International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal | www.ijtsrd.com
ISSN No: 2456 - 6470 | Volume - 2 | Issue – 6 | Sep – Oct 2018

Prediction of Coronary Artery Disease Using Text Mining

Meena Preethi. B[1], Darshna. R[2], Sruthi. R[2]
[1]Assistant Professor, [2]Student
Department of BCA and MSC.SS, Sri Krishna Arts and Science College,
Coimbatore, Tamil Nadu, India

ABSTRACT
One of the commonly occurring diseases across the world is heart disease. About 60 percent of the total population gets affected by the heart disease. Among the several kinds of heart disease, coronary heart disease is dealt in this paper. The healthcare trade gathers enormous amounts of healthcare files which, regrettably, are not mined to determine hidden information for efficient assessment creation. Since enormous sum of people get exaggerated by heart disease, the patients' case history raise to a maximum extent in hospitals, as the result analyzing becomes a difficult process for medical practitioners. In this paper, an effective method to extract the data from the large amount of documents is proposed using text mining. Using text mining techniques, the required data are extracted in the structured format. This paper uses an apriori algorithm in association rule mining, which is used for frequent item set extraction and rule generation. As the result, several rules will be generated from which the disease can be predicted.

Keywords: *Coronary Heart Disease, Text Mining, Association rule mining, Apriori*

1. INTRODUCTION
The recognition of the heart disease from diverse description or signs is a reflective crisis that is not free from false assumptions and is recurrently accompanied by unprompted effects. Due to several seasonable time changes, people get affected by more and more vulnerable diseases. This can be predicted in advanced using prediction model.[1] A significant challenge to developing models for predicting cardiac risk involves the identification of temporally related events and measurements in the unstructured text in electronic health records. The 2014 i2b2 Challenges in Natural Language Processing in Clinical Data track for identifying risk factors for heart disease over time was created to facilitate development of natural language processing systems to address this challenge. Among the various techniques text mining plays an important role in the medical field.

Text mining is the process of extracting the Hidden Knowledge from the text document. Various text mining approaches are classification, clustering, association rule mining, statistical learning; all have their significance in the medical field. [18] In association rule mining Apriori algorithm is the most efficient algorithm for extracting frequent item sets of huge data. To find out the frequent item sets, minimum support and confidence value have been used. This frequent item sets helps the user to determine the diseases at the early stage and it paves way to reduce the death rate.

The Rattle data mining tool is being used for performing the tasks of analyzing the data of the patients.

2. TEXT MINING
Text mining is outlined as a information-rigorous process in which a user cooperate with the manuscript gathering using a suite of investigation tools. It deals with converting unstructured data into structured data.[17] In the medical field, text mining algorithms are used to mine the hidden knowledge in the dataset of the medical domain.[19]

