EPH - International Journal of Science And Engineering

ISSN (Online): 2454 2016
Volume 06 Issue 04 December 2020

DOI: https://doi.org/10.53555/eijse.v6i4.104

PREDICTIVE ANALYSIS OF DISEASE USING A-PRIORI AND K- MEAN TECHNIQUE

Supriya[1]*, Asst. Prof. Manoj Kumar Singh[2]
**[1]M.Tech (CSE), BM Group of Institutions, Gurgaon*
[2]HOD (CSE), BM Group of Institutions, Gurgaon

Corresponding Author:

Abstract:-

Disease prognostication is one of the most important issues that we are facing today. A large number of patients struggle for their check-up even when it concerns of predictive diseases like heart attack possibilities, kidney damage and possibilities of lung problem. This motivates us to develop a hybrid algorithm which uses K-means and A-priori for data mining into large volumes of data and extract information that can be converted to useful knowledge and overall predict a patient for their chances of disease using a console. This console is developed with both algorithms working at back-end. This research paper is mainly focused on predicting lung and heart disease. Experimental results will show that many of the rules help in the best prediction of lung and heart disease, which even help doctors in their diagnostic decisions.

Keywords: - *Data mining, lung disease, heart disease, Apriority and K-means algorithm.*

I. INTRODUCTION

Heart and lung diseases are one of the major causes of deaths around the world [1] and major part of it is due to lack a good disease prediction system. At present doctors make disease predictions based on their learning and experience from history. The problem starts here as human intelligence itself cannot make effective predictions every time and are prone to errors, which may prove fatal for a patient. Database technologies have started to flourish and this has resulted in good record availability of large amounts of medical data regarding patients and their diseases. Data mining can be applied on these records to extract useful information for disease prediction [2] [3]. Thus an efficient prediction model consisting of data mining methods: K-means clustering and A-priori [4] is developed here for predicting those diseases. With these machine learning and data mining algorithms working on back-end, a console is made which helps to predict diseases based on information entered by a doctor or a patient who also can use it in even in the absence of a doctor. Knowledge discovery in databases is a well-defined process having several steps where data mining attributes to be a core step. Data mining results in discovery of hidden and useful information from mammoth databases via use of different perspectives. This information thus helps in providing quality services to patients i.e. patients can be diagnosed correctly and effective treatments can be provide to them. Records of millions of patients can be stored, digitalized and data mining techniques can be applied on them to answer numerous significant and critical queries related to health care. By using the hybrid of Kmeansclustering and A-priori algorithms we can find the state of lung and heart and thus can predict at what stage the disease is and even suggest doctors & patients, the effective medicines or treatments. This proves to be more efficient than using single algorithm to find useful information. Clustering is data mining approach that groups a set of nonfigurative objects into class of analogous objects. One of the leading clustering methods is K-means clustering method where each cluster is signified by the mean value of the objects in the cluster. Another approach is to find recurrent term sets from a transaction dataset and derive association rules. Finding recurrent term sets is not trivial because of its combinatorial outburst. Once recurrent term sets are obtained, it is usually straightforward to generate association rules having confidence higher than or equal to a user specified minimum confidence.

II. BACKGROUND

World health organization [1] presented the ten leading causes of death by broad income group 2008. Liao, S-C [5] presented a theory that categorical data is good for most data mining classification techniques (e.g. classification of disease and non-disease groups) and is relatively easy to use for extracting medical knowledge. In S. Vijiyarani [6] presented prediction of different types of diseases. This paper reviewed the research papers, which mainly concentrated on predicting heart disease, Diabetes and Breast cancer.My Chau Tu [7] said diagnosis of heart disease is important issue, prompting many researchers to work on development of intelligent medical decision support systems to improve the ability of physicians. M.Akhil Jabbar [8] presented experimental results, which show that most of the classifier rules help in the best prediction of heart disease, which even helps doctors in their diagnostic decisions.

N. Aditya Sundar [9] presented a training tool to train nurses and medical students to diagnose patients with heart disease. It was a user-friendly system and can be used in hospitals if they have a data warehouse for their hospital. C Y Hsu [10] presented few studies that defined how the risk of hospital acquired acute renal failure varies with the level of estimated glomerular filtration rate. Mohammed Abdul Khaleel [11] presented a methodology to discover locally frequent diseases with the help of A-priori data mining technique. Chris Ding [12] presented a Mapping data points into a higher dimensional space via kernels and showed that Kernel PCA gives solution for Kernel K-means. On learning, their results suggested effective techniques for K-means clustering. DNA gene expression and Internet newsgroups were analyzed to illustrate the results. K.R. Lakshmi [13] said diagnosis of heart disease is a significant and tedious task in medicine. The study describes algorithmic discussion of the heart disease dataset from Cleveland Heart Disease database, on line repository of large datasets. It showed better results in structural and functional based gene classification.

