MISSING DATA CLASSIFICATION OF CHRONIC KIDNEY DISEASE

Wala Abedalkhader and Noora Abdulrahman

Department of Engineering Systems and Management, Masdar Institute of Science and Technology, Abu Dhabi, United Arab Emirates

ABSTRACT

In this paper we propose an approach on chronic kidney disease classification with the presence of missing data. We implemented a classification system to solve the challenge of detecting chronic kidney diseases based on medical test data. The approach is comparing three different techniques that deals with missing data including deletion, mean imputation, and selection of best features. Each techniques is tested using the K-NN classifier, Naïve Bayes classifier, decision tree, and support vector machines (SVM). The final accuracy of each system is determined using 10-fold cross validation.

KEYWORDS

Classification, K-NN Classifier, Naïve Bayes, Data Mining, Decision Tree, Support Vector Machines & Python

1. INTRODUCTION

The threat of chronic kidney disease is increasing and to diagnose the disease from the growing volume of data is challenging. The medical examinations that are done to patients to discover the disease are enormous, which results in huge data collection. However, in the large number of records some important data are missing. Here comes a need for a classifier which produces good classification accuracy to solve the challenge of detecting chronic kidney diseases based on medical test data.

Nowadays data mining is a vital role in several applications such as manufacturing, aerospace, business organizations, government sectors, and medical industry. In the medical industry, the data mining is mostly used for disease prediction and diagnoses. Classification techniques are very useful for medical problems and have been used to diagnose many diseases over the past few decades [1]. This paper suggests a model of a chronic kidney diseases diagnosis system integrating data mining classification. The main process of the system consist of a data mining classification technique to detect the disease in case of missing data.

2. LITERATURE REVIEW

Based on studies, there are several data mining techniques that are applied to diagnose diseases including classification, clustering, association rules, summarizations, regression and etc. This section summarizes the methods that scholars used to detect medical problems.

In a recent study, Subasi et al. [2] researched different Machine Learning classifiers including random forest using both quantitative and qualitative discussions. The findings show that the random forest classifier accomplishes the near-optimal performances on the identification of chronic kidney disease subjects.

The most recent study is conducted by Sunil and Sowmya [3] to predict Chronic Kidney Disease using classification techniques such as Naïve Bayes and Artificial Neural Network. The results show that Naive Bayes technique has more accurate results than Artificial Neural Network.

DOI: 10.5121/ijdkp.2017.7604

Vijayarani and Dhayanand propos a model using classification process to classify four types of kidney diseases [4]. They compared Support Vector Machine (SVM) and Naïve Bayes classification algorithms based on the performance factors classification accuracy and execution time. They concluded that the SVM attains a better classification performance compared to Naïve Bayes classifier algorithm. However, Naïve Bayes classifier classifies the data with minimum execution time.

Vijayarani and Dhayanand also conducted another study to predict kidney diseases by using Support Vector Machine (SVM) and Artificial Neural Network (ANN) [5]. They compared the performance of these two algorithms on the basis of its accuracy and execution time. The experimental results shows that the performance of the ANN is better than the other algorithm. They also concluded that SVM classifier classifies the data with minimum execution time.
Bala and Kumar studied various data mining techniques to predict the Kidney disease and to compare them to find the best method of prediction [6]. Decision trees, ANN, Naive Bayes, and Logistic Regression, Genetic Algorithms are analyzed on Kidney disease data set. They found that decision tree, Naïve Bayes, and ANN are the well-performing algorithms used for Kidney disease. Depending on the situations some techniques perform better. However, there are situations when a combination of the best properties of some of the aforementioned DM techniques results more effective.

Shah et al. conducted a study to predict kidney disease, where they computed the average for the missing data [7]. The used decision tree classifier and tested it on aggregate data set and individual visit data set. They found that a higher accuracy value is obtained using the individual visit data set over the aggregate data set.

Huang et al. proposed a model that diagnose chronic diseases using data mining [8]. They used decision tree induction algorithm and the case association. They found that both algorithms are accurate to predict diseases.

Lakshmi et al. studied three data mining techniques to predict kidney disease including Artificial Neural Networks, Decision tree and Logical Regression [9]. The artificial neural networks was the best one among all three algorithms giving an accuracy of 93.8521%.

3. PROBLEM DEFINITION

Data Mining is used in this research to perform a classification in the presence of missing data. The dataset that is used is a set of chronic kidney disease from the UCI database. The data set consist of medical examination collected from 400 patients, where 250 of them have chronic kidney disease and 150 don't have it. These two groups are classified to chronic kidney disease class (ckd) and don't have chronic kidney disease class (notckd).

