International Journal of Management, IT & Engineering
Vol. 8 Issue 8, August 2018,
ISSN: 2249-0558 Impact Factor: 7.119
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gage as well as in Cabell's Directories of Publishing Opportunities, U.S.A

Machine Learning Algorithm Improves Accuracy for analysing Kidney Function Test Using Decision Tree Algorithm

Dr. M. Mayilvaganan*

R.Rajamani**

Abstract

The research has been focused on retrieval of relevant information from the collection of data is a challenging task in data mining techniques [4]. In this research focuses on machine learning algorithm that involves to finding the pattern of health level among the collection of kidney function test to diagnosis and decision making. Data mining is a process or method that extracts interesting knowledge from small or large amounts of dataset, here machine learning algorithm implemented for allow a computer to learn which involves consciousness. An objective of the research is to comparing decision tree algorithm of Random Forest, C4.5 with Bootstrap Aggregation algorithm for analysing the time efficiency and accuracy of the required algorithm. An experiment result is observed the performance of selecting attribute by using two algorithms and analyse the better algorithm in classification.

Keywords:

C4.5 Algorithm with Bootstrap Aggregation (Bagging) Algorithm;
Random Forest;
machine learning algorithm;
pruned;
Unpruned;

Author correspondence:

Dr. M. Mayilvaganan
Associate Professor,
Department of Computer Science,
PSG College of Arts and Science,
Coimbatore.
research.igate@gmail.com

Introduction

In this research, to find the kidney function test, the blood Test for estimate GFR of patients. Blood will be tested for a waste product called creatinine. Creatinine comes from muscle tissue. When the kidneys are damaged, they have trouble removing creatinine from your blood. Testing for creatinine is only the first step. Next, the creatinine result is used in a math formula with your age, race, and gender to find out your glomerular filtration rate (GFR).

Kidney function tests are conducted to know whether all parameters of kidney are functioning within the normal range or not. These tests tell us what is the level of blood urea, creatinine, uric acid and

* Associate Professor

** Assistant Professor

Department of Computer Science, PSG College of Arts and Science, Coimbatore.

other minerals in the body. The normal values of this test ranges are represents in the given below in the table 1.

This test estimates how well the kidneys are filtering waste. Any result lower than 60 milliliters/minute/1.73m^2 may be a warning sign of kidney disease. GFR (Glomerular Filtration Rate) is a measure of kidney function and is performed through a blood test.

Scope of Research Methodology

To propose an efficient framework for analyzing the collection of patients data of kidney function test to determines the rate by looking at factors, such as specifically creatinine levels, age, gender, race, height, weight. The Serum test also involving for the finding the fluctuation level of blood urea, creatinine, uric acid and other minerals in the body under independent variable. Based on the glomerular filtration rate (dependent variable) and based on gender wise, to classify the pattern of kidney disease test by comparing Random forest and C4.5 Bootstrap Aggregation algorithm. Current comparative studies asses the performance of the algorithms based on the results obtained in time efficiency and accuracy.

Data Collection and Pre-Processing

In Preprocessing, the raw data of numerical data is converted into nominal data for classifying process. After the preprocess, apply three algorithm of machine learning technique of Random forest, C4.5 algorithm and C4.5 Bootstrap aggregation algorithm for analysing the accuracy of classification process.

Table 1. Collection of Patients data for finding Glomerular Filtration Rate of Kidney Function test

Cases	Age	Height (Cm)	Weight (Kg)	Race	Gender
1	32	148	65	White	M
2	32	152	68	black	M
3	48	162	75	white	F
4	50	165	70	black	F
5	30	158	58	White	F
6	28	159	59	White	F
7	56	163	57	White	F
8	35	164	54	black	F
9	55	172	56	black	M
10	47	175	59	black	M
11	35	178	65	White	M
12	32	165	56	White	M
13	30	179	69	White	M
14	29	158	70	black	M
15	35	165	75	black	M
16	49	163	80	black	F
17	45	170	85	black	F
18	36	159	120	White	F
19	43	168	135	White	M
20	37	172	95	White	M

