Research Article

A Reduced Set of Features for Chronic Kidney Disease Prediction

Rajesh Misir[1], Malay Mitra[2], Ranjit Kumar Samanta[2]

[1]Department of Computer Science, Vidyasagar University, Medinipur, [2]Department of Computer Science and Application, Expert Systems Laboratory, University of North Bengal, Darjeeling, West Bengal, India

Received: 12 November 2016 Accepted: 23 February 2017 Published: 19 June 2017

Abstract

Chronic kidney disease (CKD) is one of the life-threatening diseases. Early detection and proper management are solicited for augmenting survivability. As per the UCI data set, there are 24 attributes for predicting CKD or non-CKD. At least there are 16 attributes need pathological investigations involving more resources, money, time, and uncertainties. The objective of this work is to explore whether we can predict CKD or non-CKD with reasonable accuracy using less number of features. An intelligent system development approach has been used in this study. We attempted one important feature selection technique to discover reduced features that explain the data set much better. Two intelligent binary classification techniques have been adopted for the validity of the reduced feature set. Performances were evaluated in terms of four important classification evaluation parameters. As suggested from our results, we may more concentrate on those reduced features for identifying CKD and thereby reduces uncertainty, saves time, and reduces costs.

Keywords: Chronic kidney disease, correlation, intelligent binary classification, reduced feature set, UCI database

Introduction

Disorders affecting kidney structure and function in heterogeneous form are generally termed chronic kidney disease (CKD).[1] National Kidney Foundation of America in 2002 created guidelines for a more clear definition and classification system for CKD.[2] CKD is classified into Stages I–V according to the estimated glomerular filtration rate (GFR) shown in Table 1.[2] GFR is estimated having mathematical equations using serum creatinine, age, sex, body size, ethnic origin, etc.[1,3] If the normal functionality of a kidney is degraded to an extent, wastes can build up to high levels in your blood making you feel sick. Unfortunately, this degradation is noted at the later stages of CKD making the matter complicated in much of the cases. Sometimes, it is too late when we consult a physician. Hence, the early detection is solicited for long-term survivability. It is argued that once in a year or once in 2 years, CKD-related investigations may be done. General awareness of the people has to be increased along with providing less number of pathological tests with less cost and less time.

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/jpi.jpi_88_16

Table 2 represents the CKD data set from UCI that contains 24 attributes plus one attribute for class (binary).[4] It contains 400 samples to two different classes ("CKD" - 250 cases; "NOTCKD" - 150 cases). Out of 24 attributes, 11 are numeric and 13 are nominal. The data set contains a number of missing values. After excluding tuples with missing values, 158 samples were used in this work.

This paper aims to determine the important features for CKD predictions. Features were selected using one correlation-based algorithm with eight different searching techniques. Two different intelligent classification algorithms based on artificial neural networks were applied to test the effectiveness of the feature set. Classification results were compared in terms of four model-validating parameters.

Address for correspondence: Dr. Ranjit Kumar Samanta, Department of Computer Science and Application, Expert Systems Laboratory, University of North Bengal, Darjeeling - 734 013, West Bengal, India.
E-mail: rksamantark@gmail.com

How to cite this article: Misir R, Mitra M, Samanta RK. A reduced set of features for chronic kidney disease prediction. J Pathol Inform 2017;8:24.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2017/8/1/24/208470

Methods

Correlation-based feature subset selection

Correlation-based feature subset selection (CFS) is based on the following hypothesis:

A feature of a subset is considered good which are highly correlated with the class but may be uncorrelated with other features of the class.[5]

The feature evaluation mathematical formula provides an operational definition of the above hypothesis as follows:

$$r_{fc} = \frac{k\overline{r_{fc}}}{\sqrt{(k + k(k-1)\overline{r_{ff}})}} \quad (1)$$

Where r_{fc} is the correlation between the summed features and the class variable; k is the number of features rfc and above line is the average of the correlation between the features and class variable; and rff and above line is the average intercorrelation between features.[6]

Summarily, a feature will be admitted to the subset if its correlation with the class is higher than the highest correlation between it and anyone of the already selected features.