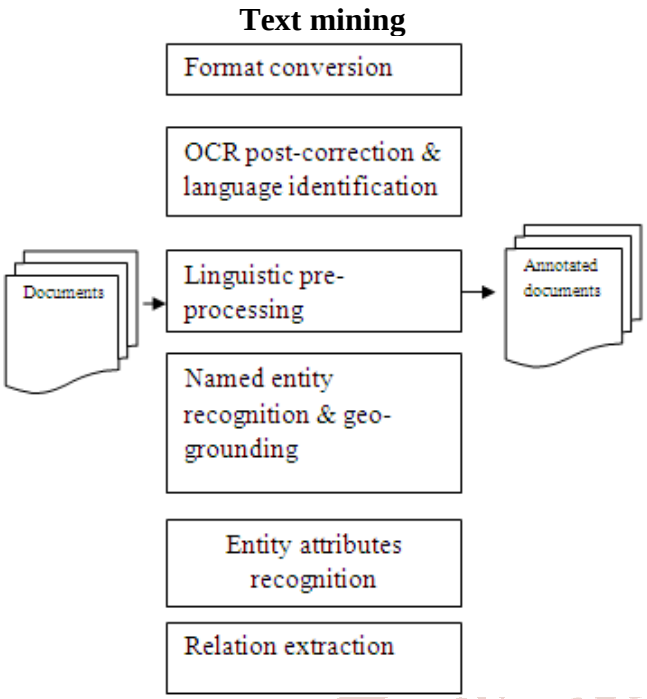

FIG1: Text mining architecture

RISK FACTORS OF CORONARY ARTERY DISEASE

There are many risk factors for CAD and some can be controlled but not others. The risk factors that can be controlled (modifiable) are: High BP; high blood cholesterol levels; smoking; diabetes; overweight or obesity; lack of physical activity; unhealthy diet and stress.[15]

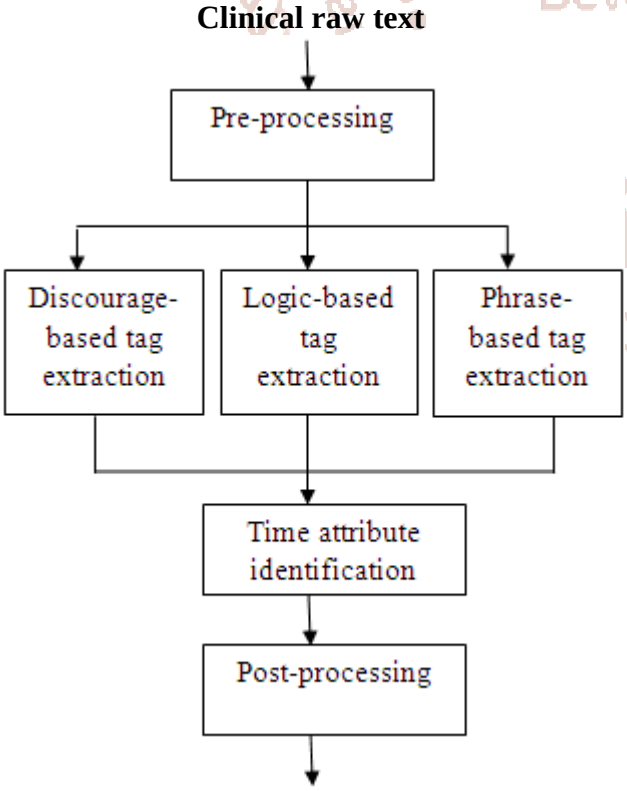

FIG2: clinical risk factors

A. Hypertension

Hypertension is one of the risks in the development of CHD. The American President Roosevelt died from cerebral hemorrhage, squeal of hypertension.[3]

Based on 20 years of surveillance of the Framingham cohort, a two-fold to threefold increased risk of clinical atherosclerotic disease was reported. It was also one of the first studies to demonstrate the higher risk of CVD in women with diabetes compared to men with diabetes. It is now accepted as a major cardiovascular risk factor. There is a clear-cut relationship between diabetes and CVD. At least 68% of inhabitants age 65 or older with diabetes die from various outline of heart disease; and 16% die of stroke.

B. Blood pressure & cholesterol

The association of Joint National Committee blood pressure and National Cholesterol Education Program cholesterol categories with coronary heart disease risk resulted that the patients were 2489 men and 2856 women 30 to 74 years old at baseline with 12 years of follow-up. [4]The target was to recognize information medically associated to heart disease threat and trail its evolution over sets of longitudinal patient medical records.

3. METHODS

A. NATURAL LANGUAGE PROCESSING

NLP defines to Artificial Intelligence method of conversing with an intelligent system using a natural language such as English.

Despite recent progress in prediction and prevention, heart disease remains a leading cause of death. One preliminary step in heart disease prediction and prevention is risk factor identification. Many studies have been proposed to identify risk factors associated with heart disease; however, none have attempted to identify all risk factors. In 2014, the National Center of Informatics for Integrating Biology and beside (i2b2) issued a clinical natural language processing (NLP) challenge that involved a track (track 2) for identifying heart illness threat factors in clinical texts over time. [2]This track intended to recognize medically appropriate information linked to heart disease risk and track the progression over sets of longitudinal patient medical records.[5] Identification of tags and attributes associated with disease presence and progression, risk factors, and medications in patient medical history were required.