Table 3: CKD EPI Equation for Estimating GFR Expressed for Specified Race, Sex and Serum Creatinine in mg/dL

Race	Sex	Serum Creatinine, S_{cr} (mg/dL)	Equation (age in years for ≥ 18)
Black	Female	≤ 0.7	$GFR = 166 \times (S_{cr}/0.7)^{-0.329} \times (0.993)^{Age}$
Black	Female	> 0.7	$GFR = 166 \times (S_{cr}/0.7)^{-1.209} \times (0.993)^{Age}$
Black	Male	≤ 0.9	$GFR = 163 \times (S_{cr}/0.9)^{-0.411} \times (0.993)^{Age}$
Black	Male	> 0.9	$GFR = 163 \times (S_{cr}/0.9)^{-1.209} \times (0.993)^{Age}$
White or other	Female	≤ 0.7	$GFR = 144 \times (S_{cr}/0.7)^{-0.329} \times (0.993)^{Age}$
White or other	Female	> 0.7	$GFR = 144 \times (S_{cr}/0.7)^{-1.209} \times (0.993)^{Age}$
White or other	Male	≤ 0.9	$GFR = 141 \times (S_{cr}/0.9)^{-0.411} \times (0.993)^{Age}$
White or other	Male	> 0.9	$GFR = 141 \times (S_{cr}/0.9)^{-1.209} \times (0.993)^{Age}$

Table 4: Data Collection of serum test in kidney function

Cases	C	U	TP	A	UA	P	CAL	BIC	PO	Sod
1	0.7	12	3	3.5	3.5	2.5	8.5	20	3.5	135
2	0.65	11	2	2.5	2.5	2	8	19.5	2.5	134.5
3	0.8	13	4	4.5	2.5	3	5	16.5	4.5	134
4	9.3	14	5	5.5	7	2	7.5	19	6	137.5
5	9.2	14	5	5.5	5.5	2	7.5	19	6	137.5
6	1	18	4	4.5	4.5	3	5	16.5	4.5	134
7	1.12	12	3	3.5	3.5	2.5	8.5	20	3.5	135
8	1	18	4	4.5	4.5	3	5	16.5	4.5	134
9	0.6	15	5	5.5	3.5	2	7.5	19	6	137.5
10	0.82	13	4	4.5	3	3	5	16.5	4.5	134
11	0.71	12	3	3.5	3.5	2.5	8.5	20	3.5	135
12	0.723	12	3	3.5	3.5	2.5	8.5	20	3.5	135
13	0.75	12	3	3.5	3.5	2.5	8.5	20	3.5	135
14	0.83	13	4	4.5	3	3	5	16.5	4.5	134
15	0.9	16	6	6	3	1.5	7	18.5	6	138