To accommodate all categories of features in equation 1, continuous features are transformed to categorical features using a discretization method.[7] The degree of associations between nominal features was estimated by a specific method.[8] For searching the feature subset in a reasonable amount of time, heuristic search strategies are often used.[9] Further, different types of feature selection methods use the correlation-based approach in different applications.[10,11]

Eight search methods were applied in the study, namely, Best First, Exhaustive Search, Genetic Search, Greedy Stepwise, Linear Forward Selection, Random Search, Scatter Search, and Subset Size Forward Selection.[12]

Incremental back propagation learning networks classifier

The normal back propagation network is not an incremental by its nature.[13] The network learns by the back propagation rule of Rumelhart *et al*. under the constraint that the change to each weight for each instance is bounded.[14] It is accomplished by introducing scaling factors which scale down all weight adjustments so that all of them are within bounds. The learning rule is now

Table 1: The stages of chronic kidney disease

Stage	Clinical features	GFR (mL/min/1.7 m²)
I	Damage with normal or increased GFR	≥90
II	Damage with a mild decrease in GFR	60-89
III	Moderate decrease in GFR	30-59
IV	Severe decrease in GFR	15-29
V	Kidney failure	<15 or dialysis

GFR: Glomerular filtration rate

Table 2: The attributes of chronic kidney disease of UCI

Attribute number	Attributes (type)	Attribute values	Attribute codes
1	Age (numerical)	Years	Age
2	Blood pressure (numerical)	mm/Hg	bp
3	Specific gravity (nominal)	1.005, 1.010, 1.015, 1.020, 1.025	sg
4	Albumin (nominal)	0, 1, 2, 3, 4, 5	al
5	Sugar (nominal)	0, 1, 2, 3, 4, 5	su
6	Red blood cells (nominal)	Normal, abnormal	rbc
7	Pus cell (nominal)	Normal, abnormal	pc
8	Pus cell clumps (nominal)	Present, not present	pcc
9	Bacteria (nominal)	Present, not present	ba
10	Blood glucose random (numerical)	mg/dl	bgr
11	Blood urea (numerical)	mg/dl	bu
12	Serum creatinine (numerical)	mg/dl	sc
13	Sodium (numerical)	mEq/L	sod
14	Potassium (numerical)	mEq/L	pot
15	Hemoglobin (numerical)	g	hemo
16	Packed cell volume (numerical)	-	pcv
17	White blood cell count (numerical)	cells/cumm	wbcc
18	Red blood cell count (numerical)	millions/cmm	rbcc
19	Hypertension (nominal)	No, yes	htn
20.	Diabetes mellitus (nominal)	No, yes	dm
21	Coronary artery disease (nominal)	No, yes	cad
22	Appetite (nominal)	Good, poor	appet
23	Pedal edema (nominal)	Yes, no	pe
24	Anemia (nominal)	Yes, no	ane
25	Class (nominal)	CKD, NOTCKD	-

CKD: Chronic kidney disease

$$\Delta W_{ij}(k) = s(k)\eta\delta_j(k)O_i(k) \quad (2)$$

Where W_{ij} is the weight from unit i to unit j, η ($0<\eta<1$) is a trial-independent learning rate, δ_j is the error gradient at unit j, O_i is the activation level at unit i, and the parameter k denotes the *k*-th iteration. The incremental back propagation learning networks proceed as follows:

Given a single misclassified instance:

Begin
Repeatedly apply the bounded weight adaptation learning rule (2) on the instance until stopping criteria are met.

If
the instance can be correctly learned, then restore the old weights and apply the bounded weight adaptation learning rule once;

Else
restore the old weights and apply the structural adaptation learning rules.

End
The stopping criteria are the instance can be correctly learned or the output error fluctuates in a small range.[13]

Levenberg–Marquardt classifier

The Levenberg–Marquardt (LM) algorithm is basically an iterative method that locates the minimum of a multivariate function that is expressed as the sum of squares of nonlinear real-valued functions.[15,16] LM can be thought of as a combination of steepest descent and the Gauss–Newton (GN) method. LM algorithm is more robust than GN algorithm which essentially means that it finds a solution even if it starts far off the final minimum. During the iterations, the new configuration of weights in step k + 1 is calculated as follows

$$W(k+1) = w(k) - (J^T J + \lambda I)^{-1} J^T \varepsilon(k) \quad (3)$$

Where J - the Jacobian matrix, λ - adjustable parameter, and ε - error vector. The parameter λ is modified based on the development of error function E. If the step causes a reduction of E, we accept it. Otherwise, λ is changed; reset the original value and recalculate w (k + 1).

Applications

This study consists of three stages: feature extraction and reduction by CFS using eight different search algorithms finding the most relevant and reduced set of features and then classification by two important classification algorithms. The schematic view of our proposed system is shown in Figure 1.

