An Ensemble approach on Missing Value Handling in Hepatitis Disease Dataset

Sridevi Radhakrishnan
Research Scholar
Department of Computer Science
Karpagam University
Tamil Nadu, India

D. Shanmuga Priyaa, PhD
Professor
Department of Information Technology
Karpagam University
Tamil Nadu, India

ABSTRACT

The Major work in data pre-processing is handling Missing value imputation in Hepatitis Disease Diagnosis which is one of the primary stage in data mining. Many health datasets are typically imperfect. Just removing the cases from the original datasets can fetch added problems than elucidations. A appropriate technique for missing value imputation can assist to generate high-quality datasets for enhanced scrutinizing in clinical trials. This paper investigates the exploit of a machine learning technique as a missing value imputation process for incomplete Hepatitis data. Mean/mode imputation, ID3 algorithm imputation, decision tree imputation and proposed bootstrap aggregation based imputation are used as missing value imputation and the resultant datasets are classified using KNN. The experiment reveals that classifier performance is enhanced when the Bagging based imputation algorithm is used to foresee missing attribute values.

General Terms

Hepatitis, Disease , Diagnosis, Health, Clinical

Keywords

data mining, prediction, knn, imputation, missing values, bagging, bootstrap

1. INTRODUCTION

The liver is a lodge formed organ. It is positioned on the upper right side of the body underneath the rib cage. This organ, which is well thought-out to be the biggest, makes up 2% to 3% of the on the entire body’s collection. Contrasting the heart or stomach, the liver doesn’t have simply one function. According to hepatologists, this solitary organ has additional than 140 functions [1]. These consist of producing bile necessary for absorption, supporting up of minerals and vitamins, supporting in blood clotting, neutralizing toxic, creating amino acids to make physically powerful muscles, amendable energy, sustaining hormonal stability and dealing out sedatives [2]. When an individual befall affected with hepatitis virus, this virus can attack his liver and source swelling and redness in it [3]. This paper handles the hepatitis dataset to improve the data quality which is one of the primary processes in the prediction of the presence of disease.

1.1 Disease of Liver

The numerous disease situation can influence the liver. Some of the diseases are Wilson's disease, hepatitis, liver cancer, and cirrhosis. Alcohol amends the metabolism of the liver, which can encompass generally harmful effects if alcohol is in use more than protracted periods of time. Hemochromatosis can reason liver harms.

1.2 1.2 Risk Factors

Hepatitis is a soreness of the liver that can be caused by a virus, hereditary disorders, and occasionally by assuring medications or toxins such as alcohol and drugs. Scientists have recognized four major kind of viral hepatitis: hepatitis A, hepatitis B, and hepatitis C, and hepatitis D. A fifth type, hepatitis E, is generally not found in India.

1.3 Problem Description

Several researchers have recognized several significant and exigent problems [4-6] for medical assessment support. Numerous real-time clinical data sets are imperfect. The difficulty with missing feature values is an extremely significant problem in Data Mining. In health data mining the crisis with the missing values has turned into a tough issue. In many clinical tryouts, the medical testimony document agrees to some attributes to be gone blank, for the reason that they are unsuitable for some class of infirmity or the person provided that the information feels that it is not suitable to record the values of some attributes. The purpose of this work is to overcome the problem for missing values in hepatitis dataset. Life diagnosis of hepatitis is somewhat a difficult chore in untimely period due to an assortment of mutually dependent features. A model can be developed to meet the criteria the dataset of hepatitis, which can be used in forecasting of life prognosis of hepatitis disease.