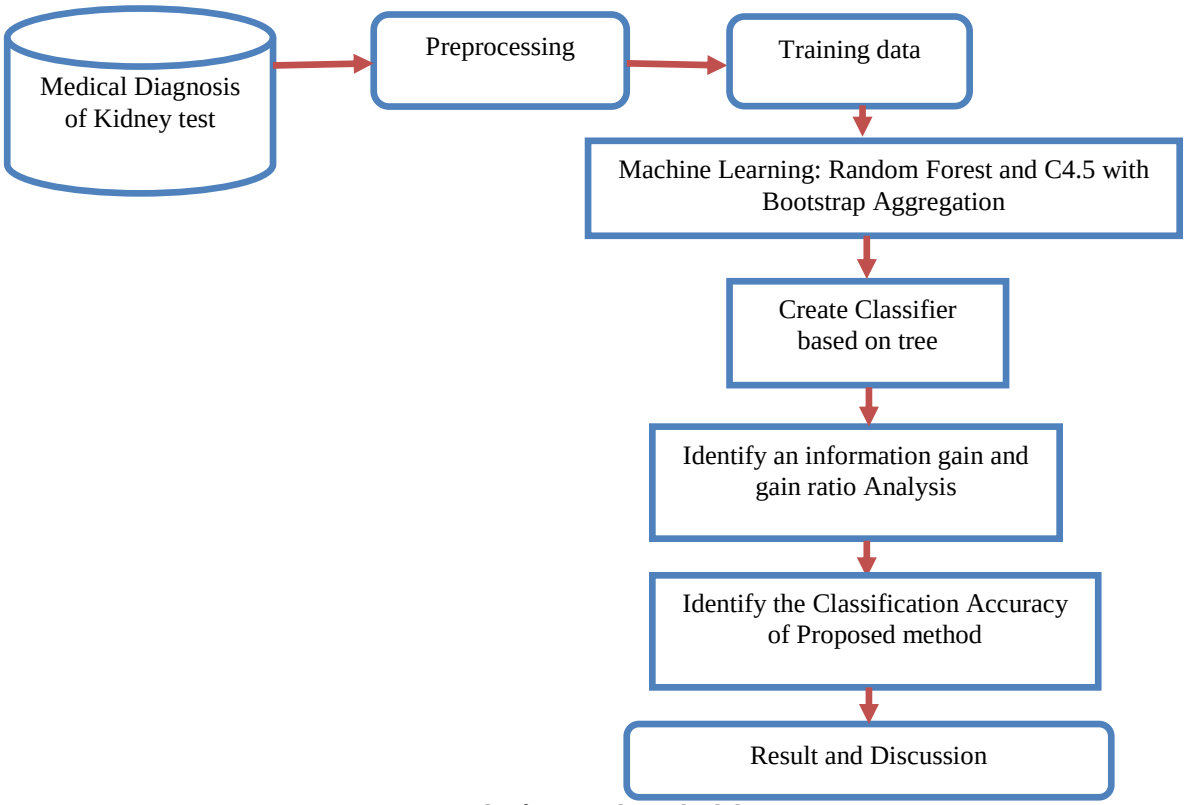

Fig. 1 Frame work of Research Methodology

Fig 1 represents the medical diagnosis of kidney function analysis with respect to the collection of Creatine, Urea nitrogen, total protein, Albumin, Uric acid in gender wise, Phosphate, Calcium, Bi-Carbonate, Potassium, Sodium, Serum in gender wise for analysing the fluctuation of ranges as shown in table 2, when predicting the data based on the Glomerular Filtration Rate it can identify the patient health condition early.

After preprocessing, apply machine learning technique of Random forest and C4.5 Bootstrap aggregation algorithm for analysing the accuracy of classification process.

Table 2: Ranges of Serum test Kidney

Serum Test for Kidney Function		**Ranges**			**Units**
		Low	**Normal**	**High**	
Creatine		<0.7	0.7- 1.4	>1.4	mg/dL
Urea nitrogen		<12	12- 20	>20	mg/dl
Total protein		<3.0	3.0-6.0	>6.0	g/dL
Albumin		<3.5	3.5- 5	>5	g/dL
Uric Acid	Female	<2.5	2.5-7	>7	mg/dL
	Male	<3.0	3.0- 3.5	>3.5	mg/dL
Phosphate		<2.5	2.5-4.5	>4.5	mg/dL
Calcium		<8.5	8.5- 10.5	>10.5	mg/dL
Bi-carbonate		<20	20- 30	>30	mEq/L
Potassium		<3.5	3.5- 5.5	>5.5	mEq/L
Sodium		<135	135- 145	>145	mEq/L
Serium Creatinine	Female	<0.6	0.6 -1.1	>1.1	mg/dL
	Male	<0.7	0.7- 1.3	>1.3	mg/dL

Table 2 represents the unit of mg/dL represents milligrams per Deciliter, unit of g/dL represents gram/deciliter and unit of mEq/L represents milli-equivalents per litre.

