasets for detection of multiple disease

Tarannum Shaila Zaman
Computer Information Science
SUNY Polytechnic Institute
Utica, USA
Tarannum.Zaman@sunypoly.edu

Anil Reddy Chepyala
Computer Information Science
SUNY Polytechnic Institute
Utica, USA
chepya@sunypoly.edu

Bavitha Raavi
Computer Information Science
SUNY Polytechnic Institute
Utica, USA
raavib@sunypoly.edu

Mamatha Narri
Computer Information Science
SUNY Polytechnic Institute
Utica, USA
narrim@sunypoly.edu

ABSTRACT

The purpose of our paper is to construct a basic prototype model which can determine and extract unknown knowledge (patterns, concepts, and relations) related with multiple disease from a past database records of specified multiple diseases. Healthcare systems are using Medical Records, promoting the use of long datasets that contain relevant information about the health of a patient. The increasing availability of data, represents an impact on the potential discovery of patterns of new diseases, helping the personalized care and increasing the quality of life. Finding if there is any kind of relationship between a drug and a diagnostic or discovering frequent hidden patterns are relevant important areas of interest in healthcare.

Our approach aims to utilize the data mining techniques: clustering and frequent pattern mining. The medical data warehouse consists of mixed attributes containing both the numerical and categorical data. These records are cleaned and filtered with the intention that the irrelevant data from the warehouse would be removed before mining process occurs. Then clustering is performed on the preprocessed data warehouse using K means clustering algorithm with K value to extract data relevant to heart diseases, asthma, and cancer. Subsequently the frequent patterns significant to these diseases' diagnosis are mined from the extracted data using the Apriori algorithm.

The system was implemented in SPMF and predicts the risk of diagnosis with a high precision of 87.6% and less time complexity compared to that of previous related.

KEYWORDS

SPMF, APRIORI, FREQUENT PATTERNS, STRUCTURED DATA.

1. INTRODUCTION

Medical data mining has a significant amount of potential for uncovering hidden patterns in the relevant data sets. One can use these patterns to obtain a clinical diagnosis. The existing raw medical data, however, are voluminous, diverse, and widely spread. This information must be gathered in a structured manner for a proper utilization. A user-oriented approach to discovering new and hidden patterns from this data is provided by data mining technologies.

The medical data warehouse is a mix of both the numerical and categorical data attributes. These data records are cleaned and filtered to get rid of irrelevant data before mining process occurs. This preprocessed data is clustered using K means clustering technique to extract only the data relevant to heart diseases, asthma, and cancer. Subsequently, the frequent patterns significant to these diseases' diagnosis are mined from the extracted data using the Apriori algorithm.

Heart diseases, asthma and cancer are the primary driver of deaths in the current world. So, we worked on data item sets with previous patients' health records. In this work, Frequent pattern mining procedures are analyzed for testing their precision and execution on preparing medicinal informational index. The classification results will be envisioned by various representation procedures like graphs. All the existing papers related to this concept, only conducted the research on a single type of disease. Our paper discusses the significant patterns for 3 diseases by analyzing the symptoms of the patients by a training from thousands of health records.

The detection of diseases from various factors or symptoms is a multi-layered issue which is not free from false presumptions often accompanied by unpredictable effects. Therefore, it is thought to be a beneficial choice to use the expertise and experience of many doctors as well as clinical screening data of patients gathered in databases to aid in the

diagnosis process. Unfortunately, the healthcare sector collects a lot of information about heart disease, are not “mined” to determine concealed information for effective decision making by healthcare practitioners.

The main goal of our paper is to achieve more precision than the existing papers. This will result in maintaining clinical documentation and awareness of patient’s health.

2 MOTIVATION

Even today, the healthcare industry is "information rich" yet "knowledge poor." Within the healthcare systems, there is a lot of data. Effective analysis tools, however, are lacking, making it difficult to find hidden links and patterns in data. Our research paper intends to provide a survey of current techniques of knowledge discovery in databases using data mining techniques that are in use in todays’ medical research. The system based on risk factors of the diseases would not only help medical professionals, but it would give patients a warning about the probable presence of the disease even before he/she visits a hospital or goes for costly medical checkups. Hence this paper presents a technique for prediction of diseases using major risk factors.
In medical In the medical industry, mining algorithms can be used to diagnose some serious diseases. Among all diseases, heart diseases, asthma and cancer are the top one cause of death in the world, claiming more lives. Thus, how to predict these diseases in real life is of great significance, both to research and application. The medical data of the patients infected by diseases is encapsulated in a dataset tabular form. This model can be deployed in hospitals for simplifying the process of admission of patient in case of emergency as our system can predict the disease by the historical data of patient.

