CKD AND ITS STAGE PREDICTION USING NAVIE BAYES AND C4.5 ALGORITHM

Tojo Mathew
Associate Professor, Dept. of CSE
National Institute of Engineering, Mysuru

Shobhitha Lankeshwar, Tanisha P Prasad, Triveni S
Computer Science and Engineering,
National Institute of Engineering, Mysuru

Abstract

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 project is to project Chronic Kidney Disease(CKD) using classification techniques like Naive Bayes and Artificial Neural Network(ANN). The experimental results implemented showed that Naive Bayes produce more accurate results than Artificial Neural Network.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. Naive Bayes classifies the patients to one of the two classes: CKD or NOT CKD. The Artificial Neural Network (ANN) is a computational model inspired by structure and function of biological neural network. It is an interconnection of artificial neurons that processes information using connected links. It has been used as it works well with noisy data and processes both numeric and categorical data.The current system uses GFR technique for stage prediction, which considers only three constraints for the prediction of the stage , hence less accurate. Stage prediction is one of the challenging tasks in today's medical field and it is the area of concern. Proposed system uses C4.5 technique for the stage prediction, which considers all the 24 constraints unlike GFR technique.

Index Terms- *Data mining , Naive bayes technique, Artificial Neural Network and C4.5 technique*

1.INTRODUCTION

Data Mining is one of the most encouraging areas of research with the purpose of finding useful information from voluminous data sets. It has been used in many domains like image mining, opinion mining, web mining, text mining, graph mining etc. Its applications include anomaly detection, financial data analysis, medical data analysis, social network analysis, market analysis etc. It has become popular in health organization as there is a requirement of analytical methodology for predicting and finding unknown patterns and information in health data. It plays a vital role for discovering new trends in healthcare industry. Data Mining is particularly useful in medical field when no availability of evidence favoring a particular treatment option is found. Large amount of complex data is being generated by healthcare industry about patients, diseases, hospitals, medical equipments, claims, treatment cost etc. that requires processing and analysis for knowledge extraction.

Data mining comes up with a set of tools and techniques which when applied to this processed data, provides knowledge to healthcare professionals for making appropriate decisions and enhancing the performance of patient management tasks. Patients with similar health issues can be grouped together and effective treatment plans could be suggested based on patient’s history, physical examination, diagnosis and previous treatment patterns. 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)

Naive bayes retrieves the data values from the servers, then it calculates the probability of data values for both CKD and NOT CKD. The two probabilities are then compared, if CKD value is larger than NOT CKD then the patient is classified into CKD class. C4.5 gets the input from naive bayes , then for the particular constraint maximum occurrence is found out, the stage corresponding to it is the predicted stage.

2. LITERATURE SURVEY

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 items in a set.

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. This research work mainly focuses on chronic kidney disease stage prediction using C4.5 technique.

