Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 7, July 2014, pg.960 – 967

REVIEW ARTICLE

A Literature Review on Kidney Disease Prediction using Data Mining Classification Technique

Suman Bala[1], Krishan Kumar[2]

[1]Department of Computer Science& Engineering, JNTU Hyderabad, India

[2]Department of Computer Science& Engineering, KUK Kurukshetra, India

[1] er.sumanverma.01@gmail.com, [2] v_krishan@rediffmail.com

Abstract--- The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making. The Healthcare industry is generally "information rich", which is not feasible to handle manually. These large amounts of data are very important in the field of data mining to extract useful information and generate relationships amongst the attributes. Kidney disease is a complex task which requires much experience and knowledge. Kidney disease is a silent killer in developed countries and one of the main contributors to disease burden in developing countries. In the health care industry the data mining is mainly used for predicting the diseases from the datasets. The Data mining classification techniques, namely Decision trees, ANN, Naive Bayes are analyzed on Kidney disease data set.

Keywords--- Data Mining, Kidney Disease, Decision tree, Naive Bayes, ANN, K-NN, SVM, Rough Set, Logistic Regression, Genetic Algorithms (GAs) / Evolutionary Programming (EP), Clustering

I. INTRODUCTION

Data mining is the non-trivial extraction of implicit previously unknown and potentially useful information about data [4]. Data Mining is one of the most vital and motivating area of research with the objective of finding meaningful information from huge data sets. In present era, Data Mining is becoming popular in healthcare field because there is a need of efficient analytical methodology for detecting unknown and valuable information in health data [1]. Medical data mining is used in the knowledge acquisition and analyses the information obtained from research reports, medical reports, flow charts, evidence tables, and transform these mounds of data into useful information for decision making[2]. This paper demonstrated the utility of Classification techniques for predicting Kidney disease with different data mining tools.

A. Organization of the paper: Section II Kidney diseases factors, symptoms and type of kidney disease. Section III describes literature reviews. Section IV describes data mining Classification techniques. Experimental results are presented in Section V and finally, Section VI concludes the paper and points out some potential future work.

II. KIDNEY DISEASE

The kidneys' functions are to filter the blood. All the blood in our bodies passes through the kidneys several times a day. The kidneys remove wastes, control the body's fluid balance, and regulate the balance of electrolytes. As the kidneys filter blood, they create urine, which collects in the kidneys' pelvis -- funnel-shaped structures that drain down tubes called ureters to the bladder. Each kidney contains around a million units called nephrons, each of which is a microscopic filter for blood. It's possible to lose as much as 90% of kidney function without experiencing any symptoms or problems. Kidney disease is a silent killer [5].

There are number of factors which increase the risk of Kidney disease:

- Diabetes
- Hypertension
- Smoking
- Obesity
- Heart disease
- Family history of Kidney disease
- Alcohol intake
- Drug abuse/drug overdose
- Age
- Race/Ethnicity
- Male sex

Symptoms of kidney disease:

- Changes in your urinary function
- Difficulty or pain during voiding
- Blood in the urine
- Swelling & Pain in the back or sides
- Extreme fatigue and generalized weakness
- Dizziness & Inability to concentrate
- Feeling cold all the time
- Skin rashes and itching
- Ammonia breath and metallic taste
- Nausea and vomiting
- Shortness of breath

Types of Kidney diseases:

- Pyelonephritis (infection of kidney pelvis): Bacteria may infect the kidney, usually causing back pain and fever. A spread of bacteria from an untreated bladder infection is the most common cause of pyelonephritis.

- Glomerulonephritis: An overactive immune system may attack the kidney, causing inflammation and some damage. Blood and protein in the urine are common problems that occur with glomerulonephritis. It can also result in kidney failure.

- Kidney stones (nephrolithiasis): Minerals in urine form crystals (stones), which may grow large enough to block urine flow. It's considered one of the most painful conditions. Most kidney stones pass on their own but some are too large and need to be treated.

- Nephrotic syndrome: Damage to the kidneys causes them to spill large amounts of protein into the urine. Leg swelling (edema) may be a symptom.

