ISSN: 0975-766X
CODEN: IJPTFI
Available Online through *Research Article*
www.ijptonline.com

PERFORMANCE ANALYSIS OF DATA MINING CLASSIFICATION TECHNIQUES FOR CHRONIC KIDNEY DISEASE

***R.Sujatha, **Dr.Ezhilmaran**
Research scholar, SITE, VIT University, Vellore - 632 014.
SAS, VIT University, Vellore - 632 014.
Email: r.sujatha@vit.ac.in

Received on 10-05-2016 *Accepted on 09-06-2016*

Abstract:

Data mining finds its application in many areas by its capability of extracting the interesting knowledge from the data set. In the health care department to predict the dreadful disease the output of data mining is highly useful. Particularly in the healthcare, the amounts and complexity of data involved are large. The decision making is highly improved by using data mining in the place of traditional method. The kidney disease cases are keeping increasing in the modern world due to life style and food habit. The work involves the performance analysis using the classification techniques namely Naïve Bayes, Multilayer Perceptron, AdaBoost, Decision Table and J48 over chronic kidney disease data set.

Keyword: Chronic Kidney Disease, Accuracy, Specificity, Sensitivity, Data mining, Classification

Introduction:

Data mining applications benefit several industries including health care industry. But certain limitations like integrating the data from several systems and settings are challenging, raw data possess missing values, require domain knowledge and support of management. [1] The different methods of handling the missing values are ignoring the row holding missing value, using global constant, using the mean and mode for the numerical and nominal values, using data mining algorithm for predicting the suitable value, filling values manually and using most possible values. The weightage of the missing value plays the role in deciding the missing value imputation method. In this work replacing missing value with mean for numeric and mode for nominal is applied. [2] Naïve Bayesian: From 1950s the extensive work is carried on in this. It is simple probabilistic classifier model in the machine learning that works based on applying Bayes theorem. It considers the features involved as the independent contributor to the probability. It is designed for supervised induction tasks. [3] Multilayer Perceptron: Rumelhart and McClelland in

1986 devised the multilayer perceptron. It is a feed forward ANN model that plots input data onto a set of appropriate outputs. It comprises of multiple layers of nodes in a directed graph, with strong inter connections. Uses supervised learning technique named backpropagation for the network training. [4] Adaboost: Yoav Freund and Robert Schapire proposed the ensemble learning method that utilises multiple learners in solving a problem. It possesses wide application and helps in accurate prediction with simplicity. [5] Decision Table:rule based classification model that is represented in tabular way for describing and analysing the decision. [6] J48: It is the simple C4.5 decision tree that helps in classification. Creation of binary tree and working over each instances in the data set and thus helps in the classification. [7] K-Fold Cross Validation: For comparing the predictive accuracy of classifiers the K-fold cross validation method is utilised in the random samples that minimize the associated bias in the data set. Otherwise called as rotation estimation where the entire dataset is split randomly into k subsets that are mutually exclusive and of almost equal in size. [4]

Methodology:

The analysing of the data set based on the attributes and its properties is always the interesting work. The value of each attribute helps in the determination of the various evaluation metrics. The taken data set used for the analysis is the chronic kidney disease from UCI data repository. As mentioned earlier the need of the hour work is to think about the health. The data set is rich with the vital things to be considered for classifying with nominal and numerical values for the attributes. The various attributes are age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, haemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema, anaemia and finally the class that tells about chronic kidney disease or not. It holds cases of 400 different age grouped people with the missing values.The work flow is illustrated in the

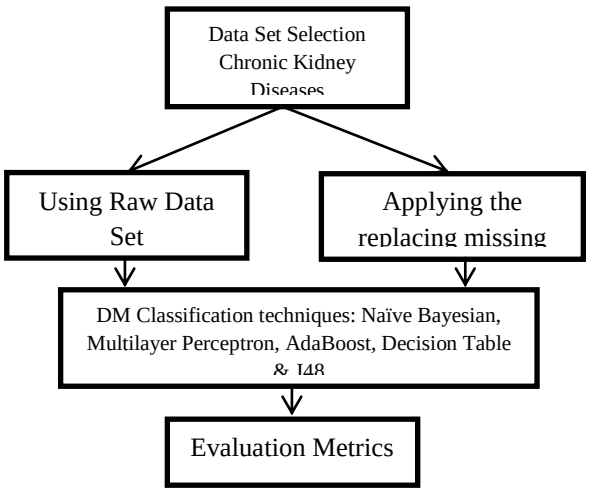

Fig 1. Work Flow

Figure 1.The classifiers selected for the analysis were done with care such that falls in different category of classification techniques namely naïve Bayesian from Bayes, Multilayer Perceptron from functions, Adaboost from meta, Decision table from rules and J48 from trees. The work is done on the K-fold cross validation varying from 2 to 6.