The data set includes 24 features which contains items like age, blood pressure, blood glucose, and so on. However, there are a large number of records for which certain feature values are missing in this data set. This research will review methods for dealing with these missing data, and to propose and implement a classifier to solve the challenge of detecting chronic kidney disease based on medical tests and to produces good classification accuracy based on cross-validation.

4. METHODOLOGY

The presence of missing data in a dataset Missing is a common problem in statistical analysis and it can affect the performance of the classifier modeled while using the dataset as a training sample. Missing data rates of less than 1% missing data are considered trivial, and 1-5% rate is manageable. However, having 5-15% rate requires sophisticated methods to handle, and more

than 15% rate may severely impact any kind of interpretation [10]. Many models have been proposed to deal with missing data and this research focuses on deletion, mean imputation, and best feature selection.

The deletion method consist of discarding all instances with missing values for at least one feature to avoid missing data. This method consists of determining and evaluating the extent of the missing data on each of the instances and attributes, then delete the instances or attributes that have high levels of missing data. However, before deleting any attribute, it is important to evaluate its relevance to the analysis.

Another method that is used in this research to deal with missing data in data mining is mean imputation. This method is the most frequently used method to deal with missing data [8]. In this research it is applied to numerical data in replacing the missing data of columns with the mean value of the available data. Missing data with the categorical data such as poor and good is replaced with the most frequent. Best feature selection technique is then applied to the data after performing the mean imputation. It consists of selecting the best column of features.

After performing the missing data techniques the experiments were carried out using four data mining classification techniques including K Nearest Neighbors, Naïve Bayes, Support Vector Machines, and Decision Tree.

KNN classifier classifies each test sample based on the majority label of the k-nearest neighbors, as determined from Euclidean distances to the test sample. In this model K is set to 5. Two functions of KNN classifier are modeled an evaluated including uniform and inverse KNN.
KNN method classifies each test sample based on the majority label of the k-nearest neighbors, the nearest points (neighbours) are determined from Euclidean distances to the missing value using the following formula.

$$Dist(A,B) = \sum (x_i - y_i)^2 / m \tag{1}$$

A small circle is created by the closest k (in this model K is set to 5) points to the missing value which is labelled with the majority of the known neighbour's label, the figure below illustrates the KNN classifier. Two functions of KNN classifier are modeled and evaluated including uniform (all points have the same weight regardless of the distance) and inverse KNN that is points are weighted according to the distance the closer the point the higher is its weight.

Naïve Bayes creates a Bayesian probability model with strong independence assumptions between features based on their frequency. This classifier is represented using three different distributions one is Gaussian distribution that follows:

$$p(x|c) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2}$$

Another distribution that is used to test Naïve Bayes classifier is Multinomial classifier that follows the following equation:

$$p(x|c_k) - \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i P_{ki}^{x_i} \tag{3}$$

The last distribution is Bernoulli Naïve Bayes that follows:

$$p(x|c_k) = \prod_{i=1}^{n} P_{ki}^{x_i} (1 - P_{ki})^{(1-x_i)} \tag{4}$$

SVM is another classification algorithm that is used in this research, which is non-probabilistic classifier. The SVM model is a representation of the training data as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. Test points are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

Decision tree algorithm uses a decision tree as a predictive model which maps observations about an item to conclusion about the item's target value. In this tree structures, class labels are represented with leaves while the branches represent conjunctions of features that lead to those class labels.

The accuracy is measured to evaluate each model and select the best model among all the four data mining algorithm used in this research. 10-fold cross validation was used to assess performance.

$$Accuracy(\%) = \frac{|correctly\ classified\ objects|}{|total\ number\ of\ objects|} \times 100$$

5. DISCUSSION AND RESULTS

We started first by exploring the data and to better understand the size of the missing data using python ("isnull().sum()") function and the results for the features as shown in Figure 1 were varying from 1 to 152 missing values however all the data had the class defined.

Even though some features have relatively much lower numbers of missing values but predicting solely on them might not necesseraliy be accurate depending on the correlations between the features and the classes.

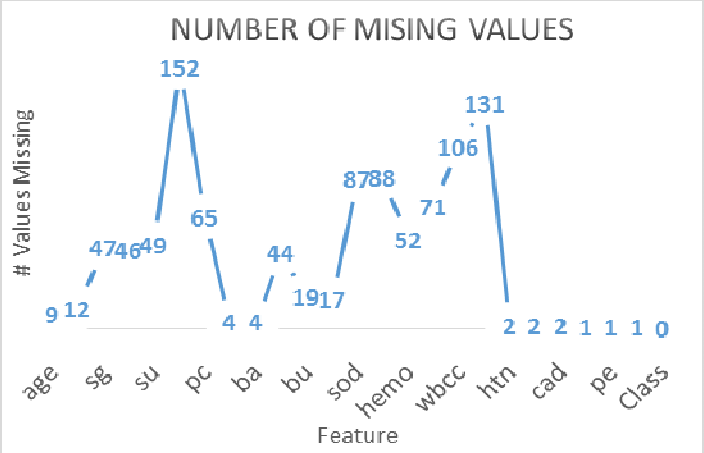

Figure 1. Number of missing values for each feature

Therefore, and as to find the best features when all data is available using the "sklearn.feature_selection" from the SKLEARN library, the following results were obtained in Figure 2.