- ❖ Polycystic kidney disease: A genetic condition resulting in large cysts in both kidneys that impair their function.
- ❖ Acute renal failure (kidney failure): A sudden worsening in kidney function. Dehydration, a blockage in the urinary tract, or kidney damage can cause acute renal failure, which may be reversible.
- ❖ Chronic renal failure: A permanent partial loss of kidney function. Diabetes and high blood pressure are the most common causes.
- ❖ End stage renal disease (ESRD): Complete loss of kidney function, usually due to progressive chronic kidney disease. People with ESRD require regular dialysis for survival.
- ❖ Diabetic nephropathy: High blood sugar from diabetes progressively damages the kidneys, eventually causing chronic kidney disease. Protein in the urine (nephrotic syndrome) may also result.
- ❖ Hypertensive nephropathy: Kidney damage caused by high blood pressure. Chronic renal failure may eventually result.
- ❖ Kidney cancer: Renal cell carcinoma is the most common cancer affecting the kidney. Smoking is the most common cause of kidney cancer.
- ❖ Interstitial nephritis: Inflammation of the connective tissue inside the kidney, often causing acute renal failure. Allergic reactions and drug side effects are the usual causes.
- ❖ Minimal change disease: A form of nephrotic syndrome in which kidney cells look almost normal under the microscope. The disease can cause significant leg swelling (edema). Steroids are used to treat minimal change disease.
- ❖ Nephrogenic diabetes insipidus: The kidneys lose the ability to concentrate the urine, usually due to a drug reaction. Although it's rarely dangerous, diabetes insipidus causes constant thirst and frequent urination.
- ❖ Renal cyst: A benign hollowed-out space in the kidney. Isolated kidney cysts occur in many normal people and almost never impair kidney function.

III.LITERATURE REVIEW

This section consists of the reviews of various technical and review articles on data mining techniques applied to predict Kidney Disease.

- DSVGK Kaladhar, Krishna Apparao Rayavarapu and Varahalarao Vadlapudi et al [6]. described in their research to understand machine learning techniques to predict kidney stones. They predicted good accuracy with C4.5, Classification tree and Random forest (93%) followed by Support Vector Machines (SVM) (91.98%). Logistic and NN has also shown good accuracy results with zero relative absolute error and 100% correctly classified results. ROC and Calibration curves using Naive Bayes has also been constructed for predicting accuracy of the data. Machine learning approaches provide better results in the treatment of kidney stones.

- J.Van Eyck, J.Ramon, F.Guiza, G.Meyfroidt, M.Bruynooghe, G.Van den Berghe, K.U.Leuven et al [7]. Explored data mining techniques for predicting acute kidney injury after elective cardiac surgery with Gaussian process & machine learning techniques (classification task & regression task).

- K.R.Lakshmi, Y.Nagesh and M.VeeraKrishna et al [8]. presented performance comparison of Artificial Neural Networks, Decision Tree and Logical Regression are used for Kidney dialysis survivability. The data mining techniques were evaluated based on the accuracy measures such as classification accuracy, sensitivity and specificity. They achieved results using 10 fold cross-validations and confusion matrix for each technique. They found ANN shows better results. Hence ANN shows the concrete results with Kidney dialysis of patient records.

- Morteza Khavanin Zadeh, Mohammad Rezapour, and Mohammad Mehdi Sepehri et al [9]l. described in their research by using supervised techniques to predict the early risk of AVF failure in patients.

They used classification approaches to predict probability of complication in new hemodialysis patients whom have been referred by nephrologists to AVF surgery.

- Abeer Y. Al-Hyari et al [10].proposed in their research by using Artificial Neural Network (NN), Decision Tree (DT) and Naïve Bayes (NB) to predict chronic kidney disease. The proposed NN algorithm as well as the other data mining algorithms demonstrated high potential in successful kidney disease.

- Xudong Song, Zhanzhi Qiu, Jianwei Mu et al [11].introduced data mining decision tree classification method, and proposed a new variable precision rough set decision tree classification algorithm based on weighted limit number explicit region.

