Proceedings of the International Conference on Inventive Computing and Informatics (ICICI 2017)
IEEE Xplore Compliant - Part Number: CFP17L34-ART, ISBN: 978-1-5386-4031-9

Predict Chronic Kidney Disease Using Data Mining Algorithms In Hadoop

Guneet Kaur , M.Tech
Student, Department of Computer Science and Engineering
Amritsar College of Engineering and Technology
Manawala , Amritsar
gunuvirdi@gmail.com

Er.Ajay Sharma
A.P. of CSE dept.
Amritsar College of Engineering and Technology
Manawala, Amritsar
ajaysharma@acetamritsar.org

Abstract

This paper presents the prediction of chronic kidney disease using data mining classifiers. Now a days, chronic diseases are escalating day by day and play a paramount role in an individual's life. To elicitate the hidden information about chronic disease from a given dataset, data mining technology is used to make decisions. Big data is another area of research used for the storage and processing of voluminous data which is structured, unstructured and semi-structured. In this paper, to predict the chronic kidney disease, two data mining classifiers are used. KNN (K-Nearest Neighbor) and SVM (Support Vector Machine). These approaches provide the following information:

i. **Accuracy**

ii. **Error %**

iii. **Elapsed Time**

The MATLAB tool is utilized for performing prediction of chronic kidney disease by accessing Hadoop in itself.

Keywords: Chronic Kidney Disease, Data Mining, Hadoop, Big Data, KNN and SVM

Nomenclature:

CKD	Chronic Kidney Disease
K-NN	K-Nearest Neighbor
SVM	Support Vector Machine
MATLAB	Matrix Laboratory
ANN	Artificial Neural Network

I. INTRODUCTION

In the 21st century, many type of chronic diseases are booming in an individual's life due to the wide changes taken place by means of their hereditarily issues and living styles. In the field of medical sciences, not much work has done on the chronic kidney diseases but as there is always a hope for the better results so there are many other ways also to detect the chronic kidney diseases and therefore, same is in the case of chronic kidney disease . Data mining has now become a current area of research

and considered as the current trend in the latest technologies all over the world. Data mining is to manage the voluminous amount of data and to create a relevant knowledge for that particular data. Data needs to be sequenced in a mannered order and is used in many real- time applications like various social-networking site, online websites and many more. Furthermore, data mining is categorized into number of domains such as graphical data mining, web data mining, text data mining, image data mining domain. All of the data mining domains are utilized in a way to make decisions and to extract some meaningful hidden data from the given dataset. Data mining extracts some data and acts as a central part of Big Data. Similarly , Big Data is the current growing technology that itself deals with structured, unstructured or semi-structured data.

The most significant application of data mining in the field of medical science is medical data analysis. In an organization of healthcare, data mining is used to detect, to predict , to diagnose and to find out the hidden figures in the medical data. Data mining is very useful in healthcare industry in a way to maintain the complex data . Data mining provides a better framework to emerge the different directions in the healthcare data organization.

In the present era, Big Data is one of the developing technology. Big Data plays a vital role in the medical data sciences.

Chronic Kidney Disease is also referred as a Chronic Renal Failure and now globally, it becomes an area of concern. Chronic Kidney Disease is a progressive loss in function of kidneys over a several years of times. Chronic Kidney Disease is a kind of situation in which kidneys get damaged and toxics cannot easily filtered out from our body. An individual realized the chronic kidney disease when kidneys function lower down 25% of the normal kidney disease is considered as one of the major global issue concerned with the person's health.

The existing work showed that the Naive Bayes and KNN data mining classifiers ignored in order to maintain the accuracy. Many of the researchers ignored these mining classifiers to achieve the best or an accurate results. Even there is no usage of any description method, for the feature extraction process. The existing model did not utilized any kind of the dimensionality reduction algorithm such as PCA(Principal Component Analysis) for attaining the elapsed time period to detect the chronic kidney disease.

In our proposed work, the data mining classifiers KNN and SVM applied in MATLAB by accessing hadoop in itself to predict the chronic kidney disease. Both of the data mining algorithm/techniques deals with an accuracy and error rate along with the time taken by a particular dataset. Similarly, on the another side, SVM is used for evaluating the features whereby for pre-analysis supervised learning is used to obtain the objective MATLAB and Hadoop is used.

