Prediction of Diabetic Chronic Kidney Disease Progression Using Data Mining Techniques

N. Afhami

Kerman Graduate University of Advanced Technology, Kerman, Iran
n.afhami@staff.kgut.ac.ir

***Abstract*—Chronic Kidney Disease is a global public health problem. In CKD, kidneys will be damaged.Kidneys may get damaged from diabetes disease.Some patients involve in diabetes and chronic kidney disease simultaneously.The objective of this paper is to predictmortality and progression of the disease to thelast stagein patient involved with the diabetic chronic kidney disease. For this purpose, different methods of data mining have been used. Also, we extract rule with using rough set theory. Different features have been applied from diabetic chronic kidney patients, and we have been predicted the disease progression using J48, naïve Byes, Bayesian Network, SVM, SMO, Bagging, Random Forest, Multilayer Perceptron methods.In conclusion, we have compared these performances according to Recall, Precision and F-Measure. Experimental results shows Random Forest has better performance in compare with other methods.**

Index Terms—Predication,Datamining, Rule Mining, Diabetic chronic kidney disease.

1. INTRODUCTION

Today, large scale data of patients are being collected by the patient care centers, that could become a source to discover disease-dependent hidden patterns. This subject makes data mining having too application in health. This research aimed to predict diabetic chronic kidney disease progression using data mining techniques. We only considered reaching the last stage (including ESRD and doubling serum creatinine) of kidney disease or patients' death for progression of diabetic chronic kidney disease.

CKD is a disease in which kidney'sefficiency will decrease gradually and can't perform well. Thus kidney will be injured and can't filter toxic futile. Now, CKD has become a world hygiene problem. It must be realized soon to decrease progression of this disease and should be tried to treat it. This disease has not any special signs andscreening test should be applied to identify. One of occurring factors of this disease is diabetes. In this research, we have been predicted the disease progression in patient involved with the diabetic chronic kidney disease.

In this paper, seven methods are studied to predict the conditions of disease progression. J48, Naïve Bayes, Bayesian Network, SVM, SMO, Multilayer perceptron, Bagging and Random Forest methods are employed for this purpose. Also In order to predict the disease progression, we used rule mining roughset theory and extract rule for disease progression.

Many researches have studied,realizing CKD using data mining techniques.[1] predicts chronic kidney disease. In this article they usedsix algorithms and choose Multilayer Perceptron for classification. In [2] they useddecision support system for diagnosing patients with Chronic Renal Failure, CRF. This article uses SVM and logistic regression for classification. In [3] and [4] they used datamining technique for chronic kidney disease prediction. [5] predicts kidney disease With SVM and ANN. Also [6,7,8,9]are researches in this field.

We have paid to introducing applied method in section 2and Experimental results have described in section 3. Experimental results are given in table related to disease predication (table 3), and extracted rule explain in figure 4.

2. METHOD

This paper have been utilized different properties of diabetic chronic kidney patients and have been studied the relationship of risk factors on mortality and reaching to the last phase of kidney disease. Data Set isutilizedfrom [10] including the information of 249 patients withdiabetic chronic kidney disease. They are in any stage of CKD except the last stage of the disease. Different properties of those can apply to prognosis the disease progression(Reaching the last stage or death) and these properties are high risk factors. Input dataset includes properties of 249 patientswhich 49 patientsdied and 40 patients reached thelast stage of the disease.Patient characteristics have displayed in Table 1.

Algorithm general levels have displayed in figure. Algorithm levels are: First, pre-processing phase on input dataset and then data mining phase; se we have been selectedthe suitable model to prognosis.To choose suitable model, different algorithms are utilized and compared on input dataset.

Table 1. Patient characteristics.

Number	Name	Description
1	Age	Age
2	Sex	Sex
3	BMI	Body mass index
4	Type2	Type 2 Diabetes
5	Duration	Duration of Diabetes
6	HTN	Hypertension
7	Hb	Hemoglobin
8	CRP	C-reactive protein
9	Alb	Albumin
10	HbA1c	Glycosylated hemoglobin
11	Ferritin	Ferritin
12	EPO	Endogenous erythropoietin
13	Chol	Cholesterol
14	Hepcidin (ngml)	Serum hepcidin levels
15	GFR	Glomerular filtration rate
16	Protein	Protein
17	Death or progression	Death or Risk class