The performance of the classifiers with missing values and by the method replacing missing values is taken into consideration for folding from 2 to 6. The Table 1 narrates about the various evaluation metrics considered.

Table-1: Evaluation Metrics.

Accuracy - It is the measure that indicates the instances that are correctly classified. $$Accuracy = \frac{TruePositive + TrueNegative}{TotalNo.ofInstances}$$
Kappa Statistics - It is a statistic that measures inter-rater agreement for the purpose of qualitative matters. The value between 0 to 1 indicating the scale of perfect agreement. 1 is considered the perfect value. ( ) $$kappa = \frac{Observed\ Agreement - Expected\ Agreement}{1 - Expected\ Agreement}$$
TN - True Negative - Predicted negative that are actually negative
FP - False Positive - Predicted Positive that are actually negative
FN - False Negative - Predicted negative that are actually positive
TP - True Positive - Predicted positive that are actually positive
Specificity - It is a measure that tells about "True Negative Rate". $$Specificity = \frac{TrueNegative}{TrueNegative + FalsePositive}$$
Sensitivity / Recall / True Positive Rate - It is measure that tells about "True Positive Rate". $$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$
Precision - It is the proportion of the predicted positive cases that were correct. $$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$
F-Measure - It is the harmonic mean of precision and recall. $$F - Measure = 2.\frac{precision.recall}{precision + recall}$$
Area under ROC curve - The accuracy of the test relies on classifying the data set into with and without chronic kidney disease. >0.5 and 1 indicates the perfect.

Results &Discussions:

The chronic kidney disease data set possess 400 patience cases. Each instance has 24 attributes with last attribute holding the class that defines chronic disease or not. The dataset is taken from the UCI data repository.[8]

The Table 2 holds the above mentioned evaluation metrics value for the raw data set for the K- fold cross validation ranging from 2 to 6 over the various classifiers.

Table-2: Evaluation metrics over chronic kidney dataset with missing values.

K	Classifiers	Accuracy (%)	kappa statistics	TN	FP	FN	TP	time to build model (sec)	Specificty	Sensitivity	Precision	F-Measure	Area under ROC Curve
2	Naïve Bayes	95.25	0.9012	231	19	0	150	0	0.924	1.000	0.8876	0.953	1
	Multilayer Perceptron	98.5	0.9679	248	2	4	146	5.96	0.992	0.973	0.9865	0.985	0.999
	AdaBoost	98	0.9572	247	3	5	145	0.02	0.988	0.967	0.9797	0.98	0.997
	Decision Table	96.5	0.9247	246	4	10	140	0.14	0.984	0.933	0.9722	0.965	0.978
	J48	96.5	0.9253	243	7	7	143	0	0.972	0.953	0.9533	0.965	0.993
3	Naïve Bayes	94.75	0.8911	229	21	0	150	0	0.916	1.000	0.8772	0.948	1
	Multilayer Perceptron	97.75	0.9526	241	9	0	150	5.85	0.964	1.000	0.9434	0.978	1
	AdaBoost	97.25	0.9411	246	4	7	143	0.02	0.984	0.953	0.9728	0.972	0.998
	Decision Table	97.5	0.9464	247	3	7	143	0.16	0.988	0.953	0.9795	0.975	0.985
	J48	98.25	0.9629	244	6	1	149	0	0.976	0.993	0.9613	0.983	0.99
4	Naïve Bayes	95	0.8961	230	20	0	150	0	0.92	1.000	0.8824	0.951	1
	Multilayer Perceptron	99.75	0.9947	249	1	0	150	5.84	0.996	1.000	0.9934	0.998	1
	AdaBoost	98	0.9573	246	4	4	146	0.02	0.984	0.973	0.9733	0.98	0.998
	Decision Table	97.5	0.9467	245	5	5	145	0.14	0.98	0.967	0.9667	0.975	0.984
	J48	98	0.9569	250	0	8	142	0.02	1	0.947	1.0000	0.98	1
5	Naïve Bayes	94.75	0.8911	229	21	0	150	0	0.916	1.000	0.8772	0.948	1
	Multilayer Perceptron	99.25	0.9841	247	3	0	150	5.92	0.988	1.000	0.9804	0.993	1
	AdaBoost	98.5	0.9682	245	5	1	149	0.03	0.98	0.993	0.9675	0.985	1
	Decision Table	96.75	0.93	247	3	10	140	0.17	0.988	0.933	0.9790	0.967	0.982
	J48	99.25	0.984	249	1	2	148	0	0.996	0.987	0.9933	0.992	0.999
6	Naïve Bayes	95.5	0.9062	232	18	0	150	0	0.928	1.000	0.8929	0.955	1
	Multilayer Perceptron	99.75	0.9947	249	1	0	150	5.84	0.996	1.000	0.9934	0.998	1
	AdaBoost	99.5	0.9894	248	2	0	150	0.03	0.992	1.000	0.9868	0.995	1
	Decision Table	98.25	0.9625	248	2	5	145	0.14	0.992	0.967	0.9864	0.982	0.994
	J48	98.75	0.9733	248	2	3	147	0	0.992	0.980	0.9866	0.987	0.999