TECHNOLOGY USED

In our proposed work , MATLAB and Hadoop are the two technologies which are used to get the accurate results depending upon the number of parameters input provided to the given data set. MATLAB itself attains hadoop accessibility for obtaining the objective of predicting the chronic kidney disease.

1. MATLAB:

Cleve Moler, the chairman of the computer science department at the University of New Mexico, started developing MATLAB in the late 1970's. Firstly, MATLAB was developed by researchers and then practitioners and now used in the field of an education. It provides variety of numerical computations in research area.

2. Hadoop:

Hadoop is the technology considered as an open-source platform build up by Apache. Hadoop is written in Java . Doug Cutting, named Hadoop as a technology by his son's yellow toy elephant . Hadoop provides storage and distributed process of cluster computations.

MATLAB has now used in Big Data for diagnosis, dialysis and analysis of chronic kidney disease datasets. In order to predict the disease there are many modules of MATLAB which is used within data mining classification algorithms. Prediction begins with the identification of symptoms and dataset classified for that in a way to detect the chronic kidney disease.

II. LITERATURE SURVEY

Many of the researchers worked on the detection and prediction of chronic kidney disease by making use

of variety of data mining approaches. But none of them yet used all the mining approaches or techniques along with the big data to predict the chronic kidney disease.

A.Sai Sabitha et al. (2016) used data mining classifiers named as ANN (Artificial Neural Network) and Naive Bayes. They used these data mining classifiers for the prediction and diagnosis of chronic kidney disease. In this research work, Rapid miner tool is used. Results obtained as Naive Bayes classifier possessed 100% accurate results whereas Artificial Neural Network has 72.73% accuracy.

M.Archana Bakare et al.(2016) applied Multirank algorithm to predict the disease of asthma by using various mining approaches for a specific data. In the experimental results, 80% accuracy has observed in the experimental results.

Prasan Kumar Sahoo et al. (2016) observed that 98% accuracy has shown in the current status of patients healthcare. They predicted health disease as in the form of asthma and cancer. In the proposed work, they build a cloud environment in their paper," Analyzing healthcare big data with the prediction for future health condition".

Lambodar Jena et al.(2015) diagnosed the chronic kidney disease using data mining techniques such as SVM(Support Vector Machine) , J48, Naive Bayes, Multi Layer perceptron , conjunctive Roole and decision table. The tool named as WEKA is used and they analyzed that the multiple perceptron has more accurate results as compared to others in order to observe any disease n human.

Anu Chadhary et al. (2014) used Apriori and K-means algorithm with 42 attributes insisted in the data set. They predicted the kidney failure disease and heart disease. In their experimental setup, machine learning tools such as attribute and distribution statistics are implemented to analyze the data.

Dr.S.Vijayarani et al. (2014) used classification techniques such as Naive Bayes and SVM (Support Vector Machine) to evaluate the prediction of heart disease. Their results showed that SVM performed better as compared to the Naive Bayes.

Tommaso Di Noia et al. (2014) predicted the end stage of kidney disease known as ESKD by applying the classifier named as Artificial Neural Network (ANN) which check the end stage probability and in turn developed a software tool for its prediction. At University of Bari, this research explored the ten networks in a 38 years. Furthermore, it can also be used as an application of android mobile phones and on the another side, we can also be utilized as an online web application. This research proved very significant for the clinical usage in healthcare organization.

Solanki A.V. et al. (2014) used WEKA data mining tool in a way to predict the sickle cell disease. In their research, they classified the cell data in a way of numerical computations.

David et al.(2013) predicted the Leukemia disease. In their experimental result, accuracy has maintained by comparing the results of the KNN, J48, Bayesian network, Random tree algorithms of data mining.

Durairaj and Ranjani (2013) compared data mining classification algorithm for predicting the disease like cancer disease and heart disease, AIDS, Diabetes, Kidney dialysis, Brain cancer, IVF and Dengue. For their prediction, J48, Naive Bayes and KNN data mining approaches are used. In the prediction of cancer, 97.77% accuracy is obtained and 70% accuracy is obtained in the IVF.

Lakshmi K.R. et al.(2013) used data mining tool named as Tanagra with the data mining techniques such as AA, Logistics reasoning, decision tree, supervised machine learning algorithms for the dialysis of kidneys. 10 fold cross validations are used in order to classify the data. In their results, they observed that ANN performed better than decision tree and Logical regression algorithm
.