Data Preprocessing

It is one of the important steps for the development of any model. We completely randomize the data sets with missing records. Missing tuples were excluded leaving 158 data sets for use in this work without missing values.

We apply CFS with the aforementioned eight search techniques using a free software named WEKA.[12] Out of the eight search algorithms, six algorithms, a majority, suggested eight common reduced attribute set as shown in Table 3.

Next, we apply the two discussed intelligent techniques on the data set to testify the effectiveness of the above reduced feature set as shown in Table 3.

Table 3: Reduced chronic kidney disease attributes using correlation based feature subset selection

n	Attributes
1	Specific gravity (sg)
2	Albumin (al)
3	Serum creatinine (sc)
4	Hemoglobin (hemo)
5	Packed cell volume (pcv)
6	White blood cell count (wbcc)
7	Red blood cell count (rbcc)
8	Hypertension (htn)

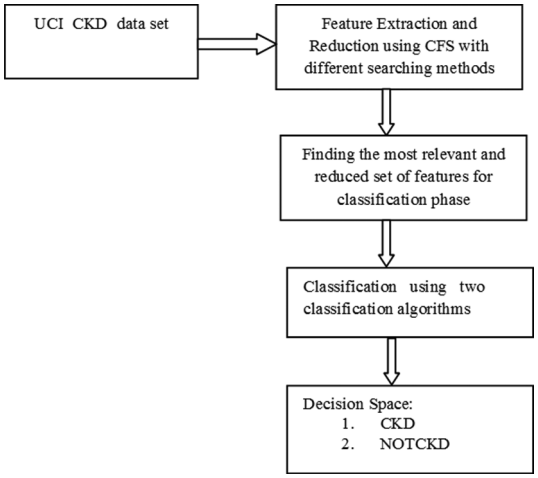

Figure 1: Block diagram of proposed system for chronic kidney disease

Table 4: Network parameters applying to UCI data set with reduced features

Classifiers	Network structure			Epochs (retrain)	Numbers patterns		
	I	HL	O		Training	Validation	Testing
CFS + IBPLN	8	6	1	2000 (10)	112	23	23
CFS + LM	8	6	1	2000 (10)	112	23	23

CFS: Correlation-based feature subset selection, LM: Levenberg–Marquardt, IBPLN: Incremental back propagation learning networks

Network Architecture

The artificial neural network model selection includes a choice of network architecture and feature selection. Theoretically, a network with one hidden layer and logistic function as the activation function at the hidden and output nodes is capable of approximating any function arbitrarily closely provided that the number of hidden nodes is large enough.[17] Hence, we used one input layer, one hidden layer, and one output layer. The number of hidden nodes is evaluated from the following formula as proposed by Huang *et al*.[18]

$$S = \sqrt{(0.43\ mn + 0.12\ n^2 + 2.54\ m + 0.77\ n + 0.35)} + 0.51 \quad (4)$$

In the present study, m = number of input nodes = 8, n = number of output = 2, and hence s = 6 after round off. Hence, in our study, we use six neurons at the hidden layer for all combinations.

Modeling Results

The classification algorithms were implemented in Alyuda NeuroIntelligence.[19] We used Intel Core 2 Duo Processor E7400 CPU (2.8 GHz Dual Core, 1066 MHz FSB, 3 MB L2 cache) with 2048 MB DDR2 RAM for implementation.

Table 4 shows the network structure, epochs, number of retrains, and number patterns used in training, validation, and testing phases. As overtraining control measure, we retain the copy of the network with the lowest validation error.

Performance evaluation

To predict the performance of the system, we computed correct classification accuracy (CCR), specificity, sensitivity, and receiver operating characteristic area under the curve (AUC) as these are very important parameters to predict the performance of the system without knowing the distribution of data. We computed true positive (TP), true negative (TN), false positive (FP), and false negative (FN) to further compute other performance parameters as discussed below:

$$\text{CCR}\,(\%) = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (5)$$

$$\text{Specificity}\,(\%) = \frac{TN}{TN + FP} \times 100 \quad (6)$$

$$\text{Sensitivity}\,(\%) = \frac{TP}{TP + FN} \times 100 \quad (7)$$

Further, we had drawn curves using sensitivity and specificity and computed the AUC. The high value of AUC indicates high-performance matching. The compiled results from 100 simulations are shown in Table 5. We observe that both the classification methods show excellent performance justifying the validity of the reduced set of features. However, out of two methods, CFS + LM shows slightly better performance in terms of CCR and specificity.