- N. SRIRAAM, V. NATASHA and H. KAUR et al [12].presented data mining approach for parametric evaluation to improve the treatment of kidney dialysis patient. Their experimental result shows that classification accuracy using Association mining between the ranges 50–97.7% is obtained based on the dialysis parameter combination. Such a decision-based approach helps the clinician to decide the level of dialysis required for individual patient.

- Jicksy Susan Jose, R.Sivakami, N. Uma Maheswari, R.Venkatesh et al [13]. Their research describes an efficient Diagnosis of Kidney Images Using Association Rules. Their approach is divided into four major steps: pre-processing, feature extraction and selection, association rule generation, and generation of diagnosis suggestions from classifier.

- Divya Jain et al [14].presented effect of diabetes on kidney using C4.5 algorithm with Tanagra tool. The performance of classifier is evaluated in terms of recall, precision and error rate.

- Koushal Kumar and Abhishek et al [15].their research describes comparison of all three neural networks such as (MLP, LVQ, RBF) on the basis of its accuracy, time taken to build model, and training data set size.

IV. DATA MINING TECHNIQUES USED FOR PREDICTIONS

Classification is a very important data mining task, and the purpose of classification is to propose a classification function or classification model (called classifier).The classification model can map the data in the database to a specific class. Classification construction methods include: Decision Tree, Naive Bayes, ANN, K-NN, Support Vector Machine, Rough set, Logistic Regression, Genetic Algorithms (GAs) / Evolutionary Programming (EP), Clustering etc [3].

Decision Tree: The decision tree is a structure that includes root node, branch and leaf node. Each internal node denotes a test on attribute, each branch denotes the outcome of test and each leaf node holds the class label. The topmost node in the tree is the root node. The decision tree approach is more powerful for classification problems. There are two steps in this techniques building a tree & applying the tree to the dataset. There are many popular decision tree algorithms CART, ID3, C4.5, CHAID, and J48.

Artificial Neural Network (ANN): is a collection of neuron –like processing units with weight connections between the units. It maps a set of input data onto a set of appropriate output data. It consists of 3 layers: input layer, hidden layer & output layer. There is connection between each layer & weights are assigned to each connection. The primary function of neurons of input layer is to divide input xi into neurons in hidden layer. Neuron of hidden layer adds input signal xi with weights wji of respective connections from input layer. The output Yj is function of $Yj = f(\Sigma\ wji\ xi)$ Where f is a simple threshold function such as sigmoid or hyperbolic tangent function.

Naive Bayes: Naive Bayes classifier is based on Bayes theorem. This classifier algorithm uses conditional independence, means it assumes that an attribute value on a given class is independent of the values of other attributes. The Bayes theorem is as follows: Let X={x1, x2... xn} be a set of n attributes. In Bayesian, X is considered as evidence and H is some hypothesis means, the data of X belongs to specific class C. We have to

determine P (H|X), the probability that the hypothesis H holds given evidence i.e. data sample X. According to Bayes theorem the P (H|X) is expressed as P (H|X) = P (X| H) P (H) / P (X).

K-Nearest Neighbour: The *k*-nearest neighbour's algorithm (*K*-NN) is a method for classifying objects based on closest training data in the feature space. K-NN is a type of instance-based learning. The *k*-nearest neighbour algorithm is amongst the simplest of all machine learning algorithms. But the accuracy of the *k*-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance.

Logistic Regression: The term regression can be defined as the measuring and analyzing the relation between one or more independent variable and dependent variable. Regression can be defined by two categories; they are linear regression and logistic regression. Logistic regression is a generalized by linear regression. It is mainly used for estimating binary or multi-class dependent variables and the response variable is discrete, it cannot be modelled directly by linear regression i.e. discrete variable changed into continuous value. Logistic regression basically is used to classify the low dimensional data having non-linear boundaries. It also provides the difference in the percentage of dependent variable and provides the rank of individual variable according to its importance. So, the main motto of Logistic regression is to determine the result of each variable correctly.

