ISSN:0975-9646

Estimation of Missing Values Using Decision Tree Approach

Gimpy[1], Dr. Rajan Vohra[2], Minakshi[3]

[1]*M.Tech. Scholar, P.D.M College of Engineering, Bahadurgarh, Haryana (India)*
[2]*H.O.D, Deptt. Of CSE, P .D.M College of Engineering, Bahadurgarh, Haryana (India)*
[3]*M.Tech. Scholar, P.D.M College of Engineering, Bahadurgarh, Haryana (India)*

***Abstract*: Data mining has made a great progress in recent year but the problem of missing data or value has remained great challenge for data mining. Missing data or value in a datasets can affect the performance of classifier which leads to difficulty of extracting useful information from datasets Dataset taken for this study is student records of university system that contains some missing values. The missing value are present in tm_10and tm_12. To estimate missing values classification algorithm (C4.5/J48) is used and then accuracy is measured through confusion matrix. Weka data mining tool is used for this analysis.**

Keywords: Data mining, Missing value, C4.5, Weka.

I. INTRODUCTION

Data mining refers to extracting knowledge from large amounts of data. The data may be spatial data, multimedia data, time series data, text data and web data. Data mining is the process of extraction of interesting, nontrivial, implicit, previously unknown and potentially useful patterns or knowledge from huge amounts of data. It is the set of activities used to find new, hidden or unexpected patterns in data or unusual patterns in data [1]. Data mining is the notion method and technique, which allow to analyses large dataset to extract discover previously unknown structure and relation out of large heap of detail these information is filtered, prepared and classified so that it will be a valuable aid for decision and strategies [2].

A. Missing values:

Missing data might occur because the value is not relevant to a particular case, could not be recorded when the data was collected, or is ignored by users because of privacy concerns. If attributes are missing in any training set, the system may either ignore this object totally; try to take it into account by, for instance, finding what is the missing attribute's most probable value, or use the value "missing", "unknown" or "NULL" as a separate value for the attribute. Missing values lead to the difficulty of extracting useful information from that data set [4]. Missing data are the absence of data items that hide some information that may be important [1].

B. Type of missing data: there is different type of missing value:

MCAR

The term "Missing Completely at Random" refers to data where the missingness mechanism does not depend on the variable of interest. Here the data are collected and observed arbitrarily and the collected data does not depend on any other variable of the dataset.[6].

MAR

It termed as "Missing at Random". We can consider an entry Xi as missing at random if the data meets the requirement that missingness should not depend on the value of Xi after controlling for another variable. [6].

NAMR

If the data is not missing at random or informatively missing then it is termed as "Not missing at Random". Such a situation occurs when the missingness mechanism depends on the actual value of missing data. Modeling such a condition is a very difficult task to achieve.This means we need to write a model for missing data and then integrate it into a more complex model for estimating missing values. [6].

C. Problem with missing data

Missing data cause a number of problems:

1. Analyses of data set with missing value are most problematic than analysis of complete data set.
2. Analyses may be inconsistent because analyst compensate for missing data or value in different way and their analysis may be based on different subset of data value "k" and the measure of similar will impact the result greatly [7]

D. Missing data imputation techniques:

1.LitwiseDeletion :

This method omits those cases (instances) with missing data and does analysis on the remains. It has two obvious disadvantages: a) A substantial decrease in the size of dataset available for the analysis. b) Data are not always missing completely at random. A variation of this method is to delete the cases (or attributes) with high missing rate. [5]

2.Mean/Mode Imputation (MMI) :

Replace a missing data with the mean (numeric attribute) or mode (nominal attribute) of all cases observed. To reduce the influence of exceptional data, median can also be used. [5]

3.K-Nearest Neighbor Imputation (KNN)

This method uses k-nearest neighbor algorithms to estimate and replace missing data. The main advantages of this method are that: a) it can estimate both qualitative attributes (the most frequent value among the k nearest neighbors) and quantitative attributes (the mean of the k nearest neighbors); b) It is not necessary to build a

predictive model for each attribute with missing data, even does not build visible models.[5]

4.Case Deletion (CD):

Also known as complete case analysis. This method consists of discarding all instances (cases) with missing values for at least one feature. A variation of this method consists of determining the extent of missing data oneach instance and attribute, and delete the instances and/or attributes with high levels of missing data random. [5]

5.Hot and Cold Deck Imputation:

A missing value is filled with a value from an estimated distribution for the missing value from the current data. Implemented in two stages-

Data are portioned into clusters.

Missing data are replaced within a cluster.

This can be calculated by mean or mode of the attribute within a cluster.