3 BACKGROUND

In this paper, we collect medical data from open-source i.e. data.gov. This data contains valuable information for training sets in research.

Next the collected data is preprocessed by following data preprocessing steps. Those steps include data cleaning, data transformation and data reduction.

Then this preprocessed data is clustered using a k-means clustering algorithm with k value as 4 i.e. 4 clusters are heart diseases, asthma, cancer and irrelevant diseases to our paper.

Clustered data is then stored in the excel file. One excel sheet for each disease dataset which includes all the attributes of the patient’s medical history.

For a better understanding on this paper, one should have the basics on how data is preprocessed, clustered, and used. And some knowledge about mining processes, to analyze algorithm application steps.

Data preprocessing:

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Data Cleaning:
The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

i. **Missing Data:** This situation arises when some data is missing in the data. It can be handled in various ways.
   Some of them are:
   a. **Ignore the tuples:** This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
   b. **Fill the Missing values:** There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

ii. **Noisy Data:** Noisy data is a meaningless data that can’t be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc.
   It can be handled in following ways :
   a. **Binning Method:** This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
   b. **Regression:** Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

Data Transformation:
This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

a. **Normalization:**
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
b. **Attribute Selection:**
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
c. **Discretization:**
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
d. **Concept Hierarchy Generation:**
   Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

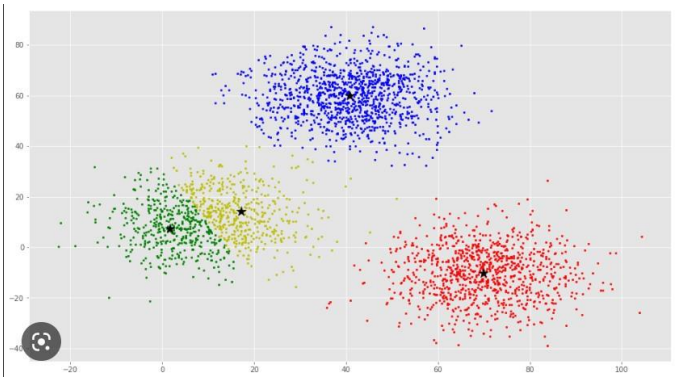
Data-reduction
Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we use data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

K-means clustering algorithm:

The algorithm will categorize the items into k groups or clusters of similarity. To calculate that similarity, we will use the euclidean distance as measurement.

The algorithm works as follows:

First, we initialize k points, called means or cluster centroids, randomly.
We categorize each item to its closest mean, and we update the mean's coordinates, which are the averages of the items categorized in that cluster so far.
We repeat the process for a given number of iterations and at the end, we have our clusters.



Frequent pattern mining algorithm:

The Apriori algorithm operates on a straightforward premise. When the support value of an item set exceeds a certain threshold, it is considered a frequent item set. Take into account the following steps. To begin, set the support criterion, meaning that only those things that have more than the support criterion are considered relevant.

Determine the level of transactional database support and establish the minimal degree of assistance and dependability. Take all of the transaction's supports that are greater than the standard or chosen support value. Look for all rules with greater precision than the cutoff or baseline standard, in these subgroups. It is best to arrange the rules in ascending order of strength.

4 APPROACH

We proposed an efficient prediction of different Diseases based on the historical and training data. This idea is to analyze and test various data-mining models and algorithms and to implement the algorithm which gives out highest degree of accuracy. The Dataset implemented ideally contains medical related attributes for each disease and approximately 800 data items for each. The diseases included in the dataset are cancer, heart disease, and asthma. The algorithms and classifier models used for are

- Data preprocessing for structuring data,
- k-means clustering
- Apriori algorithm
- Average calculation

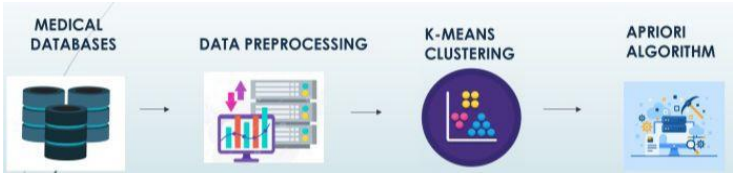


Fig 4.1: *the approach for extracting significant patterns*

4.1 EVALUATION

In our work, the three profound research questions are.