III. ALGORITHMS USED

The k -means algorithm:

Stuart Lloyd proposed this algorithm in 1957 as a technique for pulse-code modulation. [14] K-means is an unsupervised learning algorithm that solves the clustering problem. The procedure is simple and easy in the way that it classifies a given dataset through a certain „k‟ (assumption) clusters, which are fixed in starting.

The algorithm is composed of the following steps:

- Place K points into the space characterized by the objects that are being clustered. These points represent initial group centroids and are defined in a canny (better placed far away from each other) way because different locations cause different results.
- Assign each object to the group that has the closest centroid.
- When all objects have been assigned, recalculate the positions of the K centroids.
- Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

This algorithm aims at minimizing an objective function; in this case a squared error function. The objective function is:

$$k \quad n$$

$$J = x_i\, j - c_{j\, 2} \quad j=1\ i=1$$

Where, $x_i{}^{j} - c_j{}^{2}$ is a chosen distance measure between a data point x_i j and the cluster center c_j, is an indicator of the distance of the n data points from their respective cluster centers.

The A -priori algorithm:

A-priori is an algorithm proposed by R. Agrawal and R Srikant [15] for mining frequent term sets for Boolean association rule.

General Process

Association rule generation is usually split up into two separate steps:

- First, minimum support is applied to find all □ frequent item sets in a database.
- Second, these frequent item sets and the minimum confidence constraint are used to form rules.

The first step needs more attention as second is relatively easy. Finding all frequent item sets in a database is difficult since it involves searching all possible item sets (item combinations). The set of possible item sets is the power set over I and has size $2^n - 1$ (excluding the empty set which is not a valid item set). Although the size of the power set grows exponentially in the number of items n in I, efficient search is possible using the downward-closure property of support (also called anti-monotonicity) which guarantees that for a frequent item set, all its subsets are also frequent and thus for an infrequent item set, all its supersets must also be infrequent. Exploiting this property, Apriority can find all frequent item sets.

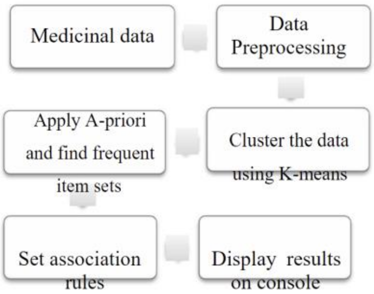
Pseudo -code:

```
Apriority (T, min Support) // T is the database and min Support is the minimum support
{
 $L_1$= {frequent items}; for (k= 2; $L_{k-1}$! =$\emptyset$; k++)
{
        Ck = candidates generated from Lk-1
//that is cartesian product Lk-1 x //Lk-1 and
            eliminating any k-1 size //item set that is not
            // frequent
            for each transaction t in database do
            {
#increment the count of all candidates in Ck that are

contained in t
                Lk = candidates in Ck with min Support
         }
         }
        Return $\cup$k Lk;
}
```

As is common in association rule mining, given a set of item sets (for instance, sets of retail transactions, each listing individual items purchased), the algorithm attempts to find subsets which are common to at least a minimum number C of the item sets. A priori 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 terminates when no further successful extensions are found. A priori uses breadth-first search and a tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length k − 1. Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates. Candidate generation generates large numbers of subsets (the algorithm attempts to load up the candidate set with as many as possible before each scan). Bottom-up subset exploration (essentially a breadth-first traversal of the subset lattice) finds any maximal subset S only after all $2^{|S|} - 1$ of its proper subsets.

IV. PROPOSED WORK


Medicinal data
Data Preprocessing
Cluster the data using K-means
Apply A-priori and find frequent item sets
Set association rules
Display results on console

A. Data per -processing:

Initially data pre-processing is done on medicinal data to reduce to attributes that we need also known as dimensionality reduction of dataset.

B. K -means application:

Next each attribute is fed individually with centrality and frequency and random cluster centroid intialization into Kmeans clustering algorithm which clusters the fed information into similar types based on which the attribute/factor is categorized like low, mid and high. Now these categories become sub -attributes

C. A priori application:

All these attributes and sub-attributes are now used in a transaction table as columns and transactions as rows& the respective column of each attribute is either marked with 1 or 0 based on the item belonging or not. This transaction table which is actually a matrix is now fed to A-priori algorithm which finds out the frequent item sets and thus the thresholds.