1.4 Literature Review

In previous methods [7] which are used for prediction of hepatitis, wrapper method is used for feature selection in which attributes are removed based on trial and error method. Life Prognosis of hepatitis patients can be predicted by using classifier such as Support Vector Machine. In support vector machine, dataset is classified into training and testing data. Support vector machine analyze the training data and makes prediction on testing data. Predictive accuracy of classifiers can be enhanced by applying the techniques of feature selection. In this paper, Wrapper methods were incorporated to remove noise features before classification. After removal of noisy attributes, accuracy of the algorithm was further increased. SVM algorithm provides more and improved accuracy with the 10 attributes identified using a wrapper method using a data mining tool called weka. Data mining concepts and techniques [8] provide us the how to preprocess data and handle with missing values. Preprocessing is important because the data collected in real world is incomplete, noisy and inconsistent. Preprocessing stages include data cleaning, data integration, data transformation and data reduction. Data cleaning is carried out for missing values and Noisy data. Data integration combines data from multiple sources into a coherent data store. Inconsistencies in attribute may result in redundancies and these redundancies can be detected by correlation analysis. Data Transformation involves smoothing, aggregation, generalization, normalization and attributes construction. Data reduction includes attribute subset selection, dimensionality reduction and discretization. [9] Feature selection is more significant for data mining algorithms for a variety of reasons such as generalization performance, running time requirements and constraints. This method is based upon finding those features which minimize bounds on the leave-one-out error. Subsets of features are selected for preserving and improving the discriminating ability of a classifier. Feature selection for SVM is computationally feasible for high dimensional datasets. [10] Wrapper methods embed the model hypothesis search within the feature subset search. The feature subset selection conducts a search for good subset using an induction algorithm such as ID3 and decision tree as a part of evaluation function. The accuracy of the induced classifiers is estimated using accuracy estimation techniques. To achieve the best possible performance with a learning algorithm on a particular training set, a feature subset selection method considers the interaction between the algorithm and the training dataset.

According to the related work it is observed that only very few research work was carried out on the missing value handling. The major problem of misclassification is due to unqualified dataset. Hence this paper concentrates on the missing value handling in hepatitis dataset.

1.5 Materials and Methods

1.5.1 Reducing the DataSet

The simplest solution for the missing values imputation problem is the reduction of the data set and elimination of all missing values. This can be done by elimination of samples (rows) with missing values [11] or elimination of attributes (columns) with missing values [12]. Both approaches can be combined. Elimination of all samples is also known as complete case analysis. Elimination of all samples is possible only when large data sets are available, and missing values occur only in a small percentage of samples and when analysis of the complete examples will not lead to serious bias during the inference. Elimination of attributes with missing values during analysis is not possible solution if interested in making inferences about these attributes. Both approaches are wasteful procedures since they usually decrease the information content of the data

2. DATA DESCRIPTION

The data available at UCI machine learning data repository contains 19 fields with one output field [13]. The output shows whether patients with hepatitis are alive or dead. The intention of the dataset is to forecast the presence or absence of hepatitis virus given the results of various medical tests carried out on a patient. The Hepatitis dataset contains 155 samples belonging to two different target classes. There are 19 features, 13 binary and 6 features with 6–8 discrete values. Out of total 155 cases, the class variable contains 32 cases that died due to hepatitis

3. DATA PREPROCESSING

Dataset used in the prediction model should be more precise and accurate in order to improve the predictive accuracy of data mining algorithms. Dataset which is collected may have missing (or) irrelevant attributes. These are to be handled efficiently to obtain the optimal outcome from the data mining process

Table 1. Attributes in dataset

Attributes	Value	Missing values
Class	die (1), live (2)	0
Age	numerical value	0
Sex	male (1), female (2)	0
Steroid	no (1), yes (2)	1
Antivirals	no (1), yes (2)	0
Fatigue	no (1), yes (2)	1
Malaise	no (1), yes (2)	1
Anorexia	no (1), yes (2)	1
Liver Big	no (1), yes (2)	10
Liver Firm	no (1), yes (2)	11
Spleen Palpable	no (1), yes (2)	5
Spiders	no (1), yes (2)	5
Ascites	no (1), yes (2)	5
Varices	no (1), yes (2)	5
Bilirubin	0.39, 0.80, 1.20, 2.00, 3.00, 4.00	6
Alk Phosphate	33, 80, 120, 160, 200, 250	29
SGOT	13, 100, 200, 300, 400, 500	4
Albumin	2.1, 3.0, 3.8, 4.5, 5.0, 6.0	16
Protime	10, 20, 30, 40, 50, 60, 70, 80, 90	67
Histology	no (1), yes (2)	0

3.1 Data Cleaning

Dataset which is collected from UCI repository may have missing values and redundant attributes. Missing values can be handled either by removing the instances or replacing them by mean, average, maximum (or) minimum. Removing the instances may further reduce the amount of data, thereby reducing the quality in prediction. When the missing values are replaced by zero, it also affect the quality of data. In this paper an enhanced imputation technique is implemented for delivering a complete dataset.