Table 5: Test results with 100 simulations

Classifiers	Test set (CCR %)			Specificity (%)			Sensitivity (%)			AUC (%)		
	Highest (frequency)	Lowest (frequency)	Average	Highest (frequency)	Lowest (frequency)	Average	Highest (frequency)	Lowest (frequency)	Average	Highest (frequency)	Lowest (frequency)	Average
CFS + IBPLN	100 (87)	95.65 (13)	99.43	100 (87)	94.74 (9)	99.33	100 (100)	-	100	100 (100)	-	100
CFS + LM	100 (95)	95.65 (5)	99.78	100 (95)	94.74 (3)	99.74	100 (100)	-	100	100 (100)	-	100

CFS: Correlation-based feature subset selection, LM: Levenberg–Marquardt, IBPLN: Incremental back propagation learning networks, CCR: Correct classification accuracy, AUC: Area under curve

Conclusions

Our findings suggest a reduced set of eight features: specific gravity, albumin, serum creatinine, hemoglobin, packed cell volume, white blood cell count, red blood cell count, and hypertension as more significant for investigating CKD. To justify the validity of the reduced feature set, we deployed two different binary classifiers. The classifiers show excellent performance. Hence, we suggest that these reduced feature set might be worthwhile to scrutiny when the final decision is made by the doctors. Reduced parameters reduce laboratory costs and time. At the same time, reduced features reduce uncertainty in decision-making. We suggest that the techniques used here could be applied to other diseases.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

References

1. Levey AS, Coresh J. Chronic kidney disease. Lancet 2012;379:165-80.
2. National Kidney Foundation. K/DOQI clinical practice guidelines for chronic kidney disease: Evaluation, classification, and stratification. Am J Kidney Dis 2002;392 Suppl 1:S1-266.
3. Anderson J, Glynn LG. Definition of chronic kidney disease and measurement of kidney function in original research papers: A review of the literature. Nephrol Dial Transplant 2011;26:2793-8.
4. Available from: https://www.archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease. [Last accessed on 2016 Aug 08].
5. Hall MA. Correlation-based Feature Subset Selection for Machine Learning. New Zealand: Hamilton; 1998.
6. Polat K, Sahan S, Kodaz H, Gunes S. A new classification method for breast cancer diagnosis: Feature selection artificial immune recognition system (FS-AIRS). Advances in Neural Computation. LNCS 3611. Berlin; Heidelberg: Springer; 2005. p. 830-8.
7. Fayyad UM, Irani KB. Multi-interval Discretization of Continuous-valued Attributes for Classification Learning. Proceedings XIIIth International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann; 1993. p. 1022-7.
8. Quinlan JR. C 4.5: Programs for Machine Learning. San Francisco; USA: Morgan Kaufmann; 1993.
9. Rich E, Knight K. Artificial Intelligence. NY; USA:McGraw-Hill; 1991.
10. Yu L, Liu H. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Vol. 20. 12th International Conference Machine Learning 2003 (ICML-2003).Washington Dc; 2003. p. 856-63.
11. Michalak K, Kwasnicka H. Correlation-Based feature selection strategy in classification problems. Int J Appl Math Comput Sci 2006;16:503-11.
12. Bouckaert RR, Frank E, Hall M, Kirkby R, Reutemann P, Seewald A, Scuse D. WEKA Manual for Version 3-6-2. The University of Waikato; 2010.
13. Fu L, Hsu HH, Principe JC. Incremental backpropagation learning networks. IEEE Trans Neural Netw 1996;7:757-61.
14. Rumelhart DE, Hinton GE, Williams RJ. Learning internal representation by error propagation. In: Rumelhart DE, McClelland JL, editors. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1. Cambridge, MA: MIT Press; 1986. p. 318-62.
15. Levenberg K. A method for the solution of certain problems in least-squares. Q Appl Math 1944;2:164-8.
16. Marquardt DW. An algorithm for least-squares estimation of nonlinear parameters. J Soc Ind Appl Math 1963;11:431-41.
17. Hornik K, Stinchcombe M, White H. Multilayer feed forward networks are universal approximator. Neural Netw 1991;2:359-66.
18. Huang ML, Hung YH, Chen WY. Neural network classifier with entropy based feature selection on breast cancer diagnosis. J Med Syst 2010;34:865-73.
19. Available from: http://www.alyuda.com. [Last accessed on 2016 Aug 08].