Vijayarani.S.et al. (2013) predicted heart disease by building a cloud network environment. In this research, artificial neural network mining classifier is used and results produced in the existence of cloud environment.

Giovanni Caocci et al. (2012) did comparison by comparing the sensitivity and specificity of logistic regression whereas ANN is used for the prediction of the kidney rejection. In their experimental results, they discriminated the ANN and Logistic regression for the prediction of long term kidney transplantation.

Xun.L. et al. (2010) predicted kidney disease by using two data mining classifiers such as Artificial Neural Network (ANN) and Naive Bayes. In their experimental results, ANN produced more accurate results as compared to the Naive Bayes.

Name of the author & Year	Disease	Techniques/Methodology
A.Sai Sabitha et al. (2016)	Kidney disease	Naive Bayes and ANN
M.Archana Bakare et al.(2016)	Asthma disease	Multirank
Prasan Kumar Sahoo et al. (2016)	Multiple diseases	Map Reduce(Big Data)
Lambodar Jena et al.(2015)	Kidney disease	WEKA data mining tool
Anu Chadhary et al.(2014)	Heart disease and Kidney disease	Apriori and K-means clustering
Dr.Vijayarani et al.(2014)	Kidney Disease	ANN and SVM
Tommaso Di Noia et al.(2014)	Kidney Disease	ANN
Solanki A.V. et al.(2014)	Sickle cell disease	WEKA data mining tool
David et al.(2013)	Leukemia disease	KNN, J48, Bayesian network and Random tree
Durairaj and Ranjani(2013)	Multiple disease	J48, KNN, Naive Bayes, and C4.5
Lakshmi K.R. et al.(2013)	Kidney disease	ANN, Decision tree and Logistic regression
Vijayarani .S. et al.(2013)	Heart disease	ANN
Giovanni Caocci et al.(2012)	Kidney disease	ANN AND Logistic regression
Xun.L.et al.(2010)	Kidney disease	Naive Bayes and ANN

Table 1. Data mining techniques for predicting multiple diseases

III. EXPERIMENTAL DESIGN

In this paper, the work has been carried upon the healthcare prediction using the k-Nearest Neighbor (KNN) and support vector machine (SVM) classification models. The predictive analysis performed in this paper is based upon the manually selected data columns, which includes age, blood pressure, RBC count and appetite fields. These four columns includes the numerical data in the case of blood pressure and age, whereas categorical data in the case RBG count and Appetite. The categorical data has been transformed into the numerical

categories in order to create the compatibility for the classification algorithms, which are the Mathematical models and unable to process the string based categorical data. The following workflow has been utilized for the classification of the data using KNN or SVM models.

Algorithm 1: Healthcare Predictive Analysis using Statistical Models (Cross-validation testing)

1. Acquire the data from the local disk
2. Select the columns manually with the column identifier IDs
3. Convert the string based categorical data to numerical category representation symbols
4. Prepare the final data matrix after the categorical conversions
5. Find the missing values in the final data matrix
6. Compute the average value of each of the column representing the variable
7. Fill the missing values with the mean values with the corresponding average value
8. Shuffle the data matrix to create non-uniform feature matrix
9. Divide the training and testing data matrices
10. Prepare the observation vectors for training and testing
11. Apply SVM over the data for predictive analysis
    a. Record the time → start Time
    b. Train the classifier with the training data matrix and training observations vector
    c. Test the classifier with the test data matrix
    d. Return the predictions (observations predicted by SVM classifier)
    e. Compute the overall performance by comparing SVM predictions and actual observations
    f. Record the time → finish Time
    g. Compute the computational time cTime ← FinishTime-startTime
12. Apply KNN over the data for predictive analysis
    a. Record the time → start Time
    b. Initialize the value of k, which represents the number of neighbors
    c. Train the classifier with the k, training data matrix and training observations vector
    d. Test the classifier with the test data matrix
    e. Return the predictions (observations predicted by KNN classifier)
    f. Compute the overall performance by comparing KNN predictions and actual observations
    g. Record the time finish Time
    h. Compute the computational time cTime ← FinishTime-startTime
13. Combine the results of both of the classifiers
14. Return the performance indication results to the user

IV. RESULTS

The obtained experimental results are represented as two data mining classifier such as KNN and SVM. On the basis of the given dataset, we did feature extraction and feature evaluation. In this section, all the outcomes are discussed. In our proposed work, we calculate the performance measures of the data mining classifiers used in this research based on three parameters which are as follows:

1. Accuracy
2. Error %
3. Elapsed Time

All the parameters are compared in the following tables where each parameter specified in a specific table for both the classification algorithms , where KNN and SVM are generated by using MATLAB itself.