3.2 Replacing missing value using MEAN AND MODE IMPUTATION

This is one of the most frequently used methods. It consists of replacing the unknown value for a given attribute by the mean(x) (quantitative attribute) or mode (qualitative attribute) of all known values of that attribute [14].

$$\bar{x} = \frac{1}{n}. \sum_{i=1}^{n} x_i$$

It replaces all missing records with a single and unique value¯, which is the mean value of that attribute.

4. PROPOSED MISSING VALUE IMPUTATION PROCESS

The original hepatitis data set is first portioned into groups. The records having missing values in their attributes are in one set and the records without any missing values are placed in a separate group. The knn classifier is trained with the complete data sets, and afterward the imperfect data is agreed to the bagging model for predicting the missing feature values. The scheme is recurrent for the whole set of attributes that have missing values. At the last part of training, this training dataset and absent value imputed datasets are combined to formulate the absolute data. The concluding dataset is then fed to the chosen k-nn classifier for classification.

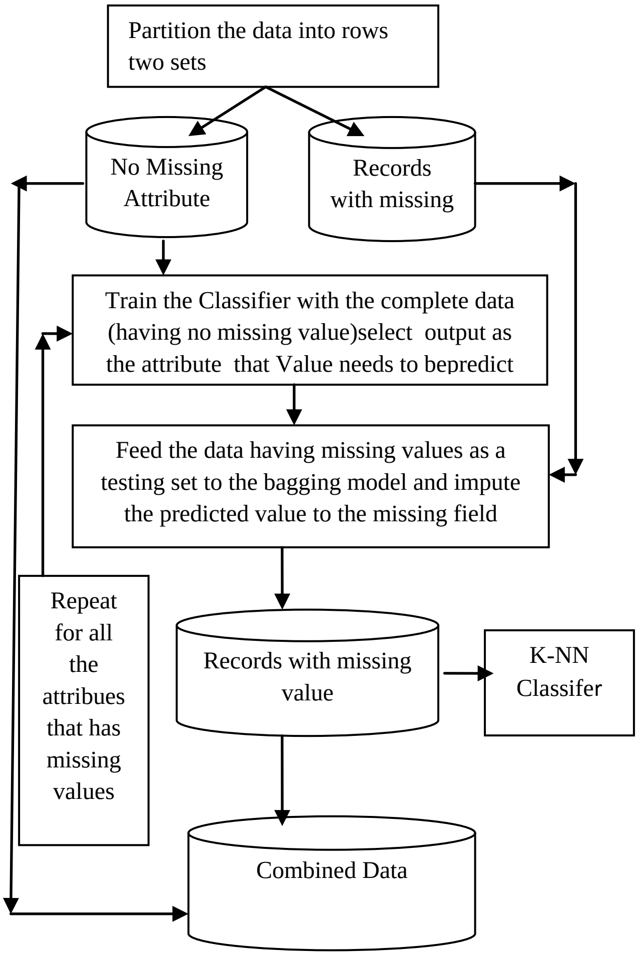
Partition the data into rows two sets
No Missing Attribute
Records with missing
Train the Classifier with the complete data (having no missing value)select output as the attribute that Value needs to bepredict
Feed the data having missing values as a testing set to the bagging model and impute the predicted value to the missing field
Repeat for all the attribues that has missing values
Records with missing value
K-NN Classifer
Combined Data

Figure 1: Working model of the proposed Work

4.1 Decision Tree

A decision tree is a tree-like graph or model. It is more like an inverted tree because it has its root at the top and it grows downwards. This representation of the data has the advantage compared with other approaches of being meaningful and easy to interpret. The goal is to create a classification model that predicts the value of a target attribute (often called class or label) based on several input attributes of the ExampleSet. In RapidMiner an attribute with label role is predicted by the Decision Tree operator. Each interior node of tree corresponds to one of the input attributes. The number of edges of a nominal interior node is equal to the number of possible values of the corresponding input attribute. Outgoing edges of numerical attributes are labeled with disjoint ranges. Each leaf node represents a value of the label attribute given the values of the input attributes represented by the path from the root to the leaf. This description can be easily understood by studying the attached Example Process.