The decision tree can be constructed from the given set of attributes. Greedy strategy that grows a decision tree by making a series of locally optimum decision about which attributes to use for partitioning the data. Typical methods of machine learning algorithm of Random Forest and C4.5 with Bootstrap aggregation can be implemented to obtain high classification accuracy and time efficiency of medical diagnosis data. The performance can be identified and diagnosis for kidney function based on Serum test. Pre-pruning involves deciding when to stop developing sub-trees during the tree building process. The minimum number of observations in a leaf can determine the size of the tree. After a tree is constructed, the C4.5 rule induction program can be used to produce a set of equivalent rules. Pruning produces fewer, more easily interpreted results.

Proposed Methodology

Random Forest Algorithm

A random forest is a collection of unpruned decision trees [4]. It combines many tree predictors, where each tree depends on the values of a random vector sampled independently. In order to construct a tree, assume that 'N' is the number of training observations and "S" is the number of attributes in a training set. In order to determine the decision node at a tree, choose N<<S as the number of variables to be selected. Select a bootstrap sample from the N observations in the training set and use the rest of the observations to estimate the error of the tree in the testing phase. Randomly choose m variables as a decision at a certain node in the tree and calculate the best split based on the m variables in the training set. Trees are always grown and never pruned compared to other tree algorithms.

C4.5 Algorithm with Bootstrap Aggregation (Bagging) Algorithm

C4.5 is a decision tree technique which is enhanced by ID3 algorithm. It is one of the most popular algorithm for rule base classification [4]. Here an attributes can be split into two partition based on the selected threshold value, all the value satisfied by the constraint it will be assigned in one child and remaining values can be store in another child respectively. It also handles missing values. Here it can be gather of all nominal tests through entropy gain and the values are sorted based on the values in continuous attribute values which are calculated in one scan. This process is repeated for each continuous attributes when the process is terminated.

Steps of the System:

1. Selecting dataset as an input to the algorithm for processing.
2. Selecting the classifiers.
3. Calculate entropy, information gain, gain ratio of attributes.
4. Processing the given input dataset according to the defined algorithm of C4.5 data mining.
5. According to the defined algorithm of improved C4.5 data mining processing the given input dataset.
6. The data which should be inputted to the tree generation mechanism is given by the C4.5 and improved

C4.5 processors. Tree generator generates the tree for C4.5 and improved C4.5 decision tree algorithm

The rule set is formed from the initial state of decision tree. Each path from the initial state, the

condition will be evaluate and simplified by the effect of rule and an outcomes will put on the required leaf, the step will continuous when it comes discarding the condition. Let freq (C_i, S) stand for the number of samples in S that belong to class C_i (out of k possible classes), and $|S|$ denotes the number of samples in the set S. Then the entropy of the set of equation 1 such as

$$\text{Info(s)}=\sum_{i=1}^{k}((\text{freq}(c_i,s)/|s|).\log_2(\text{freq}(c_i,s)/|s|)) \quad (1)$$

After set T has been partitioned in accordance with n outcomes of one attribute test X by equation 2 and 3,

$$\text{Info}_X(S)=\sum_{i=1}^{n}\frac{|S_j|}{|S|}.\text{Info}\left(S_j\right) \quad (2)$$

$$\text{Gain(x)} = \text{Info(S)} - \text{Info}_x\text{(S)} \quad (3)$$

The gain ratio "normalizes" the information gain as following equation 4,

$$\text{GainRatio}(a_i,S)=\frac{\text{InformationGain}(a_i,S)}{\text{Entropy}(a_i,S)} \quad (4)$$

Pre-pruning involves deciding when to stop developing sub-trees during the tree building process.

To integrate C4.5 algorithm combines with Bagging improves generalization error by reducing the variance of the base classifiers. The performance of bagging depends on the stability of the base classifier. After training the x classifiers, a test instance is assigned to the class that receives the highest number of votes.

Input: D, Set of S training tuples;

Classification learning scheme: C4.5 Algorithm

Output: The ensemble- a composite model, Z*.

Algorithm: Bagging

Let u be the number of bootstrap samples

for j=1 to u **do** // create w models:

Create a bootstrap sample of size M, C_i by sampling C with replacement;

Use C_i and learning scheme to derive a Model, Z_i;

endfor

To use the ensemble to classify a tuple Y:

Let each of the u models classify Y and return majority vote;

It increases accuracy because the composite model reduces the variance of the individual classifiers.