- RQ 1: What is the role of the clustering technique.
- RQ2: How efficient is the Frequent pattern mining over other algorithms.
- RQ3: What is the precision of frequent patterns obtained from the application of apriori algorithm.

The first research question explains the importance of clustering and why k-means clustering is used. The second research question describes hoe frequent pattern mining is precise. Lastly, third research question compares the various algorithms in pattern extracting.

4.2 METRICS:

Accuracy: The results obtained from the application of apriori algorithm on the medical datasets are exact and accurate compared to that of all the existing works.

Precision: As the datasets collected were more than 800, the model is trained exceptionally, and is resulting in desired outputs with 87.6%.

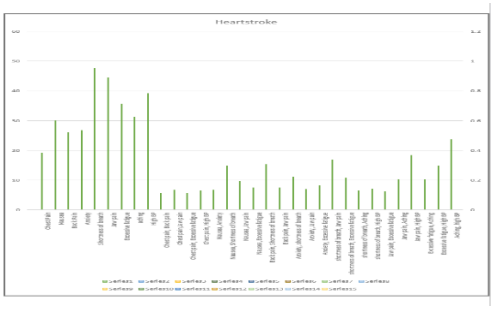


Fig 4.1 Extracted patterns percentage for Heart stroke

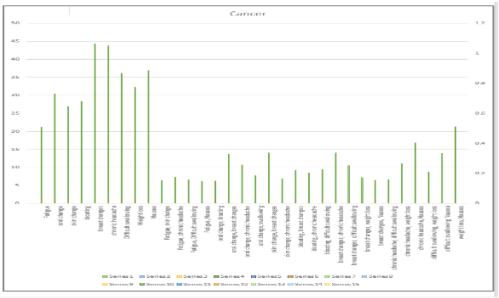


Fig 4.2 Extracted patterns percentage for Cancer

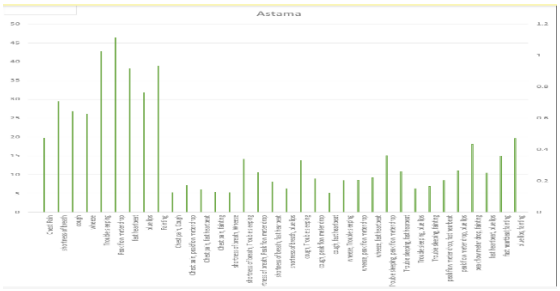


Fig 4.3 Extracted patterns percentage for asthma

5 RESULTS AND ANALYSIS

5.1 RQ1: What is the role of the clustering technique.

In our research paper, we have done the clustering on available medical datasets which are preprocessed already by data preprocessing steps.

Clustering helps in classifying heart diseases data, asthma data, cancer data from other diseases data (this is considered as irrelevant cluster to our paper). Here k=4, this results in clustered data which helps in formatting for extracting patterns.

5.1 RQ2: How efficient is the Frequent pattern mining over other algorithms.

Finding the recurring components in the data collection is mostly dependent on the frequent itemset. By putting the entire dataset into a new text file and using the output from the frequent itemset, we deleted all the repeated components, identified the frequent patterns, and eliminated them.

5.1 RQ3: What is the precision of frequent patterns obtained from the application of apriori algorithm.

Apriori prints a rule of the form s (f s) for each frequent item set f and each non-empty subset s of f if and only if the confidence of that rule is greater than the user-specified threshold. Simply put, the rule's correctness is the confidence. The precision is greater for apriori algorithm compared to that of other mining algorithms

RESULTS & ANALYSIS:

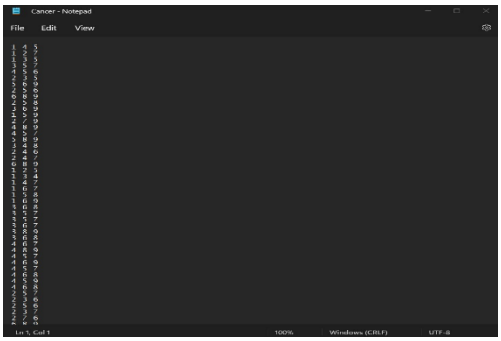


Fig.5.1 Inputs from cancer dataset

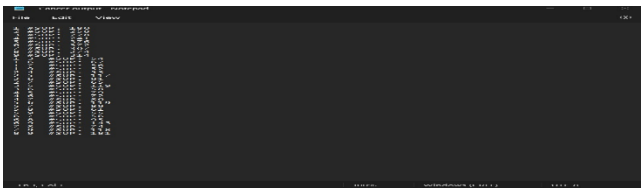


Fig.5.2 Outputs for cancer dataset

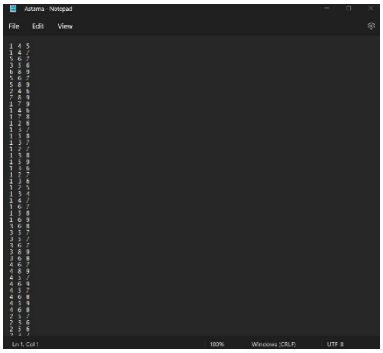


Fig.5.3 Inputs from asthma dataset

Fig.5.4 Outputs for asthma dataset

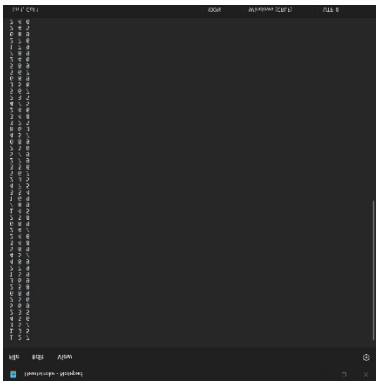


Fig.5.5 Inputs from heart disease dataset

Fig.5.6 Outputs for cancer dataset

6 THREATS TO VALIDITY

i. Threats to Internal Validity

In our paper, we worked on 3 types of diseases by analyzing their symptoms with a wide range of datasets (approximately 800 medical datasets) as it is related directly with accuracy of the results. It is to reduce the data mining bias (Data mining bias occurs when data is analyzed excessively, leading to statistically irrelevant and, non-existing patterns)

ii. Threats to External Validity

Huge datasets from open sources are considered for analysis. In this case, there's a chance of decrease in external validity as homogeneous subject is taken into account. Our proposed model could confront issues in real-world due to this, as data changes consistently.

iii. Threats to Construct Validity

There's a possibility of mono-operational bias. As we are using only Apriori algorithm, it may not be capturing the full breadth of the concept. So here, we can focus on applying multiple algorithms in future for more precision.

iv. Threats to Conclusion Validity

Sometimes, the output patterns can be incorrect or not as much as expected. In such cases, applying algorithm multiple times on the structured dataset could help obtain efficient frequent patterns.

In our paper, as we are dealing with real-time clinical data, which is dynamic in nature, may affect in deployment of this system in real-world, as we develop patterns only on a static dataset.

7 RELATED WORKS

Numerous works in literature related with heart disease diagnosis using data mining techniques have motivated our work.

i. Heon Gyu Lee suggested an unique method for developing a multi-parametric feature of the heart rate variability with both linear and nonlinear properties. The multi-parametric feature of HRV was developed using statistical and classifications methods. Moreover, they evaluated the supine, left lateral, and right lateral postures while lying in three different recumbent positions to evaluate the linear and non-linear features of HRV. They conducted several tests to assess a variety of classifiers, including Bayesian classifiers, CMAR (Classification based on Multiple Association Rules), C4.5 (Decision Tree), and SVM, on the linear and nonlinear

properties of HRV indicators (Support Vector Machine). SVM outperforms the other classifiers.

ii. Sellappan Palaniappan proposed the Intelligent Heart Disease Prediction System (IHDPS), a model was developed using data mining techniques such as Decision Trees, Naive Bayes, and Neural Networks. The results showed the distinct advantages each methodology had in recognizing the given mining objectives. IHDPS was able to provide answers to questions that could not be provided by traditional decision support systems. It helped in the formation of crucial knowledge, such as patterns and correlations among the medical variables linked to heart disease. IHDPS thrives because it is web-based, approachable, scalable, trustworthy, and expandable.

iii. Niti Guru proposed using neural networks to predict heart disease, blood pressure, and blood sugar. On a test database of patient records, experiments were carried out. Thirteen input variables, containing age, blood pressure, an angiography report, and other parameters, are used to test and train the neural network. Heart disease diagnosis has been considered for the supervised network. Back propagation technique was used to train the system. Every time the doctor input the system with unknown data, the algorithm recognized the unknown data through comparisons with the training data and produced a list of likely diseases that the patient is exposed to.

iv. Carlos Ordonez looked into the topic of identifying limited association rules for heart disease prediction. The evaluated data set comprised patient medical records with characteristics for risk factors, cardiac perfusion tests, and arterial narrowing. To reduce the amount of patterns, three limitations were added. The first situation demands that the qualities only appear on one side of the rule. The second separates qualities into various categories. The number of properties in a rule is limited by the ultimate constraint. Experiments showed that the constraints not only improved the running time but also significantly decreased the number of found rules. The existence or absence of heart disease in four distinct cardiac arteries was predicted by two sets of criteria.

