Role of Attributes Selection in Classification of Chronic Kidney Disease Patients

Naganna Chetty
Research Scholar at UTU, Dehradun
and
Dept. of CSE, MITE, Mangalore, India
nsc.chetty@gmail.com

Kunwar Singh Vaisla
Dept. of Computer Science and Engg.,
BTK Institute of Technology Dwarahat,
Uttarakhand, India
vaislaks@rediffmail.com

Sithu D Sudarsan
ABB Corporate Research, India
sdsudarsan@gmail.com

***Abstract*—In the present days the Chronic Kidney Disease (CKD) is a common problem to the public. CKD is generally considered as kidney damage and is usually measured with the GFR (Glomerular Filtration Rate). Several researchers from health care and academicians are working on the CKD problem to have an efficient model to predict and classify the CKD patient in the initial stage of CKD, so that the necessary treatment can be provided to prevent or cure CKD. In this work classification models have been built with different classification algorithms, *Wrappersubset* attribute evaluator and *bestfirst* search method to predict and classify the CKD and non CKD patients. These models have applied on recently collected CKD dataset downloaded from the UCI repository. The models have shown better performance in classifying CKD and non CKD cases. Results of different models are compared. From the comparison it has been observed that classifiers performed better on reduced dataset than the original dataset.**

Keywords—Data mining; classification; prediction; chronic kidney disease; attributes reduction

I. INTRODUCTION

The real world data gets doubles every 20 months and contain some amount of noise in it. Hence there is a need to store, manage and process this data efficiently. With the rapid advances in storage devices, it is easier to store and manage the vast amount data. Even though continuous effort has been made, the efficient processing of huge amount of data is still a challenge to the researchers, academicians, etc. This challenge can be handled with the data mining techniques. Data mining is an essential activity in KDD (Knowledge Discovery in Databases) process, which extracts patterns from observed data.

Health Informatics is producing vast amount of data and processing of this generated huge amount of data creates more possibilities for knowledge to be gained. This gained information can improve the service quality of healthcare to patients. The number of issues arise when dealing with this vast amount of data, one among them is how to analyze this data in a reliable manner. The basic goal of Health Informatics is to use real world medical data to improve our understanding of medicine and medical practice [1]. The present work emphasizes on an application of data mining, in particular, the classification techniques in health informatics to detect Chronic Kidney Disease (CKD).

CKD is usually a silent condition. Signs and symptoms, if present, are generally nonspecific and unlike several other chronic diseases (such as congestive heart failure, and chronic obstructive lung disease), they do not reveal a clue for diagnosis or severity of the condition. Typical symptoms and signs of uremia appear almost never in early stages (Stage 1 to 3, and even in Stage 4) and develop too late only in some patients in the course of CKD. Still, all newly diagnosed CKD patients, patients with an acute worsening in their kidney function and CKD patients on regular follow-up should have a focused history and physical examination. This will be the key to perceive real implications of health associated with decreased kidney function in CKD [2].

CKD is defined as damage to kidney or Glomerular Filtration Rate (GFR) < 60 mL/min/1.73 m^2 for 3 months or more, irrespective of the cause. Kidney damage in kidney related diseases can be caused by the presence of albuminuria, defined as albumin to creatinine ratio >30 mg/g in two of three spot urine specimens. GFR can be estimated from calibrated serum creatinine and estimating equations, such as the Modification of Diet in Renal Disease (MDRD) study equation or the Cockcroft-Gault formula [3].

GFR is traditionally measured as renal clearance of an ideal filtration marker, such as inulin from plasma. This measured GFR is considered the gold standard but is not practical for daily clinical use due to complexity of the measurement procedure. Estimating GFR based on a filtration marker (usually serum creatinine) is now widely accepted as an initial test. Several GFR prediction equations that use serum creatinine or some other filtration markers along with certain patient characteristics (like age, gender, and race) are giving precise estimates of GFR in various clinical settings [4].

Different stages and action plan for CKD are shown in TABLE 1. Here CKD stages from 1-5 along with the GFR reading and actions required during each stage are described.