Rough Sets: A Rough Set is determined by a lower and upper bound of a set. Every member of the lower bound is a certain member of the set. Every non-member of the upper bound is a certain non-member of the set. The upper bound of a rough set is the union between the lower bound and the so-called boundary region. A member of the boundary region is possibly (but not certainly) a member of the set. Therefore, rough sets may be viewed as with a three-valued membership function (yes, no, perhaps). Rough sets are a mathematical concept dealing with Uncertainty in data. They are usually combined with other methods such as rule induction or clustering methods.

Support Vector Machine (SVM): Support vector machine (SVM) is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. SVMs are well suited to dealing with interactions among features and redundant features.

Genetic Algorithms (GAs) / Evolutionary Programming (EP): Genetic algorithms and evolutionary programming are used in data mining to formulate hypotheses about dependencies between variables, in the form of association rules or some other internal formalism.

Support Vector Machine (SVM): Support vector machine (SVM) is an algorithm that attempts to find a linear separator (hyper-plane) between the data points of two classes in multidimensional space. SVMs are well suited to dealing with interactions among features and redundant features.

Clustering: Clustering is the process of grouping similar elements. This technique may be used as a pre-processing step before feeding the data to the classifying model. The attribute values need to be normalized before clustering to avoid high value attributes dominating the low value attributes. Further, classification is performed based on clustering.

V. EXPERIMENTAL RESULTS AND ANALYSIS

This section presents the experimental results and analysis done for this study. The data mining techniques used for kidney disease prediction has been explained in section IV. For the experiments, various different classification techniques have been applied to predict kidney disease. Table 1 shows the results for classification techniques to predict Kidney disease with different mining tools for this work.

Table1. Results of classification techniques used for kidney disease:

Author	Publication Year	Type of Kidney Disease	Tool	Techniques		Accuracy
Kaladhar et al.[6]	2012	Kidney stone	WEKA	Naive Bayes		0.99%
				Logistic		1.00%
				J48		0.97%
				Random Forest		0.98%
			ORANGE	Naive Bayes		0.79%
				K-NN		0.7377%
				Classification tree		0.9352%
				C4.5		0.9352%
				SVM		0.9198%
				Random Forest		0.9352%
K.R.Lakshmi et al.[8]	2014	Kidney dialysis	TANAGRA	ANN		93.852%
				Decision Tree(C5)		78.4455%
				Logical Regression		74.7438%
J.Van Eyck et al[7]	2012	AKI	MATLAB	Gaussian process(aROC)		0.758%
				Gussian process(RMSER)		0.408%
Morteza Khavanin Zadeh et al.[9]	2012	Early AVF Failure	WEKA	W-Simple Cart		85.11%
				WJ48		80.85%
Abeer Y. Al-Hyari et al[10]	2012	Chronic Kidney disease	WEKA	Decision tree		--
Xudong Song et al[11]	2012	Renal failure Hemodialysis	WEKA	Decision tree		60-80%
N. SRIRAAM et al[12]	2005	Kidney Dialysis	-	Association Rule		97.7%
Divya Jain et al [14]	2014	Nephrotic syndrome(total protein)	TANAGRA	C4.5		11% (error rate)
Jicksy Susan Jose et al[13]	2012	Kidney Image	MATLAB	Association Rule		92%
				Navie Bayes		
Koushal Kumar et al[15]	2012	Kidney Stone	WEKA	ANN	MLP	0.9613%
					LVQ	0.8459%
					RBF	0.8732

VI. CONCLUSIONS

The overall objective is to study the various data mining techniques available to predict the Kidney disease and to compare them to find the best method of prediction.

- We analyzed in order to improve the rough set model; the variable precision rough set model was suggested by the introduction of the error parameterβ. It allows some errors in the division process, which perfected the approximation space concept, reduced the tree's branches substantially, and improved generalization capabilities. However, the variable precision rough set decision tree construction process still possesses an obvious shortcoming in the process of calculating the explicit region: the more the number of attributes is the greater the value of explicit region. In order to solve this problem, they proposed a new weight limit number explicit region calculation method.