1. Total Accuracy :

It is observed that SVM classifier gives more accuracy i.e. 78.09% comparing to KNN classifier which gives 70% accuracy whereas SVM always possessed less accurate results in most of the works done by many researchers. Hence it is concluded that SVM perform well in given chronic kidney dataset. The following table has shown the accuracy comparison between KNN and SVM :

KNN Accuracy(in %)	SVM accuracy (in%)
69	78
71	69
71	82
74	85
67	83
71	64
69	85
70	79
71	74
71	82
66	78
TOTAL = 70%	**TOTAL = 78.09%**

Table 2. Total Accuracy of KNN and SVM

2. Total Error% :

It is observed that SVM classifier possessed less error i.e. 21.90% comparing to KNN data mining classifier which gives 30% Therefore, it is concluded that SVM perform well again in case of error performance measures in given chronic kidney dataset. The following table has shown the comparison of error between KNN and SVM:

KNN error %(in%)	SVM error %(in%)
31	22
29	31
29	18
26	15
33	17
29	36
31	15
30	21
29	26
29	18
34	22
TOTAL = 30 %	**TOTAL = 21.90 %**

Table 3. Total Error % of KNN and SVM

3. Total Elapsed Time:

It is monitored that SVM data mining algorithm takes more time for its execution i.e. 37% when comparing to KNN data mining technique which takes 13.27% execution time for its completion. Hence it is observed that SVM possessed more time as compared to KNN in the given chronic kidney dataset. The following table has shown the another performance measure comparison as in the form of total time taken between KNN and SVM:

KNN elapsed Time (in %)	SVM elapsed Time (in %)
0.14	0.30
0.14	0.38
0.13	0.37
0.13	0.36
0.14	0.38
0.14	0.35
0.13	0.38
0.12	0.38
0.12	0.39
0.14	0.39
0.13	0.40
TOTAL = 13.27%	**TOTAL = 37%**

Table 4. Elapsed Time of KNN and SVM

REFERENCES

[1] A.Sai Sabitha, Abhay Bansal, Khushboo Chandel, Veenita Kunwar, Chronic Kidney Disease Analysis

Using Data Mining Classification Techniques, 6th International Conference on Cloud System and Big Data Engineering, 2016.

[2] Agarwal.Y, Pandey, H.M, Performance evaluation of different techniques in the context of data mining-A case of an eye disease. In Confluence the Next Generation Information Technology

Summit (Confluence), 5th International Conference-IEEE, 2014.

[3] Ahmed. S, Tanzir Kabir, M.Tanzeer Mahmood, N. Rahman, R.M., Diagnosis of kidney disease using fuzzy expert system. In Software, Knowledge Information Management and Applications (SKIMA),

8th International Conference- IEEE, 2014.

[4] Alfisahrin, S.D.N.N, Mantoro, T, Data Mining Techniques for Optimization of Liver Disease Classification. In Advanced Computer Science Applications and Technologies (ACSAT),

International Conference- IEEE, 2013.

[5] Amin, S.U, Agarwal, K.Beg, R, Genetic Neural Network based data mining in prediction of heath disease using risk factors. In Information and Communication Technologies (ICT), IEEE Conference, 2013.

[6] Andrew Kusiak, Bradley Dixonb, Shital Shaha, and Predicting survival time for kidney dialysis patients: a data mining approach, Elsevier Publication, Computers in Biology and Medicine 35, page no 311-327, 2005.

[7] Anu Chaudhary, Puneet Garg, Detecting and Diagnose Heart and Kidney disease by Patient Monitoring System, International Journal of

Mechanical Engineering and Information Technology, Vol.2 Issue 6//June//Page No: 493-499, 2014.

[8] David S.K., Saeb A.T., Al Rubeaan K., Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in medical Bioinformatics, Computer Engineering and Intelligent Systems, 4(13):28-38, 2013.