Result and Discussion

Table 2 : Accuracy Comparison of Different Machine learning algorithm

Algorithm	Selected Variables (%)	Sensitivity (%)
Random Forest	61.6%	58%
C4.5 (Pruned)	71%	70.2%
C4.5 (Unpruned)	63%	61%

Table 2 represents the sensitivity is the probability that a test will indicate the classable of kidney function test condition, Random forest achieved 56% of selected variable and 58% of sensitivity variable. C4.5 Pruned achieved a classification accuracy of 71% of selected variables and 70.2% of sensitivity variable

and finally C4.5 unpruned achieved a classification accuracy of 63% in selected variable and 61% in sensitivity variable.

- **Sensitivity: True positive/(True positive + False Negative) × 100**

Specificity is the fraction of those without disease who will have a negative test result:

- **Specificity: True Negative/(True Negative + False Positive) × 100**

Here Sensitivity and specificity are characteristics of the test. The selected variables alone were used to find the sensitivity and specificity of the data mining algorithms. In C4.5 with Bootstrap Aggregation gives high accurate classification rate than C4.5 pruned, unpruned and random forest algorithm. When compared with C4.5 pruned and C4.5 with Bootstrap aggregation, it can be slightly vary for observing to determine the size of the tree.

Table 3 : Accuracy Comparison of Different Ensemble methods

Algorithm	Classified Instance Accuracy	In classified Instance Accuracy
Random Forest	70%	30%
C4.5 with Bagging Aggregation	74.2%	25.8%