- We analyzed that the most commonly used DM technique such as Decision Trees, ANN and Naïve Bayes, Logistic Regression, Genetic Algorithms (GAs) resulting as well-performing on medical databases. Also shows that DTs, ANNs and Naive Bayes are the well-performing algorithms used for Kidney disease. But it is very difficult to name a single DM technique as the best for the Kidney

diseases. Depending on concrete situations, sometime some techniques perform better than others, but there are cases when a combination of the best properties of some of the aforementioned DM techniques results more effective.

- We also analyzed that there is no single classifier which produce best result for every dataset. The performance of a classifier is evaluated using testing data set. But there are also problem with testing data set. Some time it is complex and some time it becomes easy to classify the testing data set. To avoid these problems they used cross validation method so that every record of data set is used for both training and testing.

- We also analyzed that by using different data mining classification techniques and tools not only to predict kidney disease but also the accuracy of kidney images and effect of other disease on kidney.

References

[1] H. C. Koh and G. Tan, "*Data Mining Application in Healthcare*", Journal of Healthcare Information Management, vol. 19, no. 2, 2005.

[2] K.Sudhakar & Dr. M. Manimekalai," *Study of Heart Disease Prediction using Data mining*", IJARCSSE, Volume 4, Issue 1, January 2014.

[3] J. Han and M. Kamber, "*Data mining: concepts and techniques*", 2nd Ed. The Morgan Kaufmann Series, 2006.

[4] Frawley and Piatetsky-Shapiro, 1996. *Knowledge Discovery in Databases*: An Overview. The AAAI/MIT Press, Menlo Park, C.A.

[5] http://www.webmd.com/urinary-incontinence-oab.

[6] DSVGK Kaladhar, Krishna Apparao Rayavarapu* and Varahalarao Vadlapudi,"*Statistical and Data Mining Aspects on Kidney Stones: A Systematic Review and Meta-analysis*", Open Access Scientific Reports, Volume 1 • Issue 12 • 2012.

[7] J.Van Eyck, J.Ramon, F.Guiza, G.Meyfroidt, M.Bruynooghe, G.Van den Berghe, K.U.Leuven," *Data mining techniques for predicting acute kidney injury after elective cardiac surgery*", Springer, 2012.

[8] K.R.Lakshmi, Y.Nagesh and M.VeeraKrishna,"*Performance comparison of three data mining techniques for predicting kidney disease survivability*", International Journal of Advances in Engineering & Technology, Mar. 2014.

[9] Morteza Khavanin Zadeh, Mohammad Rezapour, and Mohammad Mehdi Sepehri," Data Mining Performance in Identifying the Risk Factors of Early Arteriovenous Fistula Failure in Hemodialysis Patients", International journal of hospital research, Volume 2, Issue 1,2013, pp 49-54.

[10] Abeer Y. Al-Hyari," *CHRONIC KIDNEY DISEASE PREDICTION SYSTEM USING CLASSIFYING DATA MINING TECHNIQUES*", library of university of Jordan, 2012.

[11] Xudong Song, Zhanzhi Qiu, Jianwei Mu," *Study on Data Mining Technology and its Application for Renal Failure Hemodialysis Medical Field*", International Journal of Advancements in Computing Technology(IJACT) ,Volume4, Number3, February 2012.

[12] N. SRIRAAM, V. NATASHA and H. KAUR," *DATA MINING APPROACHES FOR KIDNEY DIALYSIS TREATMENT*" , journal of Mechanics in Medicine and Biology, Volume 06, Issue 02, June 2006.

[13] Jicksy Susan Jose, R.Sivakami, N. Uma Maheswari, R.Venkatesh," *An Efficient Diagnosis of Kidney Images using Association Rules*", International Journal of Computer Technology and Electronics Engineering (IJCTEE),Volume 2, Issue 2,april 2012.

[14] Divya Jain, Sumanlata Gautam," *Predicting the Effect of Diabetes on Kidney using Classification in Tanagra*", International Journal of Computer Science and Mobile Computing, Volume 3, Issue 4, April 2014.

[15] Koushal Kumar and Abhishek,"*Artificial Neural Networks for Diagnosis of Kidney Stones Disease*", I.J. Information Technology and Computer Science, 2012, 7, pp 20-25.