Similarly Table 3 holds for the data set that is replacing the missing values. Experimental work begins in the place of imputation method. The numeric values are replaced by mean and nominal values are replaced by mode.

Table-3: Evaluation metrics over chronic kidney dataset with replace missing values filter.

K	Classifiers	Accuracy (%)	kappa statistics	TN	FP	FN	TP	time to build model (sec)	Specificty	Sensitivity	Precision	F-Measure	Area under ROC Curve
2	Naïve Bayes	94.25	0.881	227	23	0	150	0	0.908	1.0000	0.8671	0.943	1
	Multilayer Perceptron	98	0.9578	242	8	1	150	10.58	0.968	0.9934	0.9494	0.98	0.998
	AdaBoost	96	0.9158	237	13	3	147	0.03	0.948	0.9800	0.9188	0.96	0.994
	Decision Table	97	0.9355	247	3	9	141	0.11	0.988	0.9400	0.9792	0.97	0.972
	J48	97	0.9358	245	5	7	143	0	0.98	0.9533	0.9662	0.97	0.97
3	Naïve Bayes	94	0.8756	227	23	1	149	0	0.908	0.9933	0.8663	0.941	0.998
	Multilayer Perceptron	97	0.937	238	12	0	150	10.99	0.952	1.0000	0.9259	0.97	0.999
	AdaBoost	95	0.8942	237	13	7	143	0.03	0.948	0.9533	0.9167	0.95	0.993
	Decision Table	96	0.9142	244	6	10	140	0.11	0.976	0.9333	0.9589	0.96	0.979
	J48	95.75	0.9097	240	10	7	143	0	0.96	0.9533	0.9346	0.958	0.972
4	Naïve Bayes	94.25	0.881	227	23	0	150	0	0.908	1.0000	0.8671	0.943	0.999
	Multilayer Perceptron	97.75	0.9526	241	9	0	150	10.86	0.964	1.0000	0.9434	0.978	0.999
	AdaBoost	97.5	0.9471	242	8	2	148	0.02	0.968	0.9867	0.9487	0.975	0.998
	Decision Table	97.25	0.9411	246	4	7	143	0.11	0.984	0.9533	0.9728	0.972	0.988
	J48	96.75	0.9304	245	5	8	142	0	0.98	0.9467	0.9660	0.967	0.971
5	Naïve Bayes	95.25	0.9012	231	19	0	150	0	0.924	1.0000	0.8876	0.953	0.999
	Multilayer Perceptron	97.75	0.9524	242	8	1	149	10.88	0.968	0.9933	0.9490	0.978	0.999
	AdaBoost	97.25	0.9417	242	8	3	147	0.03	0.968	0.9800	0.9484	0.973	0.998
	Decision Table	97.25	0.9408	248	2	9	141	0.11	0.992	0.9400	0.9860	0.972	0.977
	J48	96	0.9151	240	10	6	144	0	0.96	0.9600	0.9351	0.96	0.959
6	Naïve Bayes	93.25	0.8599	226	24	3	147	0.02	0.904	0.9800	0.8596	0.933	0.997
	Multilayer Perceptron	97.25	0.942	240	10	1	149	10.11	0.96	0.9933	0.9371	0.973	0.999
	AdaBoost	97.25	0.9414	244	6	5	145	0.02	0.976	0.9667	0.9603	0.973	0.998
	Decision Table	97.25	0.9409	247	3	8	142	0.11	0.988	0.9467	0.9793	0.972	0.992
	J48	96.5	0.9253	243	7	7	143	0.02	0.972	0.9533	0.9533	0.965	0.976

On comparing the various evaluation metrics from the above tables, the interpretation is quiet shocking that for this data set the missing value imputation method adopted yields the negative impact on the performance. In general ROC area is excellent.