D. Association rules:

These frequent item sets then help in generating association rules and procedure is:

- For each frequent itemset “l”, generate all nonempty subsets of l.
- For every nonempty subset s of l, output the rule “s → (l-s)” if support count (l) / support count (s) >= min_conf where min_conf is minimum confidence threshold

E. Prediction:

Finally we predict the risk level inside the console, which is hard-coded with both algorithms working in it. Risk levels are chosen as either low, or medium or high and some medications and advices may be rendered accordingly. Our model at last was tested with 100 reports and 86 of them were found to be right which evaluates to 86% correct decisions, which indeed is very exciting for our hybrid model.

V. CONCLUSION

In this paper, we have presented a Heart and Lung Disease Diagnostic System using data mining techniques. First clustering is done using K-means clustering algorithm. Apriori algorithm is then used to find the frequent item sets.
Defining the clusters and then using A-priori on them improves accuracy of disease prediction and lessens the diagnosis cost .It provides the user or patient a console to monitor themselves for their report. It is also capable of providing the suggestions for medicine if a problem is found and can also suggest consulting a doctor in critical situation.

VI.RESULTS AND DISCUSSIONS

Using K -means and Apriori The data for heart disease prediction was collected from various corporate hospitals and opinion from expert doctors. The attributes chosen for Kidney function test are age, Blood pressure, PCR, ACR, GFR.The attributes chosen for Age, chest pain, Blood pressure, Cholesterol, Thyroid, Stress Pulse, and Exang. These are chosen because these are compulsory to find different stages of the disease. The kidney is divided into two stages normal condition, destroyed. Which means that the kidney is function properly or not. Similarly, stages for heart diseases are low risk, medium risk and high risk stages for heart diseases are low risk, medium risk and high risk. For this in this thesis algorithm used are K-means and Apriori. By using both these algorithm we can find an efficient way to find out the stage at which the disease exist. Simulation Model

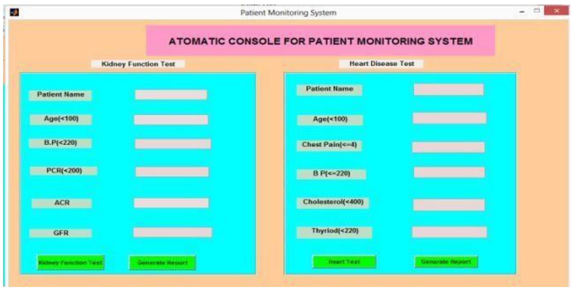
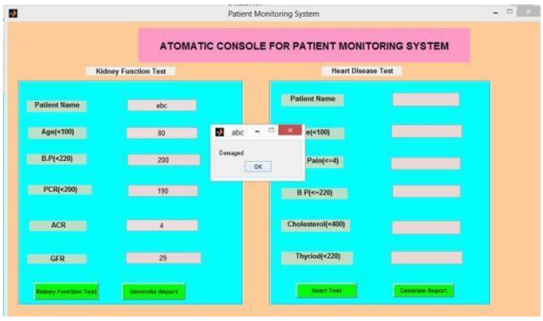


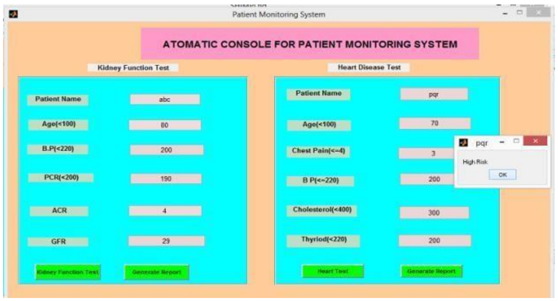
Figure 5 - Graphical User Interface developed for diseases prediction by mean of clustering & apriori algorithm