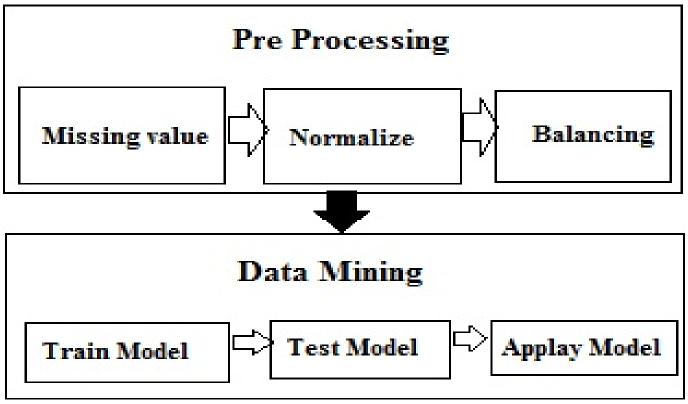


Figure 1. algorithm

We used cross validation in order to test the results. Finally, we have been selected the best method to prognosis.

2.1 Pre-processing:

Pre-processing level includes three following levels:

- Replace missing values: First missing values study and replace, character average value is used here.
- Normalize: In this level, number normalize domain is between 0 and 1.
- Balancing: at last, we studied and solved the problem of imbalanced data. To do this, we have used SMOTE algorithm.We will explain it in 2.1.1.

2.1.1 SMOTE algorithm:

Solving imbalanced data problemis applicable in data mining.Imbalanced data problem exists when sample numbers of one class is much more than the others. At this time, majority will be chosen. In order to balance the data, SMOTE algorithm has been used [11].

In this data set, the number of patients which died or reached thelast stage of the disease are lower than other samples. We used the algorithm in pre-processing level.Therefore, the effects vanished due to the class imbalance.

Two class data distributions changed in this procedure.

Of course, this algorithm doesn’t change learning algorithm. Resampling applies to data distribution changes from input data space. Minority and Majority of classes change to balance data. New data produce to minority using nearest neighbors methods. In addition, The number decrease from majority class, until exist better balancing.

2.2 Data Mining:

Learning level with different methods of data mining perform after pre-processing phase.

The dataset were classified using different methods of data mining. Various algorithms wereclassified input features

andcompared their performance and applied the best. These methods include J48, Naïve Bayes, Bayesian network, SVM, SMO, Multilayer perceptron, Bagging, Random Forest.Experimental results are provided in section 3.

2.3 Rough Set Theory:

This theory introduced by Z.Palwak in 1982 [12]. It’s used to analyze and classify uncertain and incomplete data and it’s a non-statistical method [13].Rough set theory is used in fields such as data mining, digital systems, signal processing and etc [14]. Its basic concept is diagnosing the lowest and highest of one collection borders. Approximating these borders subset is called Rough set. We use Rough set theory to mine rule prediction of patients’ death or reaching to the last stage of doubling serum creatinine.For this purpose we used the ROSE toolkit.

ROSE: Rough Set Data Explorer software [15] introduces different methods application on dataset inputs and classifies samples to positive and negative category.

3. Results and discussion:

Experiments have been done on dataset input including properties of 249 patients with diabetic chronic kidney disease. Different methods have been used. J48, Random Forest, Bagging, Multilayer perceptron, SMO, SVM, Bayesian network, Naïve Bayes have tested the dataset input to prognosis mortality and reaching the last stage of the disease in CKD patients. 10 folds cross validation is utilized to study and test the result.

In following tables2 and 3we can seethe results of the experiments.In table 2, we compared different classifiers in Percision, Recall, F- measure and ROC Area. In table 3, we explained different properties of Random Forest.

These results have been shown in following figures. In figure2we compared different methods with precision and Recall. Random Forest has reached to highest precision and Recall. Infigure 3 we displayed different methods F-measure, and method Roc Area. Random Forest is higher in four diagrams and have better performance. It’s Precision,Recall and F-measure is shown in table 3 and all of them are 85.

Table 2. Comparison of different classifiers

Classifier	Precision	Recall	F-measure	ROC Area
J48	72.90%	72.80%	72.90%	75.40%
Naïve Bayes	74.30%	74.10%	74.20%	80.40%
Bayesian Network	81.90%	81.80%	81.80%	87.10%
SVM	73.50%	71.90%	69.70%	67.40%
SMO	76.50%	76.70%	76.50%	75.20%
Multilayer Perceptron	79.40%	79.20%	79.30%	86.90%
Bagging	83%	83.10%	83%	89.80%
Random Forest	85.70%	85.60%	85.50%	90.40%

Experimental method shown in the following table displays the details of applying Random Forest algorithm. As it’s shown, Random Forest method has better efficiency from the other methods.