Decision Trees are generated by recursive partitioning. Recursive partitioning means repeatedly splitting on the values of attributes. In every recursion the algorithm follows the following steps:

- An attribute A is selected to split on. Making a good choice of attributes to split on each stage is crucial to the generation of a useful tree. The attribute is selected depending upon a selection criterion which can be selected by the criterion parameter.
- Examples in the ExampleSet are sorted into subsets, one for each value of the attribute A in case of a nominal attribute. In case of numerical attributes, subsets are formed for disjoint ranges of attribute values.
- A tree is returned with one edge or branch for each subset. Each branch has a descendant subtree or a label value produced by applying the same algorithm recursively.

In general, the recursion stops when all the examples or instances have the same label value, i.e. the subset is pure. Or recursion may stop if most of the examples are of the same label value. This is a generalization of the first approach; with some error threshold. However, there are other halting conditions such as:

- There are less than a certain number of instances or examples in the current subtree. This can be adjusted by using the minimal size for split parameter.
- No attribute reaches a certain threshold. This can be adjusted by using the minimum gain parameter.
- The maximal depth is reached. This can be adjusted by using the maximal depth parameter.

Pruning is a technique in which leaf nodes that do not add to the discriminative power of the decision tree are removed. This is done to convert an over-specific or over-fitted tree to a more general form in order to enhance its predictive power on unseen datasets. Pre-pruning is a type of pruning performed parallel to the tree creation process. Post-pruning, on the other hand, is done after the tree creation process is complete.

4.2 ID3

ID3 is the meta learning algorithm used to build a decision tree [15]. It is the ancestor of C4.5 algorithm. Using fixed set of example, it classifies future samples. It has two types of nodes. Leaf node represents class name where as non leaf node represents decision node. The attribute test is conducted on decision node. This algorithm utilizes feature selection heuristic to help it decide which attribute goes into a decision node. The required heuristic can be selected by the criterion parameter.

The ID3 algorithm can be summarized as follows:

- Select all the unused attributes and calculate their selection criterion
- Pick the attribute for which the selection criterion has the best value
- Make node containing that attribute

The benefits of ID3 are generating comprehensible prediction rules, with less time short tree is created and the test data can be pruned .

4.3 Bootstrap Aggregation

In order to improve performance of existing models a meta learning algorithm Bootstrap aggregating (bagging) is created. It also decreases discrepancy and assists to avoid overfeeding. With the help of sub process it acts as a nested operator which enhances the learner provided in the sub process. An ensemble is itself a supervised learning algorithm, because it can be trained and then used to make predictions. The trained ensemble, therefore, represents a single hypothesis. This hypothesis, however, is not necessarily contained within the hypothesis space of the models from which it is built. Thus, ensembles can be shown to have more flexibility in the functions they can represent. In this proposed Work the Boot Strap Aggregation is used to ensemble the decision tree and the ID3 for imputing missing value in the hepatitis dataset [16].

4.4 Experimental Result

The performance analysis of missing value handling of hepatitis dataset was collected from UCI machine learning data repository. In this section the performance based on accuracy, precision and recall for five different approaches is shown. To analyze the performance of each approach K-nn is used as the classifier.