978-1-4673-9354-6/15/$31.00 ©2015 IEEE

TABLE 1 STAGES AND ACTION PLAN FOR CKD [5]

Stage	Description	GFR (mL/min/1.73 m^2)	Action
-	At increased risk for CKD	>=90 with risk factors	Screening CKD risk reduction
1	Kidney damage with normal or increased GFR	>=90	Diagnosis and treatment Slow progression of CKD Treat comorbidities Cardiovascular disease risk reduction
2	Mild decrease in GFR	60-89	Estimate progression
3	Moderate decrease in GFR	30-59	Evaluate and treat complications
4	Severe decrease in GFR	15-29	Prepare for renal replacement therapy
5	Kidney failure	< 15 or dialysis	Replacement if uremic

II. LITERATURE SURVEY

Kusiak et al. [6] developed a method to predict survival period for kidney patients with dialysis using data mining approach. To obtain information on the interaction between measured values and survival of patients, they used data preprocessing, data transformations, and a data mining. Two different data mining algorithms were employed for extracting knowledge in the form of decision rules. These rules are provided as input to a decision-making algorithm, which predicts survival of new unseen patients with an accuracy of 75% to 85%. The approaches introduced in their research have been applied and tested on data collected at four dialysis sites.

Haubitz et al. [7] collected urine samples of 45 patients and tested for protein/polypeptide patterns with a novel high throughput method, capillary electrophoreses online coupled to a mass spectrometer (CE-MS). CE-MS allows the fast and accurate evaluation of up to 2000 polypeptides in one urine sample. Then they compared igAN (igA Nephropathy) results with findings in 13 patients with membranous nephropathy (MN) and 57 healthy individuals.

Yeh et al. [8], integrated temporal abstraction with data mining techniques to analyze dialysis patient's biochemical data to develop a decision support system with the purpose of reducing hospitalization rate. These mined temporal patterns are used by clinicians to predict hospitalization of hemodialysis patients and to suggest immediate treatments to avoid future hospitalization.

A first nationwide survey of predialysis CKD in Asian children carried out by Ishikura and his team indicated that the prevalence of CKD stage 3 and stage 5 in children in Japan is 2.98 for the age group 3 months to 15 years cases/100000 children. Most children with CKD presented themselves with non-glomerular disease, most frequently CAKUT. They stated improved management of CAKUT, including renoprotective treatment and urological intervention, is needed [9].

Malluche et al. [10] demonstrated the value of assessing bone from multiple views and hierarchical levels to understand CKD–MBD-related abnormalities in bone quality. Knowing the relationships between variations in material, structure, microdamage, and mechanical properties of bone in patients with CKD–MBD should aid in the development of new modalities to prevent, or treat, these abnormalities.

Chase et al. [11] developed prediction models with Naïve Bayes and Logistics Regression classification algorithms to predict the progression of stage 3 CKD. They used data extracted from Electronic Health Record (EHR) to identify two cohorts of patients in stage 3: progressors (eGFR declined >3 ml/min/1.73m2/year; n = 117) and non-progressors (eGFR declined <1 ml/min/ 1.73m2; n = 364). Initial laboratory values recorded a year before to a year after the time of entry to stage 3, reflecting metabolic complications (hemoglobin, bicarbonate, calcium, phosphorous, and albumin) were obtained. They made the comparison of average values in progressors and non-progressors.

A hybrid decision support system is developed by Neves and his team, allowed to consider incomplete, unknown, and even contradictory information. This is complemented with an approach to computing centered on Artificial Neural Networks, in order to weigh the Degree-of-Confidence in terms of reasoning procedures and knowledge representation based on Logic Programming,. Their study involved 558 patients with an age average of 51.7 years and the chronic kidney disease was observed in 175 cases. The dataset consist twenty four variables, forming five main categories [12].