Hot Deck**:** A missing value of attribute is replaced by an observed value of the attribute chosen randomly.

Cold Deck: Imputation is similar to hot deck but the data source must be other then the current data source.

II. RESEARCH BACKGROUND

Ms. r. malarvizhi , in their paper "K-NN Classifier Performs Better Than K-Means Clustering in Missing Value Imputation" K-Means and KNN methods provide fast and accurate ways of estimating missing values.KNN – based imputations provides for a robust and sensitive approach to estimating missing data [7].

Edgar Acuna, Caroline Rodriguez, in their paper "The treatment of missing values and its effect in the classifier accuracy" The presence of missing values in a dataset can affect the performance of a classifier constructed using that dataset as a training sample [10]

B. Mehala, P. Ranjit Jeba Thangaiah, and K. Vivekananda, in their paper "Selecting Scalable Algorithms to Deal with Missing Values" This work analyses the behavior and efficiency for missing data treatment: C4.5 algorithm to treat missing data and K-means for missing data imputation. [11].

Xiaoyuan su "Using Imputation Techniques to Help Learn Accurate Classifiers" The accuracy of classifiers produced by machine learning algorithms generally deteriorates if the training data is incomplete, and preprocessing this data using simple imputation methods, such as mean imputation (MEI), does not generally produce much better classifiers. [12].

Maytal Saar-Tsechansky **"**Handling Missing Values when Applying Classification Models**"** This paper first compares several different methods—predictive value imputation, the distribution- based imputation used by C4.5, and using reduced models—for applying classification trees to instances with missing values[13]

Meghali A. KalyankarProf. S. J. Alaspurka "data Mining Technique to Analyse the Metrological Data" Meteorological data mining is a form of data mining concerned with finding hidden patterns inside largely available meteorological data, so that the information retrieved can be transformed into usable knowledge.. [14]

Bhavik Doshi, "Handling Missing Values" in Data Mining Missing Values and its problems are very common in the data cleaning process. Several methods have been proposed so as to process missing data in datasets and avoid problems caused by it. [4].

III. CONCEPTUAL FRAME WORK

PROBLEM STATEMENT

In this work, the student dataset is taken that contain number of attributes such as sate of domicile family income, 10th and 12 marks, category. In these records, some of data values are missing. To handle these missing values, Classification Algorithm (C4.5/J48) is used.

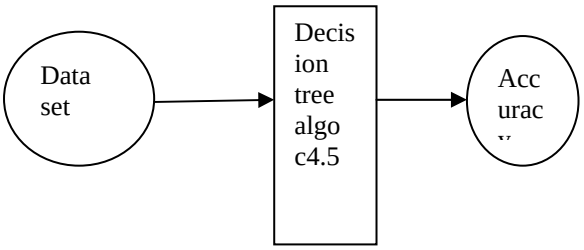

Fig1: Apply Decision Tree Algo

Dataset with missing values is taken and replace missing value technique is applied. Then on this complete dataset C4.5 algorithm is applied and finally accuracy is calculated while comparing the missing data and complete dataset.

CLASSIFICATION ALGORITHM

Classification is a supervised learning method. It means that learning of classifier is supervised in that it is told to which class each training tuples belongs. Data classification is a two step process. In the first step, a classifier is build describing a predetermined set of data classes or concepts [20].

J48 Decision Tree Classifier(Implementation of C4.5 Classifier)

C4.5 is an algorithm used to generate a decision tree. The task of constructing a tree from the training set has been called tree induction or tree building. Most existing tree induction systems adopt a greedy (non-backtracking) top-down divide and conquer manner. Starting with an empty tree and the entire training set, following algorithm is applied on the training data until no more splits are possible. The following algorithm is applied on the training data (where each tuple is associated with a class label) until no more splits are possible [3].

Algorithm:

1) Create a node N.
2) If all the tuples in the partition are of the same class then return N as a leaf node labeled with that class.
3) If attributes list is empty then return N as a leaf node labeled with the most common class in samples.
4) identify the splitting attribute so that resulting partitions at each branch are as pure as possible.
5) Label node N with splitting criterion which serves as test at that node.

6) If splitting attribute is discrete valued then remove splitting attribute from attribute list.
7) Let *Pi* be the partitions created based on the *i* outcomes on splitting criterion.
8) If any *Pi* is empty then attach a leaf with the majority class in the partition to node N.
9) Else recursively apply the complete process on each partition.
10) Return N [23].