Patient Name: abc
Age(<100): 80
B.P(<220): 200
PCR(<200): 190
ACR: 4
GFR: 29

abc
Demaged
OK

Figure 6:-Kidney Function Result

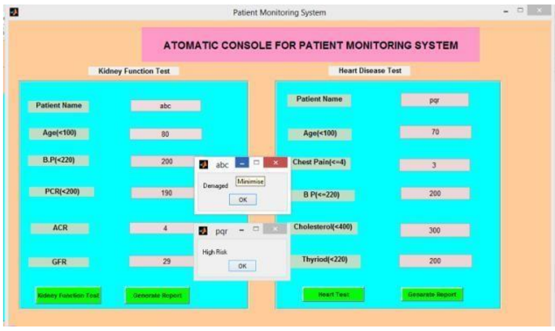


Patient Name: abc
Age(<100): 80
B.P(<220): 200
PCR(<200): 190
ACR: 4
GFR: 29

Patient Name: pqr
Age(<100): 70
Chest Pain(<=4): 3
B P(<=220): 200
Cholesterol(<400): 300
Thyriod(<220): 200

pqr
High Risk
OK

Figure 7:-Heart Diseases Test



Patient Name: abc
Age(<100): 80
B.P(<220): 200
PCR(<200): 190
ACR: 4
GFR: 29

Patient Name: pqr
Age(<100): 70
Chest Pain(<=4): 3
B P(<=220): 200
Cholesterol(<400): 300
Thyriod(<220): 200

abc
Demaged
Minimise
OK

pqr
High Risk
OK

Figure 8:-Kidney Function & Heart Disease Test

Table 5 Testing Under simulation

Serial No.	Patient Name	Age	Chest Pain	Bp	Cholesterol	Thyriod	Actual Report	Our Model Test	Match
1	M1	99	4	180	380	200	Hishest risk	Highest	Yes
2	M2	32	2	150	350	200	Highest risk	Medium	No
3	M3	2	4	110	150	50	Low risk	Low	yes
4	M4	82	3	170	399	219	Highest risk	Highest	yes
5	M5	26	4	210	300	200	Medium risk	Highest	no
6	M6	24	1	123	200	100	Low risk	Low	yes
7	M7	26	1	120	200	100	Low risk	Low	yes

The above result shows the good accuracy to detect the heart disease
Probability of detecting the disease as per actual report P (e) =N(e)/N(s)=5/7= .71
In term of percentage 71 % which is very exciting for our model.

VII. FUTURE SCOPE

More reports can be fed to improve accuracy of the algorithms as machine-learning algorithms learn more when fed with more examples (cases here). Amalgamation of more optimization techniques may further improve accuracy .A large part of remote population needs doctor. But their deficiency creates problem so console does play an important role as it facilitates the users to predict the heart and lung conditions by them self even if they are at remote location and it is hard for them to reach doctors regularly. Thus it will be less costly and time saving method if integrated to web portals.

REFERENCES

[1]. Jyoti Soni, Sunitha Soni. Predictive data mining for Medical Diagnosis: An Overview of Heart Disease Prediction; International Journal of Compute Applications (0975-8887) Volume 17-No.8, March 2011.

[2]. Fariba Shadabi, Dharmendra Sharma, Artificial Intelligence and Data Mining Techniques in Medicine-Success Stories, International Conference on Bio-Medical Engineering & Informatics, vol. 1, pp.235-239, 2008

[3]. Bodon.F. A Fast A-priori Implementation, FIMI‟03, November 2003.

[4]. Liao, S.C. and I.N. Lee, Appropriate medical data categorization for data mining classification techniques. MED. INFORM., 2002. Vol. 27, no. 1, 59–67.

[5]. S.Vijiyarani, Disease Prediction in Data Mining Technique–ijcait, Vol. II, Issue I, January 2013 (ISSN: 2278-7720)

[6]. My Chau Tu, Dongil Shin, Dongkyoo Shin, Effective Diagnosis of Heart Disease through Bagging Approach, 2nd International Conference on Biomedical Engineering and Informatics, 2009.

[7]. M.Akhil Jabbar, Heart Disease Prediction System using Associative Classification and Genetic Algorithm. ICECIT, 2012

[8]. N. Aditya Sundar, Performance analysis of classification data mining techniques over heart disease database, Volume-2, Issue-3, 470 – 478.

[9]. C Y Hsu, J D Ordonez. The risk of acute renal failure in patients with chronic kidney disease. 2 April 2008

[10]. Mohammed Abdul Khaleel, Sateesh Kumar Pradhan J Finding Locally Frequent Diseases Using Modified A-priori Algorithm, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 10, October 2013.

[11]. Chris Ding, Xiaofeng He,K-means Clustering via Principal Component Analysis, Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 947208

[12]. K.R. Lakshmi, Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability .International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013.

[13]. Lloyd, S. P. (1982). "Least squares quantization in PCM.IEEE Transactions on Information Theory28 (2): 129–137.

[14]. R. Agrawal and R. Srikant-Fast algorithms for mining association rules. In 1994. pp. 487-49 VLDB'94, Santiago, Chile, Sept.1994. pp. 487-49.