v. Data mining methods may help clinicians predict the prognosis of patients and, as a result, adapt their treatments. Every medical treatment or health problem might be addressed using Franck Le Duff's approach, and a decision tree could be began building using information on a service or a doctor. A comparison of traditional analysis and data mining analysis showed the data mining method's role in variable sorting and came to a conclusion regarding the importance of the information and variables and their impact on the study's conditions. Knowledge acquisition and the necessity to obtain sufficient data to develop a suitable model became two major drawbacks of the procedure.

vi. Boleslaw Szymanski proposed a new heuristic for the quick computation of the sparse kernel in SUPANOVA. It was used to enhance the diagnosis of diseases in the population using a novel, non-invasive measurement of the heart activities based on the magnetic field produced by the human heart, which is a socially significant issue. Outperforming the results achieved by Support Vector Machine and equivalent kernels, 83.7% of the predictions made on the results were accurate. The benchmark dataset for the Boston housing market yielded results using the spline kernel that were just as good.

vii. Latha Parthiban et al. devised a method for the prediction of cardiac illness relying on the coactive neuro-fuzzy inference system (CANFIS). By integrating the flexible neural network capabilities with the qualitative fuzzy logic approach and further integrating with genetic algorithm, the CANFIS model proved able to recognize the presence of illness. The CANFIS model's performance was evaluated depending on its training results and classification accuracy. The results showed that the CANFIS model has possibilities for forecasting cardiac disease.

8 REFERENCES

[1] C. M. Antunes, and A. L. Oliviera, Temporal Data Mining an overview, Workshop on Temporal Data Mining-7th ACM SIGKDDInt'l Conf. on Knowledge Discovery and Data Mining, (2001)

[2] H. Manilla, H. Toivonen and I. Verkamo, Discovery of frequent episodes in event sequences, Data Mining and Knowledge Discovery: An International Journal 1(3), pp. 259-289, (1997).

[3] M. A. Khaleel, S. K. Pradhan, G.N.Dash, F. A. Mazarbhuiya, ASurvey of Data Mining Techniques on Medical Data for Finding Temporally Frequent Diseases, International Journal of Advanced Research in Computer and Communication Engineering, Vol.2,Issue 12, pp. 4821-4824, (December 2013)

[4] C. Catley, H. Stratti and C. McGregor, Multi-Dimensional Temporal Abstraction and Data Mining of Medical Time Series Data: Trends and Challenges, In proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2008, pp.4322-4325¸ (February, 2008)

[5] A. Olukunle and S. Ehikioya, A Fast Algorithm for Mining Association Rules in Medical Image Data. IEEE. p1-7, (2002).[17] G. N. Pradhan and B. Prabhakaran, Association Rule Mining In Multiple, Multidimensional Time Series Medical Data. IEEE. pp.1-4, (2009).

[6] T. Revathi S. Jeevitha, "Comparative Study on Heart Disease Prediction System Using Data Mining Techniques ",International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013)

[7] Monika Gandhi, Dr.Shailendra Narayan Singh ,"Predictions in Heart Disease Using Techniques of Data Mining" , 2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE-2015), IEEE 2015.

[8] Amrender Kumar, "ARTIFICIAL NEURAL NETWORKS FOR DATA MINING" ,I.A.S.R.I., Library Avenue, Pusa, New Delhi-110 012.

[9] Nuzhat F. Shaikh, Dharmpal D. Doye,"An Adaptive Central Force Optimization (ACFO) and Feed Forward Back Propagation Neural Network (FFBNN) based iris recognition system", Journal of Intelligent and Fuzzy Systems 30 (2016) 20832094 DOI:10.3233, IOS Press,2083.

[10] R. Spence, L. Tweedie, H. Dawkes, and H. Su, "Visualization for functional design", in Proc. Int. Symp. on Information Visualization (InfoVis 95), 1995, pp. 410

[11] Nuzhat F. Shaikh, Dharmpal D. Doye, "Improving the Accuracy of Iris Recognition System using Neural Network and Particle Swarm Optimization" International Journal of Computer Applications (0975 – 8887)Volume 79 – No3, October 2013

[12] Ankita Dewan, Meghna Sharma , "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification" , 2015 2ndInternationalConference on Computing for Sustainable Global Development (INDIACom), IEEE 2015.