3. METHODOLOGY

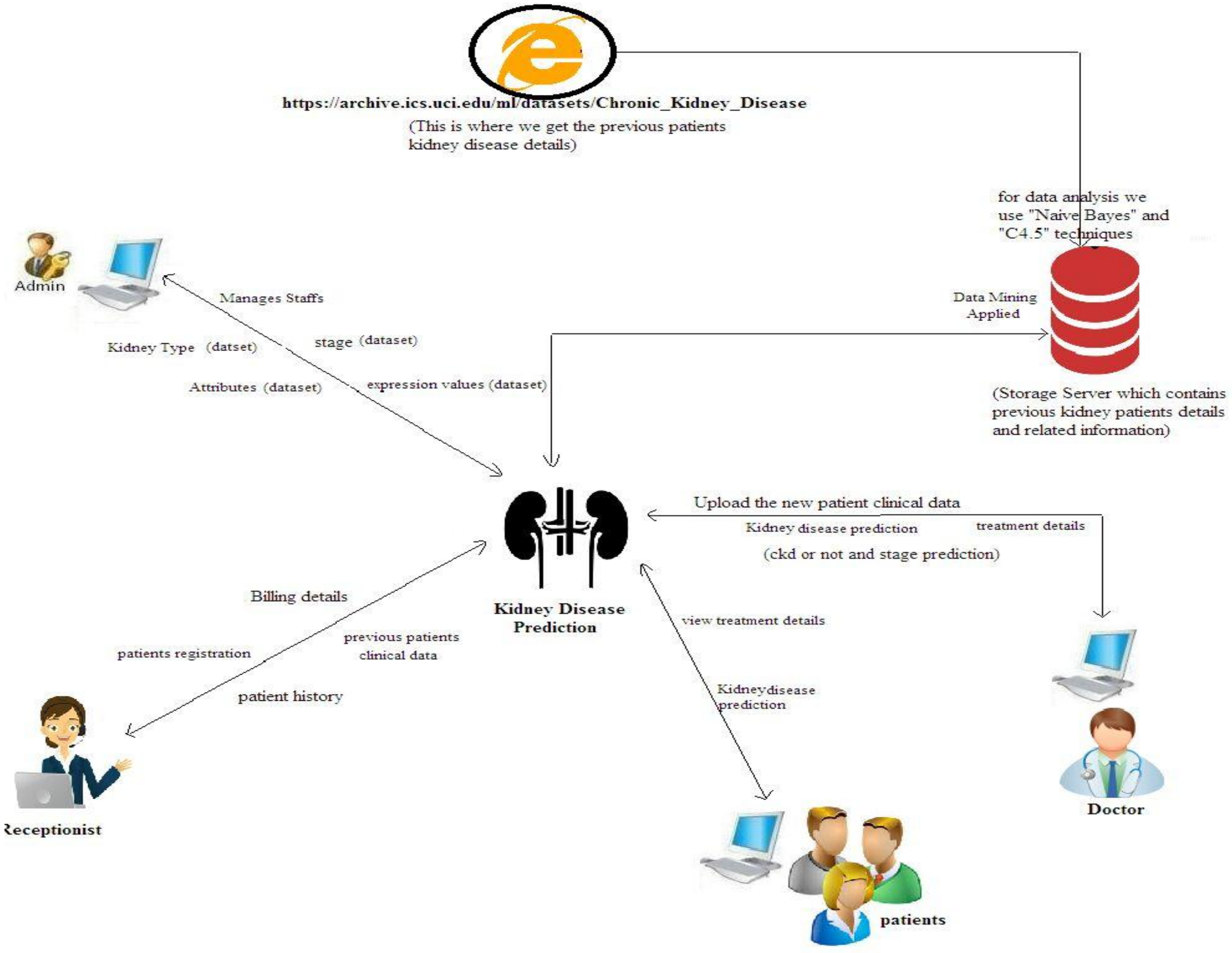
https://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease
(This is where we get the previous patients kidney disease details)
for data analysis we use "Naive Bayes" and "C4.5" techniques
Admin
Manages Staffs
Data Mining Applied
Kidney Type (datset)
stage (dataset)
Attributes (dataset)
expression values (dataset)
(Storage Server which contains previous kidney patients details and related information)
Upload the new patient clinical data
Kidney disease prediction
treatment details
(ckd or not and stage prediction)
Billing details
Kidney Disease Prediction
view treatment details
previous patients clinical data
patients registration
patient history
Kidneydisease prediction
Doctor
Receptionist
patients

FIG -1 Architecture diagram of the actors

Actors:

Admin - Administrator is a one who maintains the entire application. Administrator is a owner of the application. Admin can login module, add doctors and receptionists, set Id and password of staffs, add diseases types – dataset (chronic kidney disease), add diseases – dataset (subtypes), add constraints [types and values] – [dataset]

Doctor - Doctor is a one who specifies the necessary inputs for chronic kidney disease prediction. Doctor is a service receiver. The key service given by the system is “chronic kidney disease prediction” based on the medical data. Doctor can login module, upload patient data [clinical Data] old and new patient, chronic kidney disease prediction module [new patient – Navie Bayes Algorithm], upload treatment details, view patients history

Receptionist - Receptionist is one who maintains the patients registration, billing and treatment details. Receptionist can login module, patient registration (set Id and pwd for patients], manage patients history

Patient - Patient is a one who receives the services from the application. Patient can login module, patient history, treatment module

3.1 NAVIE BAYES ALGORITHM

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.