The most representative work concerning clinical concept recognition is the 2010 i2b2 clinical NLP challenge, where various machine learning-based, rule-based, and hybrid methods were proposed. Phenotypes that include diseases and some observable characteristics have also been widely investigated.[6]

B. ASSOCIATION RULE:

Association rule learning is a law-based mechanism learning technique for realizing motivating associations between variables in outsized databases. It is projected to categorize strong rules discovered in databases using some process of interestingness.[7]

Let {\displaystyle I=\{i_{1},i_{2},\ldots ,i_{n}\}} I=\{i_{1},i_{2},\ldots ,i_{n}\} be a set of {\displaystyle n} n binary attributes called items.

Let {\displaystyle D=\{t_{1},t_{2},\ldots ,t_{m}\}} D=\{t_{1},t_{2},\ldots ,t_{m}\} be a set of transactions called the database.

Every contract in {\displaystyle D} D has a exclusive transaction ID and surround a subset of the items in {\displaystyle I} I.

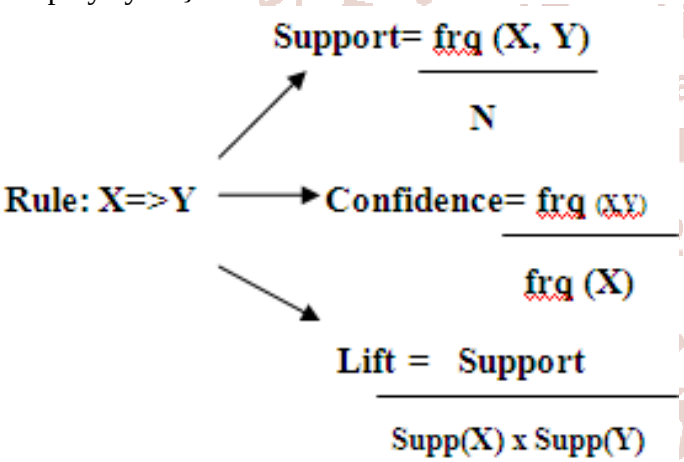

FIG3**:** Association rule

A rule is defined as an implication of the form: {\displaystyle X\Rightarrow Y} X\Rightarrow Y, where {\displaystyle X,Y\subseteq I} X,Y\subseteq I.

In Agrawal, Imieliński, Swami[2] a rule is defined only between a set and a single item, {\displaystyle X\Rightarrow i_{j}} {\displaystyle X\Rightarrow i_{j}} for {\displaystyle i_{j}\in I} {\displaystyle i_{j}\in I}.

Every rule is composed by two different sets of items, also known as itemsets, {\displaystyle X} X and {\displaystyle Y} Y, where {\displaystyle X} X is called antecedent or left-hand-side (LHS) and {\displaystyle Y} Y consequent or right-hand-side (RHS).[8]

The standard problem of mining association rules is to find all rules whose metrics are equal to or greater than some specified minimum support and minimum confidence thresholds. A k-item set with sustain more than the smallest amount threshold is called frequent. We use a third significance metric for association rules called lift : [13]

lift (X * Y) = P(Y |X)/P (Y) = confidence(X * Y)/support (Y).

Lift quantifies the predictive power of Xε Y; we are interested in rules such that lift (X * Y) > 1.

C. FP-GROWTH ALGORITHM

In Data Mining the mission of discovering repeated pattern in huge databases is exceedingly essential and has been premeditated in outsized scale in the past few years. Regrettably, this mission is systematically costly, especially when a large number of patterns survive.

