International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 4 | ISSN : 2456-3307

Chronic Kidney Disease Analysis using Data Mining

Sunil D, Prof. B. P. Sowmya

Department of Master of Computer Application, PES College of Engineering, Mandya, Karnataka, India

ABSTRACT

Data mining has been a current trend for attaining Diagnostic results. Huge amount of unmined data is collected by the healthcare industry in order to discover hidden information for effective diagnosis and decision making. Data mining is the process of extracting hidden information from massive dataset, categorizing valid and unique patterns in data. There are many data mining techniques like clustering, classification, association Analysis, regression etc. The objective of our paper is to predict Chronic Kidney Disease (CKD) using classification techniques Like Naive Bayes and Artificial Neural Network (ANN). The Experimental results implemented in Rapidminer tool show that Naive Bayes produce more accurate results than Artificial Neural Network.

Keywords : Data mining, Classification, Chronic Kidney, disease, Naive Bayes, Artificial Neural Network.

I. INTRODUCTION

Chronic kidney disease (CKD) has become a global health Issue and is an area of concern. It is a condition where Kidneys become damaged and cannot filter toxic wastes in the body. Our work predominantly focuses on detecting life Threatening diseases like Chronic Kidney Disease (CKD) Using Classification algorithms like Naive Bayes and Artificial Neural Network (ANN).

II. EXISTING SYSTEM

Features of existing system

- Nowadays, health care industries are providing several benefits like fraud detection in health insurance.
- Disease detection is also one of the significant areas of research in medical.
- In medical facilities to patients at inexpensive prices, identification of smarter treatment methodologies.
- To construction of effective healthcare policies, effective hospital resource management, better customer relation, improved patient care and hospital infection control.

Drawbacks of existing system

- Extraction of irrelevant information
- Less Reliable
- Less Efficient
- Manual Approach
- Requires Medical Equipments
- More Expensive
- Lack of user satisfaction
- Less Efficient
- Less Accurate

III. PROPOSED SYSTEM

Chronic kidney disease (CKD) has become a global health issue and is an area of concern. It is a condition where kidneys become damaged and cannot filter toxic wastes in the body. Our work predominantly focuses on detecting life threatening diseases like Chronic Kidney Disease (CKD) using Classification algorithms. Proposed system is an automation for chronic kidney disease prediction using classification technique “naïve bayes” and artificial neural network technique

IV. Concepts Under Study

Nowadays, health care industries are providing several benefits like fraud detection in health insurance, availability of medical facilities to patients at inexpensive prices, identification of smarter treatment methodologies, construction of effective healthcare policies, effective hospital resource management, better customer relation, improved patient care and hospital infection control. Disease detection is also one of the significant areas of research in medical.

Data mining approaches have become essential for healthcare industry in making decisions based on the analysis of the massive clinical data. Data mining is the process of extracting hidden information from massive dataset. Techniques like classification, clustering, regression and association have been used by in medical field to detect and predict disease progression and to make decision regarding patient's treatment. Classification is a supervised learning approach that assign objects in a collection to target classes. It is the process which classifies the objects or data into groups, the members of which have one or more characteristic in common. The techniques of classification are SVM, decision tree, Naive Bayes, ANN etc. Clustering involves grouping of objects of similar kinds together in a group or cluster. Some of its techniques include K-means, Kmedoids, agglomerative, divisive, DBSCAN etc. Association states the probability of occurrence of Association states the probability of occurrence of items in a set. Apriori is an example of association.

Figure 1 describes about various data mining techniques used over last 15 years for investigating various diseases.