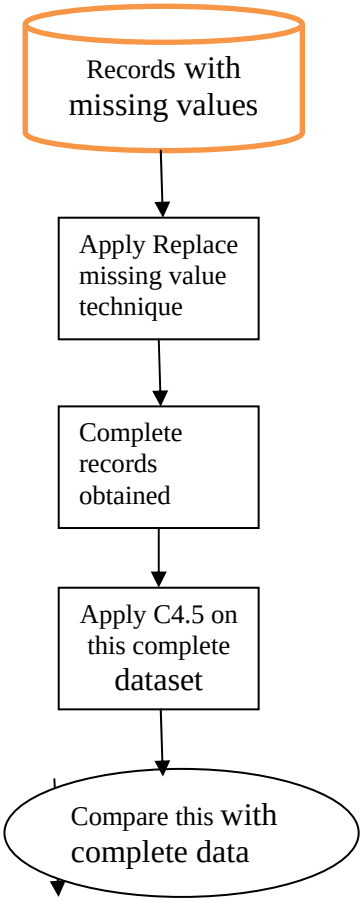

Fig 2: Flow Diagram of methodology for Missing Data Imputation

WEKA

The Weka or woodhen (Gallirallus australis) is an endemic bird of New Zealand. The WEKA project aims to provide a comprehensive collection of machine learning algorithms and data preprocessing tools to researchers and practitioners alike. It allows users to quickly try out and compare different machine learning methods on new data sets. Its modular, extensible architecture allows sophisticated data mining processes to be built up from the wide collection of base learning algorithms and tools provided. Extending the toolkit is easy thanks to a simple API, plug-in mechanisms and facilities that automate the integration of new learning algorithms with WEKA's graphical user interfaces.

IV. RESULT AND ANALYSIS

The database that is taken for this research work contains student records .These are 200 record of student having 13 attributes. Some of the data values in these records are missing. The attributes that contain missing values are marks obtained in 10th and 12th class. This database is designed in MS excel format.

First the file is saved and then converted into .csv format. After converting into .csv format the data set is loaded into weka tool. The main weka explorer interface with the data file is loaded using preprocessing panel which is shown as:

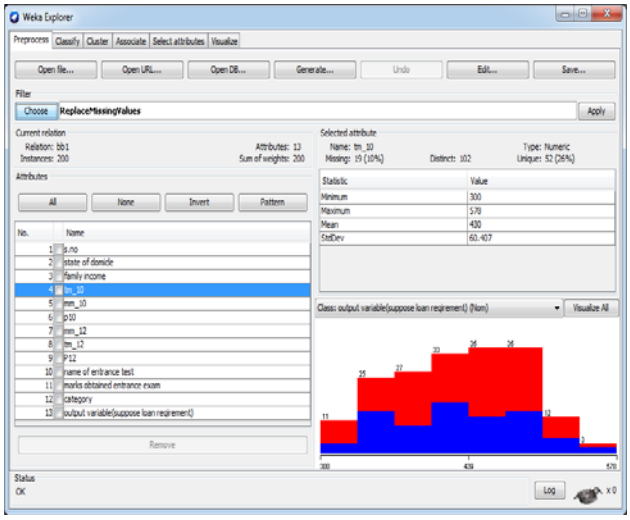

Fig3: Dataset having missing value loaded into weka tool

Then select the "classification" and click on choose button. Now choose the j48 classifier and in the classification mode panel use cross validation option is selected and then click on start. The resulting window is as follow :

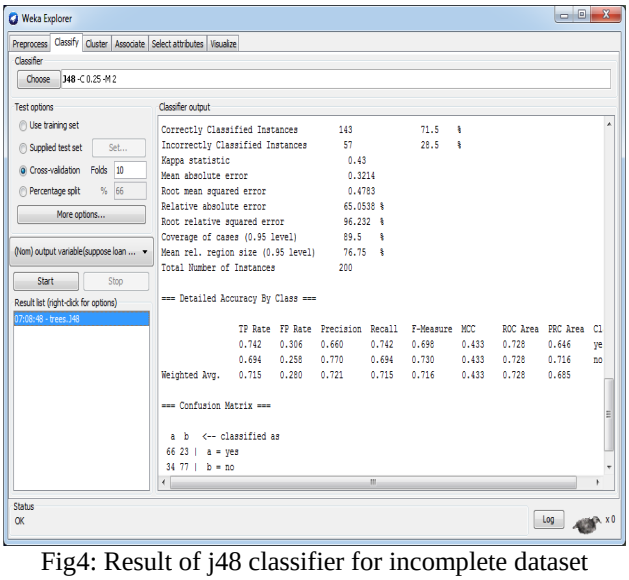

Fig4: Result of j48 classifier for incomplete dataset

A confusion matrix comes with attributes a and b. Now there are some attributes missing from dataset. Then load this dataset into weka tool which has missing value applying replace missing value technique:

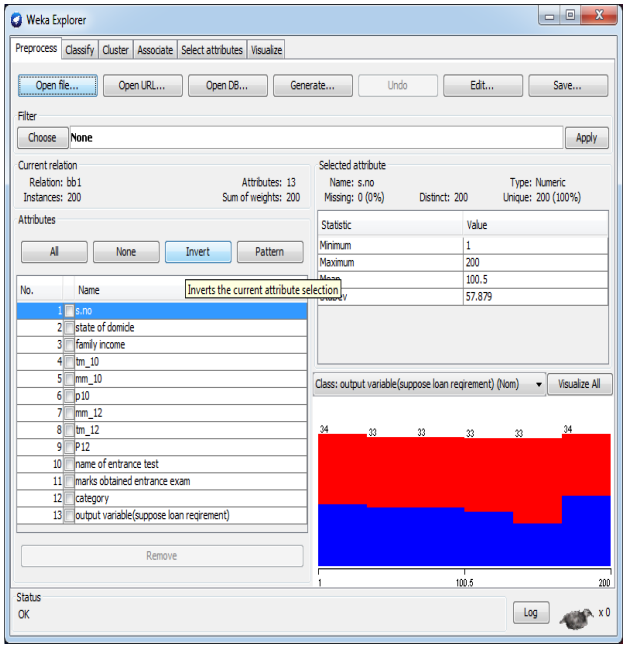

Fig5: Applying Replace Missing value technique on dataset

And then again J48 classifier technique is on the imputed data set. The result will be as shown below :

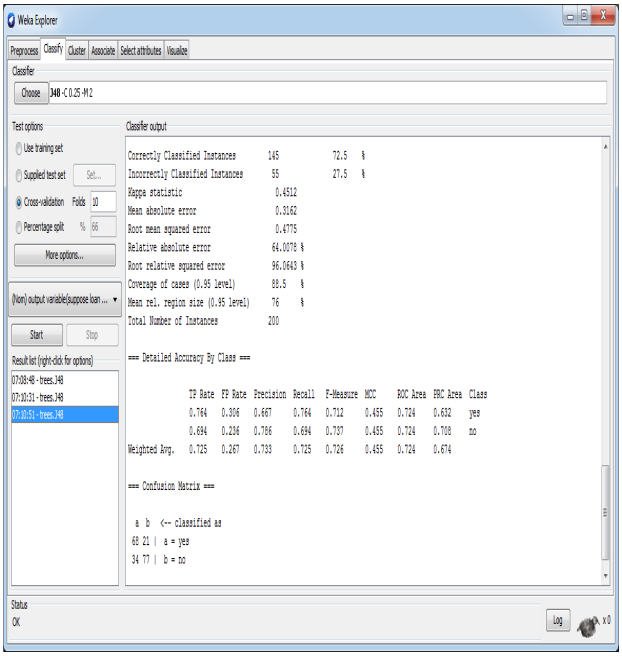

Fig6: Result of j48 classifier on imputed dataset

V. EXPERIMENTAL RESULT

The student data set having 13 attribute and 200 instances. The attribute that having missing value are tm_10and 1m_12. Classification is evaluated by using confusion matrix.

Confusion matrix

A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier [24].

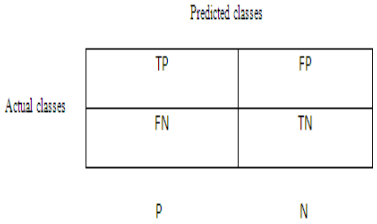

Table1: Confusion Matrix

- True positive (TP)- These are the positive tuples that were correctly labeled by the classifier [20].If the outcome from a prediction is p and the actual value is also p, then it is called a true positive (TP)[24].
- True Negative (TN)-These are the negative tuples that were correctly labeled by the classifier [20].
- False Positive (FP)-These are the negative tuples that were incorrectly labeled as positive [20]. However if the actual value is n then it is said to be a false positive (FP) [20].
- False Negative (FN)-These are the positive tuples that were mislabeled as negative [20].
- **Accuracy is calculated as** = (TP+TN)/(P+N) where, P=TP+FN and N=FP+TN. Or TP+TN/(TOTAL).

Algorithm used C 4.5 classifier	Correctly classified instances	Incorrectly classified instances	Accuracy	Mean absolute error
Missing data	143	57	71.5%	0.324
Imputed data	145	55	72.5%	0.316

Table2: Comparison based on J48/C4.5 Classification Algorithm

According to experimental results, correctly classified instances for missing value data or incomplete data is 143 and for imputed dataset is 145 which is 72.5% and is greater than previous one. Accuracy of C4.5 for imputed dataset is greater than missing dataset.