- Replacing missing value using Mean imputation
- Replacing missing value using ID3 Imputation
- Replacing Missing value using Decision Tree Imputation
- Replacing missing value using Bagging ID3 Imputation
- Replacing Missing value using Bagging Decision Tree Imputation

4.5 K-NN Classification Accuracy with Missing Value

Table 2. Accuracy: 79.79% +/- 10.88% (mikro: 80.00%)

	true 2	**true 1**	**class precision**
pred. 2	**109**	**17**	**86.51%**
pred. 1	14	15	51.72%
class recall	88.62%	46.88%	

Table 3. Replace missing values using Median :Accuracy: 79.25%

	true 2	**true 1**	**class precision**
pred. 2	109	18	85.83%
pred. 1	14	14	50.00%
class recall	88.62%	43.75%	

Table 4. Replacing missing value using ID3 Imputation:Accuracy: 81.88% +/- 7.18% (mikro: 81.94%)

	true 2	**true 1**	**class precision**
pred. 2	111	16	87.40%
pred. 1	12	16	57.14%
class recall	90.24%	50.00%	

Table 5. Replacing missing value using Decision Tree Imputation:Accuracy: 81.88% +/- 7.18% (mikro: 81.94%)

	true 2	**true 1**	**class precision**
pred. 2	111	16	87.40%
pred. 1	12	16	57.14%
class recall	90.24%	50.00%	

Table 6. Replacing missing value using Baging Decision Tree Imputation: Accuracy: 82.37% +/- 12.22% (mikro: 82.58%)

	true 2	**true 1**	**class precision**
pred. 2	112	16	87.50%
pred. 1	11	16	59.26%
class recall	91.06%	50.00%	

5. CONCLUSION AND FUTURE STUDIES

In this paper missing value in hepatitis dataset is investigated. This research work proposed a bootstrap aggregating based imputation approach for missing value handling to qualify the dataset of life prognosis of Hepatitis disease. The performance of this approach is analyzed by computing the classification accuracy of K-nearest neighbor with and without missing values. The result shows that before bagging process the ID3 performs better than the Decision Tree and Median Imputation. After the application of bagging in both ID3 and Decision Tree it's performance are enhanced to equal ratio.

6. REFERENCES

[1] WHO, Hepatitis C (Fact Sheet No. 164), World Health Organization, Geneva, 2000.

[2] WHO, Hepatitis C global prevalence (update), Weekly Epidemiological Record (World Health Organization), 74, 1999, pp. 421–428.

[3] Information regarding hepatitis C from the staff of Mayo Clinic; available at: http://www.mayoclinic.com/health/hepatitis-c/DS00097

[4] D. F. Sittig, A. Wright, J. A. Osheroff, B. Middleton, J. M. Teich, J. S. Ash, et al., "Grand challenges in clinical decision support," in J Biomed Inform. vol. 41, ed United States, 2008, pp. 387-92.

[5] J. Fox, D. Glasspool, V. Patkar, M. Austin, L. Black, M. South, et al., "Delivering clinical decision support services: there is nothing as practical as a good theory," in J Biomed Inform. vol. 43, ed United States, 2010, pp. 831-43.

[6] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: Current issues and guidelines," International Journal of Medical Informatics, vol. 77, pp. 81-97, Feb 2008.

[7] Roslina, A.H. and Noraziah, A "Prediction of Hepatitis Prognosis Using Support Vector Machine and Wrapper Method", Seventh International Conference on Fuzzy

Systems and knowledge Discovery (FSKD 2010), 978-1-4244-5934-6/10, 2010 IEEE.

[8] Jiawei Han and Micheline Kamber. "Data Mining: Concepts and Techniques",Data Preprocessing, Third Edition, 2011

[9] Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T. and Vapnik, V., " Feature Selection For SVMs", Advances in Neural Information processing Systems, MIT Press 2001, pg 668- 674.

[10] Ron Kohavai and George H. John., "Wrappers for feature subset selection" , Artificial Intelligence

[11] Kantardzic M. 2003: Data Mining – Concepts, Models, Methods, and Algorithms, IEEE, pp. 165-176.

[12] Lakshminarayan, K., Harp S. A. & Samad, T., 1999: Imputation of Missing Data in Industrial Databases, Applied Intelligence 11, pp. 259–275.

[13] Blake, C. L., & Merz, C. J. (1996). UCI repository of machine learning databases. Available from: <http://www.ics.uci.edu./ ~mlearn/MLReporsitory.html>.

[14] Liu Peng, Lei Lei , A Review of Missing Data Treatment Method

[15] http://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm

[16] http://docs.rapidminer.com/studio/operators/modeling/classification_and_regression/meta/bagging