[False False True True False True False False False False False False
False False False False False True True True True False True]
[15 11 1 1 5 1 4 9 6 16 13 14 12 10 2 8 17 3 1 1 1 1 7 1]

Figure 2. Results

Indicating that the best features that can be used to predict the classes are: “sg, al, hemo, htn, dm, cad, appet, ane”, which mostly has very few missing data and are mostly categorical.
10-fold Cross Validation was used to test the methods applied, and the accuracy was used to get the order of the methods and decide on the best method used.

Before testing the data all categorical values (rbc, pc, pcc, ba, htn, dm, cad, appet, pe, ane, and class) were replaced with binary values of (0,1) representing the categories each feature had, (“Yes/No”, “Present/NotPresent” , “Normal/Abnormal”, “Good/Poor”, “CKD/not CKD”) using the “fit_transform” function from the pandas library.

In dealing with missing data two main methods were used: deletion and mean imputation. Some of the most popular classification functions were used to correlate data from the Scikit-learn Python library, as shown in Table 1 Naive base (Bernoulli, Multinomial and Gaussian), K-NN (uniform and inverse), Decision tree and Support Vector Machines.

Deletion Also known as complete case analysis, is basically disregarding all cases with missing values for at least one feature, even though this have produced good accuracies that reached up to 100% using NB Gaussian, we could not consider it as practical nor truly representative of the data because of the relatively large number of missing values, looking back at Figure 1 the maximum missing values for one of the features (rbc) was 152 values which indicates that using deletion a minimum of 152 cases will be lost that is about 38% of the total data or more and that is a huge loss of data, also, taking into consideration that the data represents a medical problem “Kidney Diseases” which needs to be detected, failing to detect it due to some missing information cannot be tolerated.

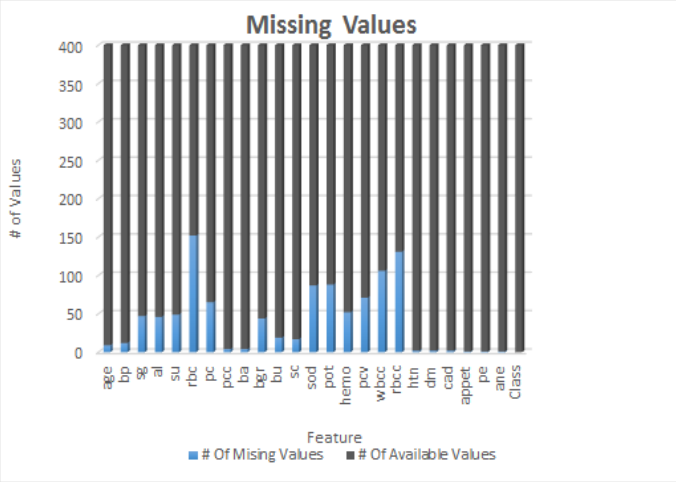

Figure 3. Missing values compared to total number of values for each feature

Therefore, we used the mean imputation to solve the missing data issue that is one of the most frequently used methods, for the numerical values the mean value replaced the missing values, while for the categorical values the most frequent category replaced the missing data.

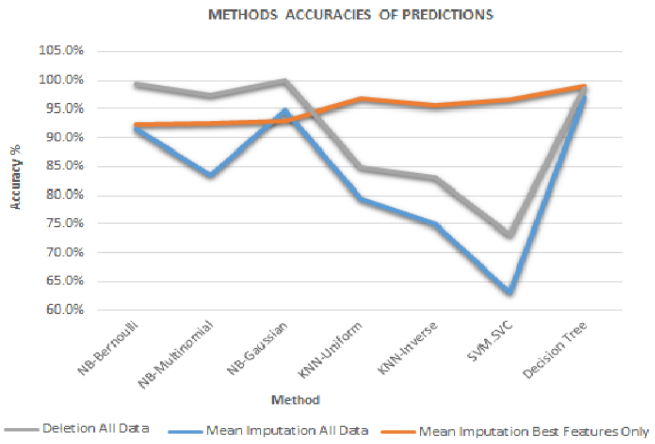

Figure 4. Accuracies of Predictions in Cross Validation for each method

After testing on data with replaced missing values using mean imputation we used the beast features only to predict the classes and that gave the best results reaching up to 99% using Decision tree.