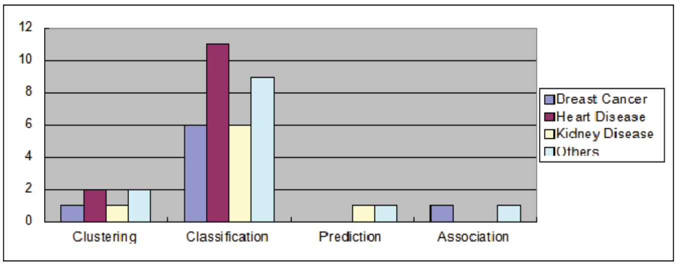

Figure 2. Data Mining techniques used for disease detection unwanted waste from blood causing smooth functioning of body organs.

S.No	Author	Year	Disease	Technique
1	Ju-Hsin Tsai [1]	2008	Cancer breast	CLUSTERING (AGNES)
2	Mostafa Ghannad Rezaie et al[2]	2008	Temporal	SVM(classification)
3	Jenn-Lung Su et al. [3]	2001	breast tumour	Bayesian Network , DT,
4	Paolo Bonato et al. [4]	2004	Parkinson	clustering
5	S Wang et al,[5]	2005	breast cancer	decision tree
6	Yanwei Xing et al. [6]	2007	coronary heart	SVM ,ANN ,DT
7	Sellappan Palaniappan et al. [7]	2008	heart disease	(DT, naive bayes,ANN)
8	Heon Gyu Lee et al. [8]	2008	coronary heart	classification (SVM)
9	K.Srinivas et al. [9]	2010	heart disease	DT, Naïve Bayes, ANN
10	Narin Watanasusin [10]	2011	ear	ANN,Naive Bayes
11	Debabrata Pal et al. [11]	2011	heart disease	Classification (DT)
12	T.John Peter [12]	2012	heart disease	DT, NB, K-NN and NN
13	Jenn-Long Liu [13]	2012	cardiac	GA, K-Means algorithm
14	Geeta Yadav [14]	2012	Parkinson	DT,Regression, SVM
15	M. Ilayaraja [15]	2013	multiple	Apriori algorithm
16	Sivagowry .S et al. [16]	2013	heart disease	Classification (DT, ANN)
17	K. Vasantha Kokilam[17]	2012	genetic	clustering &
18	Syed Umar Amin et al. [18]	2013	heart disease	genetic neural network
19	Girija D.K [19]	2013	fibroid	ANN
20	Juliet Rani Rajan [20]	2013	lung cancer	ANN
21	Sa'diyah Noor Novita Alfisahrin	2013	liver	DT,Naive Bayes
22	Ranganatha S. et al. [22]	2013	heart disease	ID3,Naive Bayes
23	Yukti Agarwal [23]	2014	eye disease	Fuzzy logic , ANN
24	M.A.Nishara Banu [24]	2014	heart disease	k-means,c4.5
25	X Xiong et. al [25]	2005	Breast Cancer	DT,association rules
26	Susan Maskery et al. [26]	2006	Breast Cancer	Bayesian network
27	Menolascina F et al. [27]	2007	Breast Cancer	J48 and Naïve Bayes
28	Qi Fan,Chang-jie Zhu [28]	2010	Breast Cancer	Pre-classification method
29	Abdelaal [29]	2010	Breast cancer	Classification (SVM),DT
30	Vijayarani, S. Et al.[30]	2015	Kidney	SVM and
31	Chiu, R. K et al. [31]	2012	Kidney	ANN
32	Lakshmi, K. R et al. [32]	2014	Kidney	(classification)DT,ANN,
33	Xun, L et. Al [33]	2010	Kidney	ANN
34	Ravindra, B. V. et al. [34]	2014	Kidney	K-means clustering
35	Ahmed, S et al. [35]	2014	Kidney	Fuzzy Logic

Table 1. Data Mining Techniques used for Disease detection

CKD is a condition that describes loss of kidney function over time making it difficult for them to filter poisonous wastes from the body. Researchers in their recent study have addressed the use of data mining techniques for CKD detection .