Vanlede et al. [13] examined the urinary excretion of monosaccharides and polyols in children with different CKD degrees, but without known metabolic or renal tubular disorders. Urinary concentrations of 18 monosaccharides and polyols were measured by gas chromatography–mass spectrometry (GC-MS) in random urinary samples of 25 patients with CKD stage1– stage 5. A comparison of these values are made with age-related reference values. Serum creatinine was measured at the time of the urine sample, and the height-independent estimated glomerular filtration rate (eGFR-Pottel) was calculated. Urinary excretions of monosaccharides and polyols were above the reference values in 8–88 % of all patients. A significant difference between

CKD stage 1– stage 2 compared with CKD stage 3–stage 5 was found for allose, arabitol and sorbitol ($p<0.05$) and for arabinose, fucose, myoinositol, ribitol, xylitol, and xylose ($p<0.01$). They concluded that the excretion of polyols and sugars depends on eGFR, which warrants a cautious interpretation of the results in patients with CKD.

Dubey et al. [14] adopted the K-Means Clustering Algorithm with a single mean vector of centroids, to classify and make clusters of varying probability of likeliness of suspect being prone to CKD. They observed and stated that the suspects falling in clusters K1 or K3 are surely suffering from CKD. The probability of a suspect lying in K2 cluster to fall in the class of CKD is 0.50545, which implies that the suspect cannot be classified by their L-factor classifier. However, suspects from clusters K1 & K3 were found to be falling in CKD class with full probability.

III. PROPOSED METHODOLOGY

In this section we propose a methodology for mining patters from the CDK dataset. Fig. 1 shows the proposed framework for mining patterns.
The proposed methodology consists of:

- A framework for pattern mining
- A methodology for attributes selection
- Applying classification method to original and reduced datasets.

Attribute selection is a process of reducing the dimension of a dataset by eliminating the attributes of less importance. A framework for attribute selection is shown in Fig. 2. We have used *wrappersubseteval* attribute evaluator with *bestfirst* search method to select features for Naïve Bayes, SMO (Sequential Minimal Optimization) and IBK classifiers using WEKA data mining tool.

Naïve Bayes, SMO and IBk classifiers are used to classify original and reduced CKD datasets. Naïve Bayes classifiers are simple classifiers with probability based on Baye's theorem. This possess strong (naive) independence assumptions between the features. SMO is an algorithm which solve problem of quadratic programming that occurs at the time of training support vector machines. IBk is one of the instance-based learner using class of the nearest *k* training instances for the class of the test instances. Parameter *k* is the number of nearest neighbors to use for prediction.

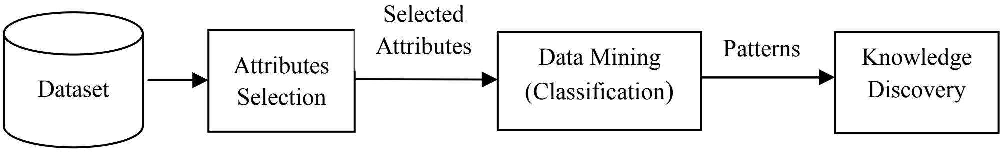

Fig. 1. Proposed Framework for Mining Patterns

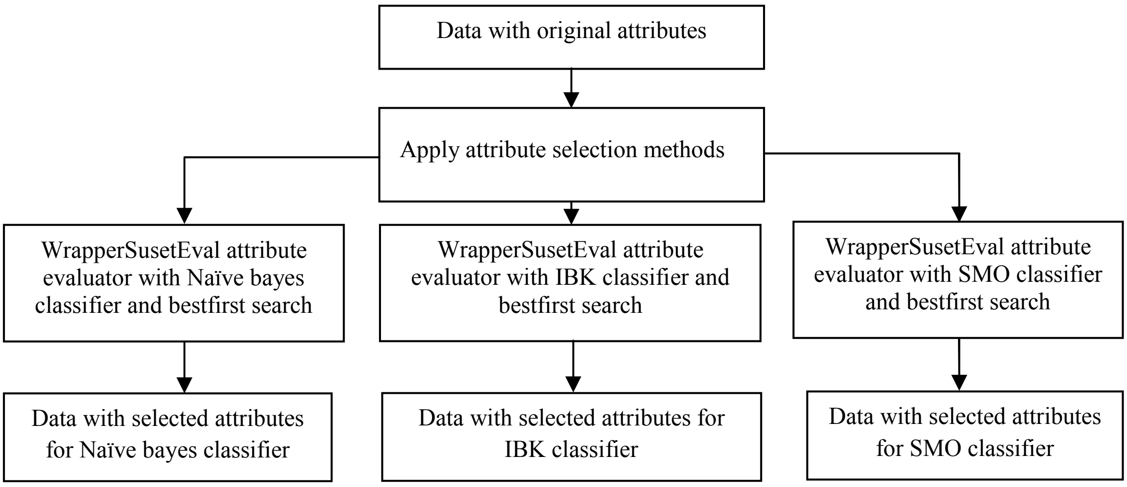
Data with original attributes
Apply attribute selection methods
WrapperSusetEval attribute evaluator with Naïve bayes classifier and bestfirst search
WrapperSusetEval attribute evaluator with IBK classifier and bestfirst search
WrapperSusetEval attribute evaluator with SMO classifier and bestfirst search
Data with selected attributes for Naïve bayes classifier
Data with selected attributes for IBK classifier
Data with selected attributes for SMO classifier