Table 3. Properties of using Random Forest

True Positive Rate	**False Positive Rate**	**Precision**	**Recall**	**F-Measure**	**Roc Area**	**Class**	
0.914	0.227	0.854	0.914	0.883	0.904	0	
0.773	0.086	0.861	0.773	0.815	0.904	1	
0.856	0.169	0.857	0.856	0.855	0.904		**Weighted Average**

The result of these two tables are shown in below diagram. Figure 2 shows Precision and Recall and figure 3 shows F-measure and ROC Area of different methods.

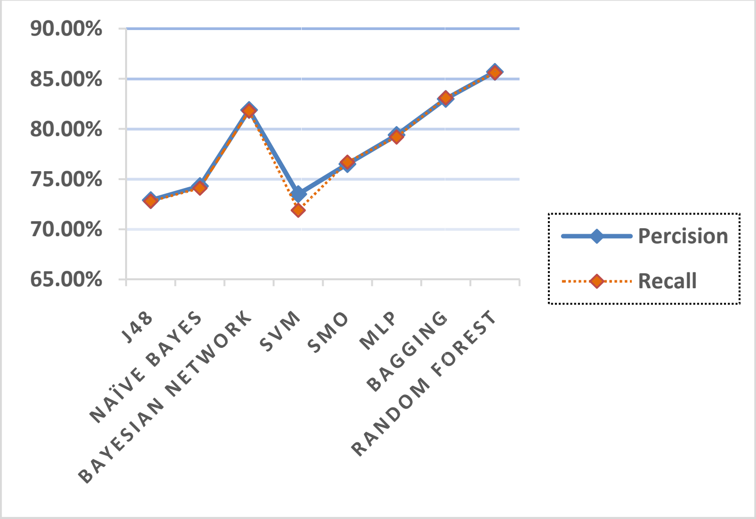


Figure 2. Precision and Recall using different methods

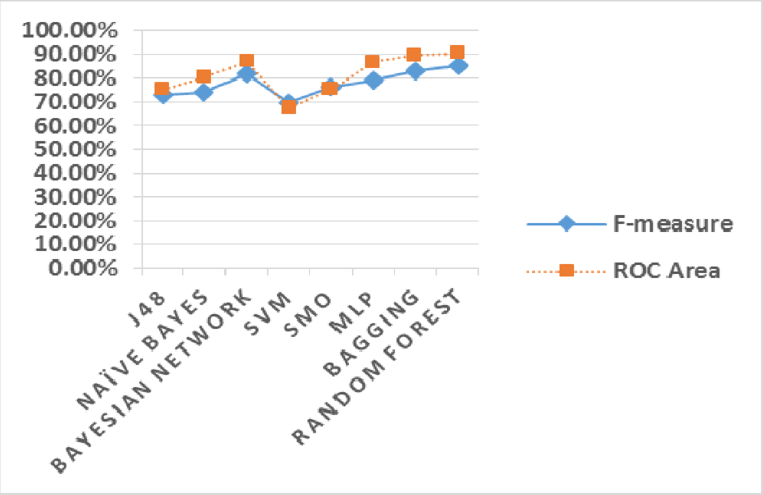
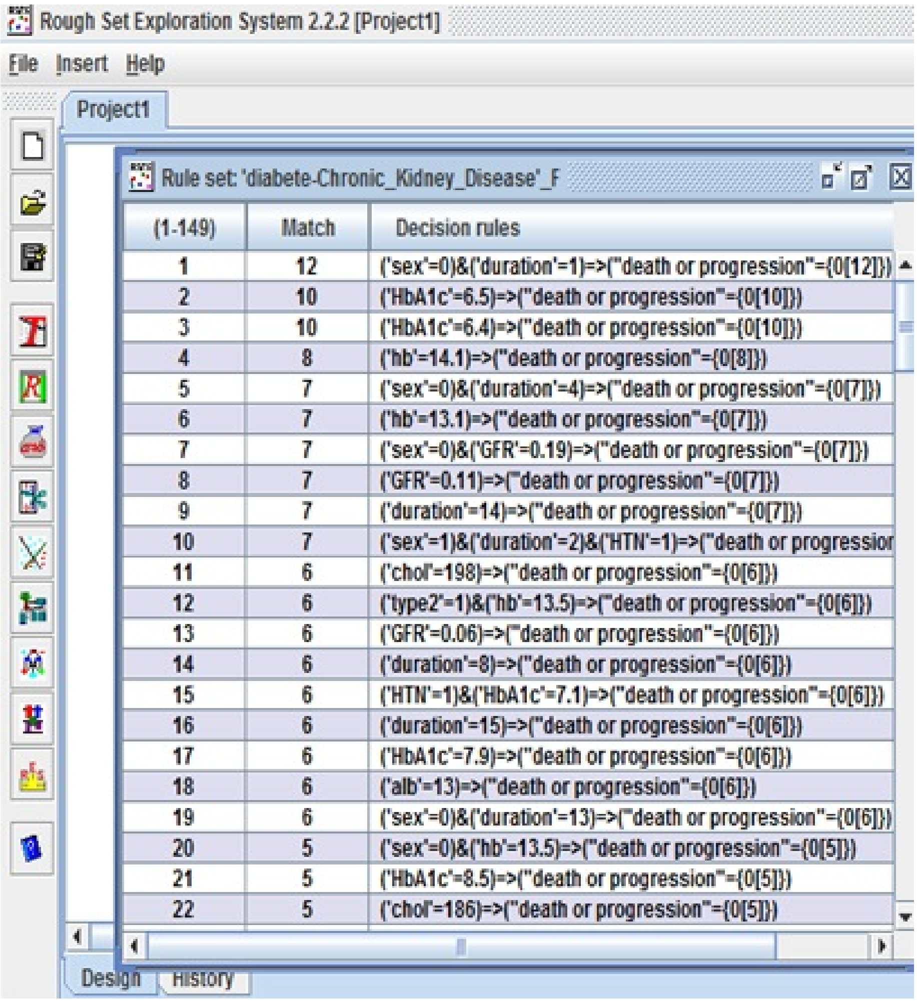


Figure 3. F-measure and ROC Area using different methods

Results of rule mining by rough set theory for prognosis disease progression or patient death displayedin figure4. We used ROSE toolkit [15]to implement rule mining with rough set theory.


Rough Set Exploration System 2.2.2 [Project1]
File
Insert
Help
Project1
Rule set: 'diabete-Chronic_Kidney_Disease'_F
Design
History

(1-149)	Match	Decision rules
1	12	('sex'=0)&('duration'=1)=>("death or progression"={0[12]})
2	10	('HbA1c'=6.5)=>("death or progression"={0[10]})