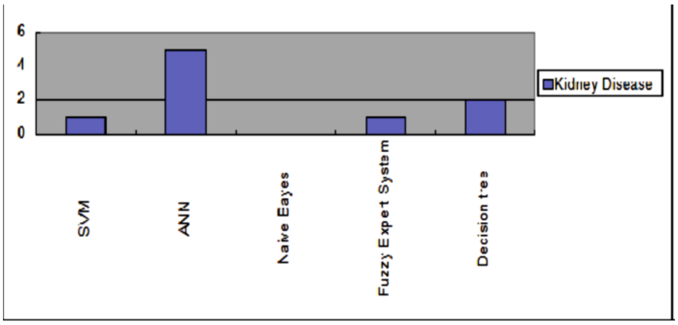

Figure 3. Classification techniques used for detecting kidney disease

It has been observed that classification algorithms have widely been used for identifying and investigating kidney disease. Figure 3 shows that many research work has been conducted using ANN while other techniques like SVM, Fuzzy logic has been used the least. It has also been observed that Naive Bayes has rarely been used. In this research work Naive Bayes approach, an important classification algorithm which uses Bayes Theorem has been used. It is particularly suited when the dimensionality of inputs is high. In this work the dimensionality of dataset is 25.

The performance of Naive Bayes has also been compared with ANN algorithm. Naive Bayes is a probabilistic classifier based on Bayes Theorem. It assumes variables are independent of each other.
The algorithm is easy to build and works well with huge data Sets. It has been used because it makes use of small training Data to estimate the parameters important for classification.

Bayes Theorem states the following:

P (A|X) = P (X|A) ·P(A) / P(X).
P(X) is constant for all classes.
P(A) = relative frequency of class A samples a such that p is
increased=c Such that P (X|A) P(A) is increased
Problem: computing P (X|A)

V. METHODOLOGY

Classification Rules

Classification is a process of finding a model (or function) that describes and distinguishes data classes or concepts. The model is derived based on the analysis of a set of training data (i.e., data objects for which the class labels are known). The model is used to predict the class label of objects for which the class label is unknown.

Naïve Bayes Algorithm Steps

Step 1: Scan the dataset (storage servers)
Retrieval of required data for mining from the servers such as database, cloud, excel sheet etc.

Step 2: Calculate the probability of each attribute value. [n, n_c, m, p]
Here for each attribute we calculate the probability of occurrence using the following formula. (Mentioned in the next step). For each class (disease) we should apply the formulae.

Step 3: Apply the formulae

P(attributevalue(ai)/subjectvaluevj)=(n_c + mp)/(n+m)
Where:
n = the number of training examples for which v = vj
nc = number of examples for which v = vj and a = ai
p = a priori estimate for P(aijvj)
m = the equivalent sample size

Step 4: Multiply the probabilities by p
For each class, here we multiple the results of each attribute with p and final results are used for classification.

Step 5: Compare the values and classify the attribute values to one of the predefined set of class.

Sample Example
Attributes (Constraints) – S1, S2, S3 [m=3]
Subject (Disease) – CKD, NOT CKD [p=1/2=0.5]

Training Dataset

Patient Name	S1 (X,Y,Z)	S2 (A,B,C)	S3 (P,Q,R)	Disease (subject)
Anil	X	A	P	CKD
Ajay	X	B	Q	CKD
Arun	Y	B	P	NOT CKD
Kumar	Z	A	R	CKD
Naveen	Z	C	R	NOT CKD

New Patient data – Akash Constraints (S1 -X,S2-A,S3-R) Disease – CKD / NOT CKD
P=[n_c + (m*p)]/(n+m)