Fig. 2. Framework for Attributes Selection

IV. RESULTS

The dataset used in the present work is CKD, downloaded from the UCI repository [15]. The CKD dataset consists of 400 instances and 25 attributes viz. age, blood pressure, sugar, etc., including class label. TABLE 2 shows result of attributes reduction by the attribute evaluator with different classifiers. *WrapperSusetEval* attribute evaluator with Naïve bayes classifier and *bestfirst* search selects only 6 attributes from 25 total attributes with 76% of attributes reduction. *WrapperSusetEval* attribute evaluator with SMO classifier and *bestfirst* search results in 12 attributes from 25 total attributes with 52% of attributes reduction. *WrapperSusetEval* attribute evaluator with IBK classifier and *bestfirst* search selects only 7 attributes from 25 total attributes with 72% of attributes reduction. The graphical representation for the result of attributes selection is shown in Fig. 3.

After reducing the attributes, different classifier models have been applied on both original and reduced datasets. The result of classification on the original dataset is shown in the TABLE 3. The Naïve Bayes classifier classified 380 instances correctly and 20 instances incorrectly from total 400 instances of CKD dataset with the classification accuracy of 95%. SMO classifier classified 391 instances correctly and only 9 instances incorrectly from total 400 instances of CKD dataset with the classification accuracy of 97.75%. The IBK classifier classified 383 instances correctly and 17 instances incorrectly from total 400 instances of CKD dataset with the classification accuracy of 95.75%. Graphical representation of classification accuracies of different classifiers is shown in Fig. 4.

TABLE 4 shows the result of classification on reduced dataset. Naïve Bayes classifier classified 396 instances correctly and only 4 instances incorrectly from total 400 instances of CKD dataset with the classification accuracy of 99%. SMO classifier classified 393 instances correctly and only 7 instances incorrectly from total 400 instances of CKD dataset with the classification accuracy of 98.25%. The IBK classifier classified all the 400 instances correctly, this lead to the classification accuracy of 100%. Graphical representation of classification accuracies of different classifiers on reduced dataset is shown in Fig. 5.

TABLE 2 TABULAR REPRESENTATION OF ATTRIBUTES REDUCTION

Attribute Evaluator	**Initial Attributes**	**Selected Attributes**	**Attributes Reduction**
WrapperSusetEval attribute evaluator with Naïve bayes classifier and bestfirst search	25	06	76%
WrapperSusetEval attribute evaluator with SMO classifier and bestfirst search	25	12	52%
WrapperSusetEval attribute evaluator with IBK classifier and bestfirst search	25	07	72%