3	10	('HbA1c'=6.4)=>("death or progression"={0[10]})
4	8	('hb'=14.1)=>("death or progression"={0[8]})
5	7	('sex'=0)&('duration'=4)=>("death or progression"={0[7]})
6	7	('hb'=13.1)=>("death or progression"={0[7]})
7	7	('sex'=0)&('GFR'=0.19)=>("death or progression"={0[7]})
8	7	('GFR'=0.11)=>("death or progression"={0[7]})
9	7	('duration'=14)=>("death or progression"={0[7]})
10	7	('sex'=1)&('duration'=2)&('HTN'=1)=>("death or progressio
11	6	('chol'=198)=>("death or progression"={0[6]})
12	6	('type2'=1)&('hb'=13.5)=>("death or progression"={0[6]})
13	6	('GFR'=0.06)=>("death or progression"={0[6]})
14	6	('duration'=8)=>("death or progression"={0[6]})
15	6	('HTN'=1)&('HbA1c'=7.1)=>("death or progression"={0[6]})
16	6	('duration'=15)=>("death or progression"={0[6]})
17	6	('HbA1c'=7.9)=>("death or progression"={0[6]})
18	6	('alb'=13)=>("death or progression"={0[6]})
19	6	('sex'=0)&('duration'=13)=>("death or progression"={0[6]})
20	5	('sex'=0)&('hb'=13.5)=>("death or progression"={0[5]})
21	5	('HbA1c'=8.5)=>("death or progression"={0[5]})
22	5	('chol'=186)=>("death or progression"={0[5]})

Figure 4. Rule mining with ROSE toolkit.