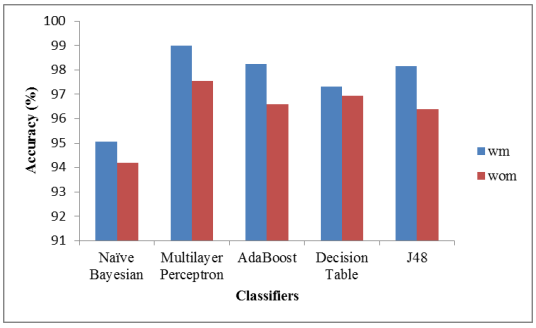

Fig 2. Accuracy

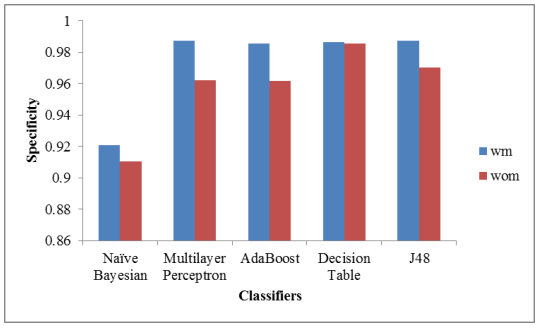

Fig 3. Specificity

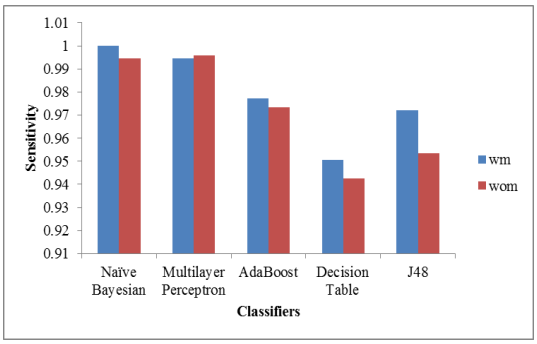

Fig 4. Sensitivity

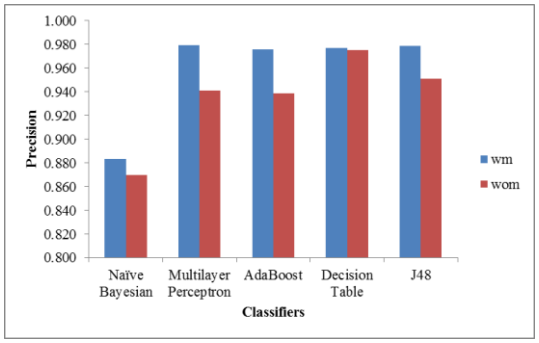

Fig 5. Precision

The various performance like accuracy, specificity, sensitivity and precision is dropped is indicated in the Fig. 2,3,4 and 5 on the comparison between the raw data set and processed data set ie ., data imputation with mean and mode make data sets without missing value.

Conclusion:

The work clearly illustrated that choosing the method for the missing data set handling should be done with the utmost care. When tried to use the ignoring method due to the lot of missing values the output is not appreciable. The values of the data set needs to be analysed well to get the highly interesting patterns and best performance values.

References:

1. Hian Chye Koh, Gerald Tan, Data mining applications in health care, Journal of Healthcare Information Management , 2005, Vol. 19, No. 2, pp. 64-72.
2. J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques: Morgan kaufmann, 2006.
3. Sau Loong Ang, Hong Choon Ong and Heng Chin Low, Classification Using the General Bayesian Network, Pertanika J. Sci. & Technol, 2016, 24 (1), pp. 205 - 211.
4. http://archive.ics.uci.edu/ml/datasets/Chronic_Kidney_Disease

5. Dursun Delen, Glenn Walker, Amit Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, Artificial Intelligence in Medicine ,2005, 34, pp. 113—127.
6. Xindong Wu et. al., Top 10 algorithms in data mining, Knowledge Information System, 2008, 14,pp.1–37.
7. Vikas Chaurasia, Saurabh Pal, Early Prediction of Heart Diseases Using Data Mining Techniques, Caribbean Journal of Science and Technology, 2013,Vol.1,pp. 208-217.
8. Tina R. Patil et. al., Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification, International Journal of Computer Science And Applications, April 2013, Vol. 6, No.2, pp. 256-261.

Corresponding Author:
R.Sujatha*,
Email: *r.sujatha@vit.ac.in*