[9] Dhamodharan S, Liver Disease Prediction Using

Bayesian Classification, Special Issues, 4th National Conference on Advance Computing, Application Technologies, May 2014.

[10] Vijayarani, S., & Dhayananad, M., S. Kidney Disease Prediction using SVM and ANN algorithms, unpublished.

[11] Duraiaj M, Ranjani V, Data Mining applications in healthcare sector a study. Int. J. Sci. Technol. Res. IJSTR, 2(10), 2013.

[12] Giovanni Caocci, Roberto Baccoli, Roberto Littera, Sandro Orru, Carlo Carcassi and Giorgio La Nasa, Comparison Between an ANN and Logistic Regression in Predicting Long Term Kidney Transplantation Outcome, Chapter 5, an open access article distributed under the terms of the Creative Commons Attribution License, http://dx.doi.org/10.5772/53104,2012.

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

[14] Joshi J, Rinal D, Patel J, Diagnosis and Prognosis of Breast Cancer using Classification Rules, International Journal of Engineering Research and General Science, 2(6): 315-323, October 2014.

[15] Kumar M.N., Alternating Decision trees for early diagnosis of dengue fever .arXiv preprint arXiv: 1305.7331, 2013.

[16] Lakshmi, K.R., Nagesh, Y., & VeeraKrishna, M, Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability, International Journal of Advances in Engineering & Technology, Mar., Vol.7, Issue 1, Page No: 242-254, 2013.

[17] Lambodar Jena, Narendra Ku. Kamila, Distributed Data Mining Classification Algorithms for Prediction of Chronic Kidney Disease, International Journal of Emerging Research in Management & Technology ISSN: 2278-9359(Vol 4, Issue-11), IJERMT, 2015

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

[19] M Archana Bakare, Prof. R.V. Argiddi, Prediction of Diseases using Big Data Analysis, International Journal of Innovative Research in Computer and Communication Engineering(Vol.4, Issue 4), ISSN, April 2016.

[20] M.A. Nishara Banu, Gomathy B., Heart Disease Prediction using Data mining Techniques. In

Intelligent Computing Applications (ICICA), International Conference on (pp.130-133), IEEE, March 2014.

[21] M.Cottle, W.Hoover, S. Kanwal, M. Kohn, T. Strome and N. W. Treister, Transforming Health Care Through Big Data, 2013.

[22] Maskery, S., Zhang, Y., Hu, H., Shriver, C., Hooke, J., & Liebman, M. (2006, June), Caffeine intake, race, and risk of invasive breast cancer lessons learned from data mining a clinical database.

In Computer- Based Medical Systems, CBMS, 19th IEEE International Symposium on (pp. 714-718), IEEE, 2006.

[23] Prasan Kumar Sahoo, Suvendu Kumar Mohapatra, Shih-Lin Wu, Analyzing Healthcare Big Data with Prediction for Future Health Condition, Journal Publications , IEEE, 2016.

[24] Shakil K.A. and Alam M., Data Management in Cloud Based Environment using K-median Clustering Technique, IJCA Proceedings on 4th International IT Summit Confluence - The Next Generation Information Technology Summit Confluence , 2013.

[25] Vijayarani, S., Sudha, S., Comparative Analysis of Classification Function Techniques for Heart Disease Prediction, International Journal of

Innovative Research in Computer and
Communication Engineering, 1(3): 735-741, 2013.

[26] Watanasusin, N., & Sanguansintukul, S., Classifying chief complaint in ear diseases using data mining techniques. In Digital Content, Multimedia

Technology and its Applications (IDCTA), 7th International Conference on (pp. 149-153), IEEE, August 2011.

[27] Xiong, X., Kim, Y., Baek, Y., Rhee, D. W., & Kim, S. H., Analysis of breast cancer using data mining & statistical techniques. In Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, First ACIS International Workshop on Self-Assembling Wireless Networks. 6th International Conference on (pp. 82-87), IEEE, May 2005.

[28] Xun, L., Xiaoping, W., Ningshan, L., & Tanqi, L., Application of radial basis function neural network to estimate glomerular filtration rate in Chinese patients with chronic kidney disease. In Computer Application and System Modeling (ICCASM), International Conference on (Vol. 15, pp. V15-332), IEEE, October 2010.