4. Conclusion:

This paper,we predicate the disease progression in patients with diabetic chronic kidney. It has been done based on properties of 249 patients. First stage isinput dataset pre-process and in this stage missing value replacing, normalizing and balancing are done. Then in second stage, different techniques appliedon dataset. J48, Naïve Bayes, Bayesian Network, SVM, SMO, Random Forest, Bagging, Multilayer perceptron were employed on input dataset. Efficiency of different methods studied and compared using precision, Recall, F-measure. Random Forest had better performance.Also we extract rulefor prediction of the disease progression to thelast stage of renal disease or mortality.

In future works, We can study patient condition and the disease progression, over a period of time.

References

[1] Jena, L. and Kamila, N. Ku., 2015, “Distributed Data Mining Classification Algorithms for Prediction of Chronic Kidney Disease”, International Journal of Emerging Research in Management and Technology, 4(11), pp.110-118.

[2] Al-Hyari, A. Y., Al-Taee, A. M., Al-Taee, M. A., 2014, “Diagnosis and Classification of Chronic Renal Failure Utilising Intelligent Data Mining Classifiers”, International Journal of Information Technology and Web Engineering, 9(4), pp. 1-12.

[3] Rubini, L. J. and Eswaran, Dr. P., 2015, “Generating Comparative Analysis of Early Stage Prediction ofChronic Kidney Disease”, International Journal of Modern Engineering Research, 5(7), pp. 49-55.

[4] Kunwar, V.,Chandel, Kh.,Sabitha,A. S. and Bansal, A.,2016,“Chronic Kidney Disease analysis using data mining classification techniques”. Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference, pp. 300-305.

[5] Vijayarani, Dr. S., Dhayanand, Mr. S., 2015,“KIDNEY DISEASEPREDICTION USING SVM AND ANN ALGORITHMS”.International Journal of Computing and Business Research (IJCBR),1(3), pp. 1765-1771

[6] Chiu,R. K.,Chen,R. Y., Wang,S. A. andJian,S. J. ,2012,“Intelligent systems on the cloud for the early detection of chronickidney disease”. InMachine Learning and Cybernetics (ICMLC), 2012International Conference on(Vol. 5, pp. 1737-1742). IEEE

[7] Ravindra,B. V.,Sriraam,N. andGeetha, M.,2014,“Discovery of significant parameters in kidney dialysis data sets by Kmeansalgorithm”. InCircuits, Communication, Control and Computing(I4C), 2014 International Conference on (pp. 452-454). IEEE.

[8] Ahmed, S.,TanzirKabir, M., TanzeemMahmood, N. andRahman,R.M., 2014,“Diagnosis of kidney disease using fuzzy expertsystem”. InSoftware, Knowledge, Information Management andApplications (SKIMA), 2014 8th International Conference on (pp. 1-8).IEEE.

[9] Pellakuri, V. and Rao, D. R., 2016, “Rough Set Reasoning Based Classification Model Generating Decision Rule on Early Stage of Chronic Kidney Disease”.International Journal of Advance Research in Computer Science and Management Studies ,4(1), pp. 202-207

[10] Wagner, M.,Ashby, DR., Kurtz, C.,Alam, A.,Busbridge, M.,Raff, U., Zimmermann, J.,Heuschmann,PU.,Wanner, C. and Schramm, L., 2015,“Hepcidin-25 in diabetic chronic kidney disease is predictive for mortality and progression to last stage renal disease”. PLOS ONE 10(4): e0123072. http://dx.doi.org/10.1371/journal.pone.0123072

[11] Chawla, N. V., Bowyer, K. W., Hall, L. O. andKegelmeyer, W. P., 2002 ,“Smote: Synthetic minority over-sampling technique”. Journal of Artificial Intelligence Research, 16:321-357.

[12] Palwak, Z., 1982, “Rough Sets”, International Journal of Computer & Information Sciences, vol. 11, no. 5, pp. 341-356

[13] Pattaraintakorn, P. andCercone, N., 2008, “Integrating rough set theory and medical applications”, Applied Mathematics Letters,21(4), pp. 400-403

[14] Kaneiwa, K. and Kudo, Y., 2011, “A sequential pattern mining algorithm using rough set theory”,International Journal of Approximate Reasoning, vol. 52, no. 6, pp. 881-893

[15] ROSE 2.0. http://www.idss.cs.put.poznan.pl/rose. 1999