The FP-Growth Algorithm, projected by Han, is a proficient and scalable method for mining the entire set of recurrent patterns by pattern section growth, using an unmitigated prefix-tree structure for stockpiling compacted and decisive information about common patterns named frequent-pattern tree (FP-tree).[12]

First it compresses the input database creating an FP-tree instance to represent frequent items. After this foremost step it partitions the compacted database into a set of provisional databases, each one linked with one numerous pattern. Finally, each such database is mined separately. Using this technique, the FP-Growth diminishes the investigated costs looking for diminutive patterns recursively and then concatenating them in the elongated recurrent patterns, offering superior selectivity.

In large databases, it's not achievable to embrace the FP-tree in the central memory. The approach to manage with this difficulty is to initial partition the record into a position of lesser databases (called projected databases), and then construct an FP-tree from each one of these smaller databases.

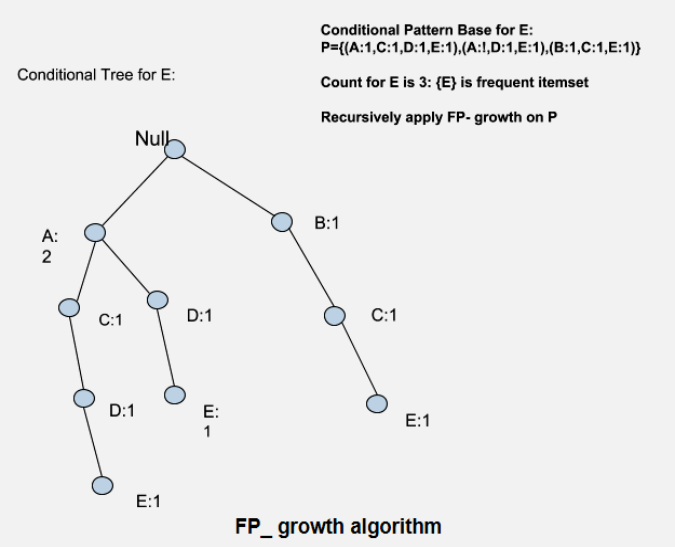

FIG4: FP- growth algorithm

D. Construction of the FP-Tree

The FP-Tree is a compacted illustration of the input. While understanding the data resource each matter t is mapped to a trail in the FP-Tree. As dissimilar transaction can have numerous objects in frequent, their path may overlie. With this it is probable to constrict the configuration. [14]

E. APRIORI ALGORITHM

Given the set of all frequent (k-1) item-sets. We want to generate superset of the set of all frequent k-item-sets. The perception behind the apriori applicant making method is that if an item-set X has smallest amount support, so do all subsets of X.[9] after all the (l+1)- applicant progression have been produced, a new scrutinize of the transactions is ongoing (they are read one-by-one) and the sustain of these new candidates is resolute.

Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as candidate generation), and groups of candidates are tested against the data. The algorithm lapses when refusal additional thriving lean-to are found.

Apriori uses breadth-first search and a Hash tree structure to count candidate item sets efficiently. It engender applicant item situate of length {\displaystyle k} k commencing item sets of length {\displaystyle k-1} k-1. Then it prunes the candidates which have an intermittent sub pattern.[10] According to the descending conclusion lemma, the applicant set include all recurrent {\displaystyle k} k-length item sets. Following that, it examines the contract database to establish recurrent item sets amongst the applicants.

The apriori principle can lessen the quantity of item sets we need to inspect. Set plainly, the apriori principle defines that if an itemset is intermittent, then all its supersets must also be intermittent.[11]