CKD	NOT CKD
X P=[n_c+(m*p)]/(n+m) n=2, n_c=2,m=3,p=0.5 p=[2+(3*0.5)]/(2+3) p=0.7	X P=[n_c+(m*p)]/(n+m) n=2, n_c=0,m=3,p=0.5 p=[0+(3*0.5)]/(2+3) p=0.3
A P=[n_c+(m*p)]/(n+m) n=2, n_c=2,m=3,p=0.5 p=[2+(3*0.5)]/(2+3) p=0.7	A P=[n_c+(m*p)]/(n+m) n=2, n_c=2,m=3,p=0.5 p=[2+(3*0.5)]/(2+3) p=0.3
R P=[n_c+(m*p)]/(n+m) n=2, n_c=1,m=3,p=0.5 p=[1+(3*0.5)]/(2+3) p=0.5	R P=[n_c+(m*p)]/(n+m) n=2, n_c=1,m=3,p=0.5 p=[1+(3*0.5)]/(2+3) p=0.5

CKD – 0.7 * 0.7 * 0.5 * 0.5 (p) =0.1225

NOT CKD – 0.3 * 0.3 * 0.5 * 0.5 (p) =0.0225
Since CKD > NOT CKD

So this new patient is classified to CKD

VI. CONCLUSION

Recommendation systems are very useful and powerful tool used to make Chronic Kidney Disease has been predicted and diagnosed using data mining classifiers: ANN and Naive Bayes. Performances of these algorithms are compared using Rapidminer tool. The obtained results showed that Naïve Bayes is the most accurate classifier with 100% accuracy When compared to ANN having 72.73% accuracy. In this Research study, some of the factors considered were age, Diabetes, blood pressure, RBC counts etc. The work can be Extended by considering other parameters like food type, Working environment, living conditions, availability of clean Water, environmental factors etc for kidney disease detection. Further studies can be conducted using other classifiers like Fuzzy logic, KNN.

VII. REFERENCES

[1]. Tsai, J. H. (2008). Data Mining for DNA Viruses with Breas Cancer and its Limitation. INTECH Open Access Publisher.

[2]. Ghannad-Rezaie, M., & Soltanian-Zadeh, H. (2008). Interactive Knowledge discovery for temporal lobe epilepsy. INTECH Open Access Publisher.

[3]. Su, J. L., Wu, G. Z., & Chao, I. P. (2001). The approach of data mining methods for medical database. In Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE (Vol. 4, pp. 3824-3826). IEEE.

[4]. Bonato, P., Sherrill, D. M., Standaert, D. G., Salles, S. S., & Akay, M. (2004, September). Data mining techniques to detect motor fluctuations in Parkinson's disease. In Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE (Vol. 2, pp. 4766-4769). IEEE.

[5]. Wang, S., Zhou, M., & Geng, G. (2005). Application of fuzzy cluster analysis for medical image data mining. Mechatronics and Automation, 2, 631-636.

[6]. Xing, Y., Wang, J., Zhao, Z., & Gao, Y. (2007, November). Combination data mining methods with new medical data to predicting Outcome of coronary heart disease. In Convergence Information Technology, 2007. International Conference on (pp. 868-872). IEEE.

[7]. Palaniappan, S., & Awang, R. (2008, March). Intelligent heart disease prediction system using data mining techniques. In Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on (pp. 108-115). IEEE.

[8]. Lee, H. G., Noh, K. Y., & Ryu, K. H. (2008, May). A data mining approach for coronary heart disease prediction using HRV features and carotid arterial wall thickness. In BioMedical Engineering and Informatics, 2008. BMEI 2008. International Conference on (Vol. 1, pp.200-206). IEEE.

[9]. Srinivas, K., Rao, G. R., & Govardhan, A. (2010, August). Analysis ofcoronary heart disease and prediction of heart attack in coal miningregions using data mining techniques. In Computer Science andEducation (ICCSE), 2010 5th International Conferenceon (pp. 1344-1349). IEEE.

[10]. Watanasusin, N., & Sanguansintukul, S. (2011, August). Classifyingchief complaint in ear diseases using data mining techniques. In Digital Content, Multimedia Technology and its Applications (IDCTA), 20117th International Conference on (pp. 149-153). IEEE.