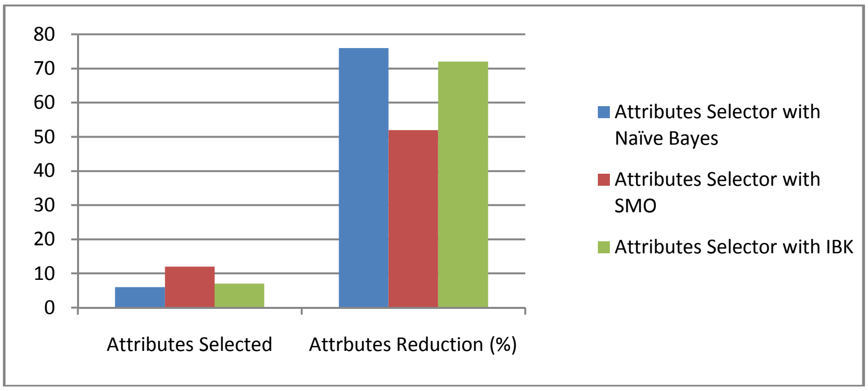

Fig. 3. Result of Attributes Reduction

TABLE 3 RESULT OF CLASSIFICATION ON ORIGINAL DATASET

Classifier	Correctly classified instances	Incorrectly classified instances	Classification accuracy
Naïve Bayes	380	20	95%
SMO	391	9	97.75%
IBK	383	17	95.75%

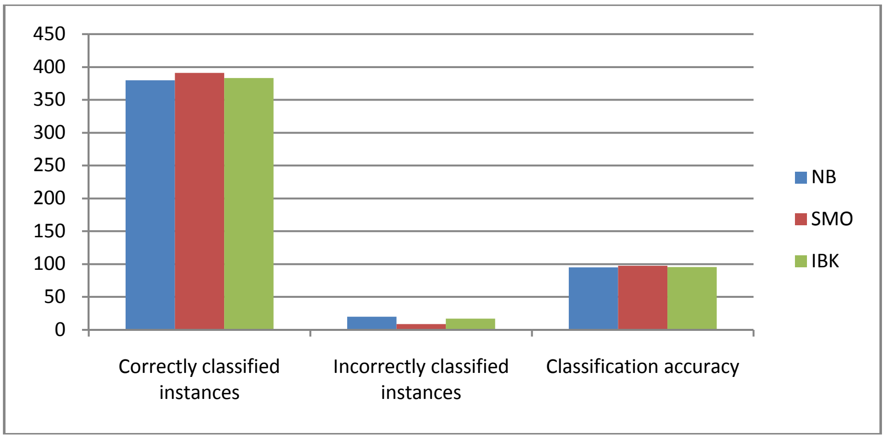

Fig. 4. Result of Classification on Original Dataset

TABLE 4 RESULT OF CLASSIFICATION ON REDUCED DATASET

Classifier	Correctly classified instances	Incorrectly classified instances	Classification accuracy
Naïve Bayes	396	4	99%
SMO	393	7	98.25%
IBK	400	0	100%