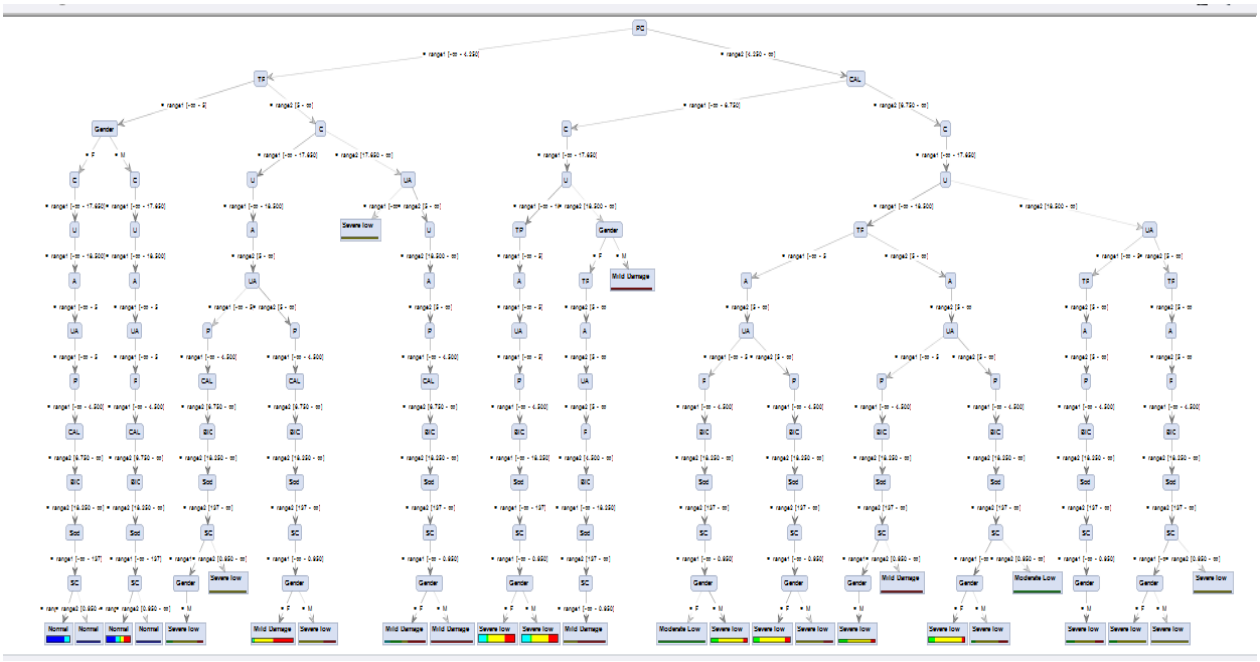

Fig. Tree view of C4.5 with Bagging Aggregation

Table 3 represents an accuracy of classification of instance with different ensembles methods, in this implementation 68% of data can be classified in random forest and C 4.5 with Bagging Aggregation classified 74.2 % of instance. The classification of Tree View based on C4.5 with Bagging Aggregation algorithm.

Conclusion

From this research, it can be concluded that to evaluate an accurate method of machine learning technique by applying medical diagnosis data with two contributions. From the research, C4.5 with Bootstrap Aggregation gives high accurate classification rate rather than C4.5 pruned, C4.5 unpruned and random forest algorithm.

References

[1] Hall, Mark, et al. "The WEKA data mining software: an update." *ACM SIGKDD* Explorations Newsletter 11.1 (2009): 10-18.

[2] T.F. Gonzales. Clustering to minimize the maximum inter cluster distance. Theoretical Computer Science,1985,38(2-3):293-306.

[3] Kannan, M., S. Prabhakaran, and P. Ramachandran. "Rainfall forecasting using data mining technique."(2010)

[4] Arun K Pujari, 2003, “Data mining techniques”, University Press (India).

[5] Jiawei Han Micheline Kamber, 2006, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publisher an imprint of Elsevier

[6] L. Breiman, J. Friedman, R. Olshen and C. Stone. “Classification and Regression Trees”, Wadsworth International Group,
Belmont, CA, 1984.

[7]. Quinlan, J.R. (2003). “C5.0 Online Tutorial”, http://www.rulequest.com.

[8]. Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.,
Steinbach, M., Hand, D. J and Steinberg, D (2008). “Top 10 Algorithms in Data Mining”, Knowledge and Information Systems, 14
(1): 1-37.

[9]. Wolpert, D. (1992). “Stacked generalization”, Neural Networks, 5: 241-259.

[10]. Schapire, R. (1990). “The strength of weak learnability”, Machine Learning, 5(2): 197-227. [11]. Breiman, L. (1996a). “Bagging
Predictors”, Machine Learning, 24(2): 123-140.

[11]. Breiman, L (2001). "Random Forests". Machine Learning 45 (1): 5–32.

[12]. Freund, Y. Schapire, R. (1996). “Experiments with a new boosting algorithm”, In Proceedings of the Thirteenth International
Conference on Machine Learning, 148-156 Bari, Italy.

[13]. Dietterich, T. G. (2000). “An experimental comparison of three methods for constructing ensembles of decision trees: bagging,
boosting and randomization”. Machine learning, 40: 139-157.

[1]. Kotsiantis, S and Pintelas, P. (2004). “Local Boosting of Weak Classifiers”, Proceedings of Intelligent Systems Design and
Applications (ISDA 2004), August 26-28, Budapest, Hungary.

[15]. Opitz, D and Maclin, R (1999) "Popular Ensemble Methods: An Empirical Study", 11: 169-198.

[16]. Chap T. Le (1997). “Applied survival analysis”, Wiley, New York.

[17]. Quinlan, J. R. (1996) “Bagging, Boosting and C4.5”, AAAI/IAAI, 1: 725-730.

[18]. Endo, A, Shibata, T and Tanaka, H (2008) “Comparison of Seven Algorithms to Predict Breast Cancer Survival”, Biomedical Soft
Computing and Human Sciences, 13(2), pp.11-16.

[19]. Banfield, R.E, Hall, L.O, Bowyer, K.W and Kegeimeyer, W. P. “A comparison of decision tree ensemble creation techniques” (
(2007), IEEE Transactions on pattern analysis and machine intelligence, 29: 173-180.