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

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. Compare the values and classify the attribute values to one of the predefined set of class. 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) NOT CKD – 0.3 * 0.3 * 0.5 * 0.5 (p)

=0.1225 =0.0225

Since CKD > NOT CKD the new patient is classified to CKD

3.2 C4.5 ALGORITHM

Scan the dataset (storage servers), for each attribute a, calculate the gain [number of occurrences] Let a_best be the attribute of highest gain [highest count] Create a decision node based on a_best – retrieval of nodes[patient] where the attribute values matches with a_best. recur on the sub-lists [list of patient] and calculate the count of outcomes[Stages] – termed as subnodes. Based on the highest count we classify the new node. Attributes(Features) – F1,F2,F3 [m=3] Subject (stages) – S1,S2 [p=1/2=0.5]

Training Dataset

Name	F1(X,Y,Z)	F2(A,B,C)	F3(P,Q,R)	Stage (subject)
Anil	X	A	P	S1
Kumar	X	B	Q	S1
Ajay	Y	B	P	S2

Naveen	Z	A	R	S1
Akash	Z	A	Q	S2

New Patient Features – Akul F1-X,F2-A,F3-R Which Stage - ?

Feature Count (X) in the dataset = 2, Feature Count (A) in the dataset = 3, Feature Count (R) in the dataset = 1

Sort(); Reverse();

Feature	Count
A	3
X	2
R	1

A – S1 (2) & S2(1); This algorithm is based on single attribute values.

Output

Stage	Priority
S1	2
S2	1

4. CONCLUSION

In this article, we study that using data mining prediction of the CKD can be determined within seconds instead of waiting for a long time for the reports.The modules are defined in such a way that all the treatment report of the particular patient can be written in the system. Since its an web based application, it can accessed anywhere using user id and password. Not only the prediction of the CKD buut also the stage of the CKD in positive patient is classified using C4.5 algorithm. Hence we conclude that within no time the report is ready with the stage prediction for the usefulness regarding to medical field.

5.REFERENCES

[1] Tsai, J. H. (2008). Data Mining for DNA Viruses with Breast 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 of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. In Computer Science and Education (ICCSE), 2010 5th International Conference on (pp. 1344- 1349). IEEE.

[10] Watanasusin, N., & Sanguansintukul, S. (2011, August). Classifying chief complaint in ear diseases using data mining techniques. In Digital Content, Multimedia Technology and its Applications (IDCTA), 2011 7th International Conference on (pp. 149-153). IEEE.

[11] Pal, D., Chakraborty, C., & Mandana, K. M. (2011, November). Data mining approach for coronary artery disease screening. In Image Information Processing (ICIIP), 2011 International Conference on (pp. 1-6). IEEE.

[12] Peter, T. J., & Somasundaram, K. (2012, March). An empirical study on prediction of heart disease using classification data mining techniques. InAdvances in Engineering, Science and Management (ICAESM), 2012 International Conference on (pp. 514-518). IEEE.

[13] Liu, J. L., Hsu, Y. T., & Hung, C. L. (2012, June). Development of evolutionary data mining algorithms and their applications to cardiac disease diagnosis. InEvolutionary Computation (CEC), 2012 IEEE Congress on (pp. 1-8). IEEE.

[14] Yadav, G., Kumar, Y., & Sahoo, G. (2012, November). Predication of Parkinson's disease using data mining methods: A comparative analysis of tree, statistical and support vector machine classifiers. In Computing and Communication Systems (NCCCS), 2012 National Conference on (pp. 1-8). IEEE.

[15] Ilayaraja, M., & Meyyappan, T. (2013, February). Mining medical data to identify frequent diseases using Apriori algorithm. In Pattern Recognition, Informatics and Mobile Engineering (PRIME), 2013 International Conference on(pp. 194-199). IEEE.