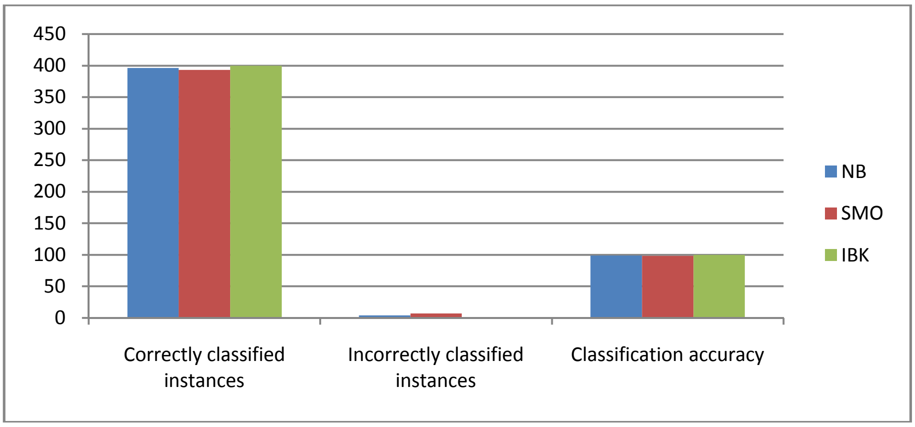

Fig. 5. Result of Classification on Reduced Dataset

V. CONCLUSION

Attributes evaluator and classification models have been applied on CKD dataset, uploaded very recently to UCI repository. From TABLE 2 it can be observed that attribute evaluator model has performed well by reducing the attributes from 25 to 6, 12 and 7 with NB, SMO and IBK classifiers respectively. TABLE 3 shows the classification accuracy for CKD original dataset as 95%, 97.75% and 95.75% for NB, SMO and IBK classifiers respectively. Here SMO classifier performs better than other two. Table 4 shows the classification accuracy for CKD reduced dataset as 99%, 98.25% and 100% for NB, SMO and IBK classifiers respectively. Here IBK classifier performs better than other two. Finally by referring to the results of Table 3 and Table 4, we can conclude that our models resulted in better classification accuracy on the reduced dataset than the original dataset.

REFERENCES

[1] M. Herland, T. M. Khoshgoftaar and R. Wald, “A review of data mining using big data in health informatics,” *Journal of Big Data,* pp. 1-35, 2014.

[2] M. Arici, “Clinical Assessment of a Patient with Chronic Kidney Disease,” *Management of Chronic Kidney Disease, Springer-Verlag Berlin Heidelberg,* pp. 15-28, 2014.

[3] A. S. Levey, K. Eckardt, U. Tsukamoto, A. Levin, J. Koresh, J. Rossert, D. D. Zeeuw, T. H. Hostetter, N. Lameire and G. Eknoyan, “Definition and classification of chronic kidney disease: A position statement from Kidney Disease: Improving Global Outcomes (KDIGO),” *Kidney International,* Vol. 67, pp. 2089-2100, 2005.

[4] L. A. Stevens, J. Coresh, T. Greene, A. S. Levey, “Assessing kidney function-measured and estimated glomerular filtration rate,” *N Engl J Med*, 354(23), pp. 2473-83, 2006.

[5] M. A. Parazilla and R. F. Reilly, “Chronic Kidney Disease: A New Classification and Staging System,” *Hospital Physician*, pp. 18-22, 45, March 2003.

[6] A. Kusiak, B. Dixon, S. Shah, “Predicting survival time for kidney dialysis patients: a data mining approach,” *Computers in Biology and Medicine 35*, pp. 311–327, 2005.

[7] Haubitz et al., “Urine protein patterns can serve as diagnostic tools in patients with igA nephropathy,” *Kidney International*, Vol.67, pp. 2313-2320, 2005.

[8] J. Y. Yeh, T. H. Wu, C. W. Tsao, “Using data mining techniques to predict hospitalization of hemodialysis patients,” *Decision Support Systems,* 50, pp. 439–448. 2011.

[9] Ishikura et al., “Pre-dialysis chronic kidney disease in children: results of a nationwide survey in Japan,” *Nephrol Dial Transplant,* pp.1-11, 2013.

[10] H. H. Malluche, D. S. Porter, and D. Pienkowski, “Evaluating bone quality in patients with chronic kidney disease,” *Nat Rev Nephrol,* 9(11), pp. 671-680, 2013.

[11] H. S. Chase, J. S. Hirsch, S. Mohan, M. K. Rao and J. Radhakrishnan, “Presence of early CKD-related metabolic complications predict progression of stage 3 CKD: a case-controlled study,” *BMC Nephrology,* 15:187, 2014.

[12] J. Neves, M. R. Martins, J. Vilhena, J. Neves, S. Gomes, A. Abelha, J. Machado, H. Vicente, “A Soft Computing Approach to Kidney Diseases Evaluation,” *J Med Syst, Springer,* 39: 131, 2015.

[13] K. Vanlede, L. A. J. Kluijtmans, L. Monnens, E. Levtchenko, “Urinary excretion of polyols and sugars in children with chronic kidney disease,” Pediatr Nephrol, 30, pp. 1537–1540, 2015.

[14] A. Dubey, “A Classification of CKD Cases Using MultiVariate K-Means Clustering,” *International Journal of Scientific and Research Publications*, Vol. 5, Issue 8, 2015.

[15] P. Soundarapandian and L. J. Rubini, http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease, UCI Machine Learning Repository, Irvine, 2015.