Table 1. Accuracies of Predictions in Cross Validation for each method.

Method	**Mean Imputation**		**Deletion**
	All Features	**Best Features Only**	**All Features**
NB-Bernoulli	91.5%	92.4%	99.5%
NB-Multinomial	83.5%	92.6%	97.3%
NB-Gaussian	94.7%	93.0%	100.0%
KNN-Uniform	79.5%	96.8%	84.80%
KNN-Inverse	75.0%	95.5%	83.10%
SVM.SVC	63.0%	96.6%	73.20%
Decision Tree	96.9%	99.0%	98.60%

On average the SVM method resulted in the least accuracies followed by K-NN Uniform then K-NN inverse, Naïve Bayes had relatively better accuracies and the Decision Tree had the best accuracies.

During testing SVM had the longest execution time of few minutes which was much longer than the other methods that had very similar timings of about few seconds.

All methods except for Naïve Bayes had better accuracies with mean imputation, while naïve bayes had better values for Deletion. However, overall on average Mean Imputation-Best features only returned the best accuracies as shown in Figure 4 above.

Also Decision tree scored very high accuracies, among some of the advantages of Decision trees in classification is Decision trees implicitly do feature selection. Also Decision trees require relatively little effort for data preparation. Additionally, nonlinear relationships between

parameters do not affect tree performance and the visual aspect and ease of interpretation and understanding. We believe that the implicit feature selection and being not affected by nonlinearity were the main reasons for its good accuracies.

Using traditional methods like linear regression will not be as effective as KNN approach. Linear regression will give biased data and will not represent the data properly as KNN, as the study has many category variables and there are correlations between the variables.

6. CONCLUSIONS

In this paper, we applied two missing data handling techniques – Deletion and mean imputation. Then different data mining methods were applied KNN, Naïve Bayes, Decision Tree and Support Vector Machines to identify patterns and then to predict the classification of a Kidney disease presence or not based on the results of a medical examination. In the KNN approach we applied Uniform and inverse version of the technique while for the Naïve Bayes approach we used three different types; Bernoulli, Gaussian and Multinomial.

Best Features selection was applied on the mean imputed data and again all methods were tested as well. The results of the mean imputed data, best features only Decision Tree approach score was the best among the applied techniques based when the results are evaluated based on the accuracy where it scored 99%.

In handling missing data other techniques could be implemented such as Median Imputation, KNN Imputation and Regression Imputation. Also in classification of data other techniques could be applied such as Neural Networks (ANN, INN and FLNN).

REFERENCES

[1] A. A. Freitas, "A survey of evolutionary algorithms for data mining and knowledge discovery," in Advances in evolutionary computing, ed: Springer, 2003, pp. 819-845.

[2] Subasi, A., Alickovic, E. and Kevric, J. (2017). Diagnosis of Chronic Kidney Disease by Using Random Forest. *IFMBE Proceedings*, pp.589-594.

[3] Sunil, D., and B. P. Sowmya. "Chronic Kidney Disease Analysis using Data Mining." (2017).

[4] V. S and D. S, "Data Mining Classification Algorithms for Kidney Disease Prediction," International Journal on Cybernetics & Informatics, vol. 4, pp. 13-25, 2015.

[5] S. Vijayarani and M. S. Dhayanand, "KIDNEY DISEASE PREDICTION USING SVM AND ANN ALGORITHMS."

[6] S. Bala and K. Kumar, "A Literature Review on Kidney Disease Prediction using Data Mining Classification Technique," 2014.

[7] S. Shah, A. Kusiak, and B. Dixon, "Data Mining in Predicting Survival of Kidney Dialysis Patients-Invariant object approach," Proceedings of Photonics West-Bios, vol. 4949, 2003.

[8] M.-J. Huang, M.-Y. Chen, and S.-C. Lee, "Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis," Expert Systems with Applications, vol. 32, pp. 856-867, 2007.

[9] K. Lakshmi, Y. Nagesh, and M. VeeraKrishna, "Performance Comparison of Three Data Mining Techniques for Predicting Kidney Dialysis Survivability," International Journal of Advances in Engineering & Technology (IJAET), vol. 7, pp. 242-254, 2014.

[10] E. Acuna and C. Rodriguez, "The treatment of missing values and its effect on classifier accuracy," in Classification, Clustering, and Data Mining Applications, ed: Springer, 2004, pp. 639-647.