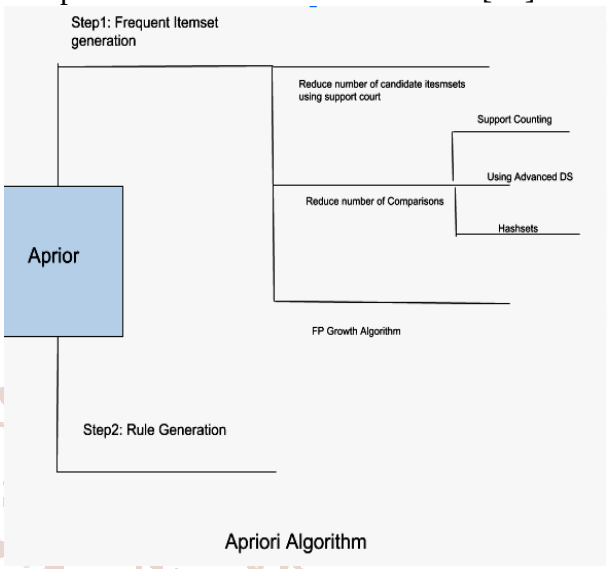

FIG5: Apriori algorithm

Finding item sets with high support:

Using the apriori principle, the number of item sets that have to be examined can be pruned, and the list of popular item sets can be obtained in these steps:[18]

Step0. Start with item sets containing just a single item.

Step1. Determine the support for item sets. Keep the item sets that meet your minimum support threshold, and remove item sets that do not.

Step2. Using the item sets you have kept from Step 1, generate all the possible item set configurations.

Step3. Repeat Steps 1 & 2 until there are no more new item sets.

4. CONCLUSION

It can be concluded from this project that if text mining is used for large amounts of text documents, the results will be accurate and efficient. It will be very easy for the users to understand. Since the apriori algorithm is used, the results are predicted accurately. The enclosure narrative hazard factors like obesity in existing risk-assessment programs like FRS are enormously compulsory as they have been proved to be univariate indicators of CHD. An primary advance would be to include the recently revealed principle into existing estimation programs, thereby, recuperating the calculation result of embryonic CHD. The Data Mining organization policy were used to forecast numerous related target elements, for heart disease diagnosis. The aim was to find organization

policy predicting healthy arteries or diseased arteries, given patient risk factors and medical dimensions. Intervention from both Government and Nongovernment organizations is necessary to properly combat the current cardiac crisis.

REFERENCE LINKS

1. https://www.sciencedirect.com/science/article/pii/S153204641500194X(intro)
2. https://bmccardiovascdisord.biomedcentral.com/articles/10.1186/s12872-017-0580-8(Natural Lang processing)
3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5686931/(risk factors)
4. https://www.ncbi.nlm.nih.gov/m/pubmed/9603539(risk factors)
5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4977226/(nlp)
6. https://www.sciencedirect.com/science/article/pii/S153204641500194X(nlp)
7. https://pdfs.semanticscholar.org/7fc5/30279dba68aecbbeb392a706bd8eb3fab0c9.pdf(association rule)
8. https://www.researchgate.net/publication/306030128_ASSOCIATION_RULE_MINING_ON_MEDICAL_DATA_TO_PREDICT_HEART_DISEASE
9. https://en.m.wikipedia.org/wiki/Apriori_algorithm
10. https://www.slideshare.net/mobile/INSOFE/apriori-algorithm-36054672
11. https://www.kdnuggets.com/2016/04/association-rules-apriori-algorithm-tutorial.html/2
12. https://link.springer.com/chapter/10.1007/978-0-387-35300-5_3
13. http://www.rroij.com/open-access/a-review-on-association-rulemining-algorithms.php?aid=43382
14. https://en.m.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Frequent_Pattern_Mining/The_FP-Growth_Algorithm.
15. “An automatic system to identify heart disease risk factors in clinical texts over time”, Qingcai Chen a, Haodi Li a, Buzhou Tang a,⇑, Xiaolong Wang a, Xin Liu a, Zengjian Liu a, Shu Liu a, Weida Wanga, Qiwen Deng b, Suisong Zhu b, Yangxin Chen c, Jingfeng Wang c, 1 September 2015.
16. “Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models”, Jay Urbain, 7 August 2015
17. “Text Mining: Natural Language techniques and Text Mining applications, *M. Rajman, R. Besan*.
18. Web-Based Heart Disease Decision Support System using Data Mining Classification Modeling Techniques Sellappan Palaniappan and Rafiah Awang.
19. https://www.researchgate.net/publication/301661056