VI. CONCLUSION

Missing values are regarded as serious problem in most of the information system due to unavailability of data and must be impute before the dataset is used. Here student dataset is taken in which some of the values are missing. The classification algorithm is used and accuracy is calculated for both incomplete data and the imputed data. And as a result accuracy is greater for imputed dataset as compared to incomplete dataset.

Future scope: This work handles missing values only for numerical attributes. Further it can be extended to handle a categorical attribute. Different classification algorithm can be used for comparative analysis of missing data techniques. Missing data technique can be implemented in mat lab. The automation algorithm for retrieval of missing information in any domain is a very significant area for further research work.

ACKNOWLEDGEMENTS

Author would like to thanks to her head Dr. Rajan Vohra, Head Of Department of CSE & I.T department, PDMCE, Bahadurgarh for their valuable support and help.

REFERENCES

[1] Dinesh J. Prajapati ,Jagruti H. Prajapat, "Handling Missing Values: Application to University Data Set" .Issue 1, Vol.1 (August-2011), ISSN 2249-6149.

[2] Johannes Gambier, Andreas Rudolph," Techniques of Cluster Algorithms in Data Mining. Data Mining and Knowledge Discovery", 303–360, 2002.

[3] Osmar R. Zaïane, "Chapter I: Introduction to Data Mining CMPUT 690 Principles of Knowledge Discovery in Databases", 1999.

[4] Luai Al Shalabi, Mohannad Najjar and Ahmad Al Kayed, A framework to Deal with Missing Data in Data Sets . Journal of Computer Science 2 (9): 740-745, 2006 ISSN 1549-363.

[5] Alireza Farhangfar*a* , Lukasz Kurgan*b* , Witold Pedrycz "Experimental analysis of methods for imputation of missing values in databases

[6] Bhavik Doshi, Handling Missing Values in Data Mining. Data Cleaning and Preparation Term Paper.

[7] Liu Peng, Lei Lei , A Review of Missing Data Treatment Methods

[8] Ms.R.Malarvizhi, Dr.Antony Selvadoss Thanaman, "K-NN Classifier Performs Better Than K-Means Clustering in Missing Value Imputation", IOSR Journal of Computer Engineering

[9] Edgar Acuna, Caroline Rodriguez, "The treatment of missing values and its effect in the classifier accuracy".

[10] B. Mehala1 P. Ranjit Jeba Thangaiah2, and K. Vivekanandan," Selecting Scalable Algorithms to Deal With Missing Values", International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009

[11] Xiaoyuan Su," Using Imputation Techniques to Help Learn Accurate Classifiers"

[12] Maytal Saar-Tsechansky, "Handling Missing Values when Applying Classification Models ", journal of Machine Learning Research 8 (2007) 1625-1657

[13] Meghali A. KalyankarProf. S. J. Alaspurka, "data Mining Technique to Analyse the Metrological Data"

[14] M. Sandhya, Dr. A. Kangaiammal, Dr. C. Senthamarai "A Comparative Study on Decision Rule Induction for incomplete data using Rough Set and Random Tree Approaches" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 9, Issue 3 (Mar. - Apr. 2013), PP 06-10

[15] Santosh Dane, Dr. R. C. Thool "Imputation Method for Missing Value Estimation of Mixed-Attribute Data Sets" ,International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 5, May 2013

[16] .Sapna Jain 2.M Afshar Aalam3. M. N Doja,"K-MEANS CLUSTERING USING WEKA INTERFACE", Proceedings of the 4th National Conference; INDIACom-2010 Computing For Nation Development, February 25 – 26, 2010

[17] Rohanizadeh.s "A proposed data mining methodogy application to industrial procedures"

[18] S.Vijiyarani, S. Sudha," Disease Prediction in Data Mining Technique – Survey", International Journal of Computer Applications & Information Technology Vol. II, Issue I, January 2013

[19] Neelamadhab Padhy, Dr. Pragnyaban Mishra, and Rasmita Panigrahi, "The Survey of Data Mining Applications And Feature Scope" International Journal.

[20] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining Concepts and Techniques" Third edition .

[21] Sharad Verma, Nikita Jain," Implementation of ID3 – Decision Tree Algorithm".

[22] Anjana Gosain, AmitKumar," Analysis of Health Care Data Using Different Data Mining Techniques". lAMA 2009.

[23] R. R. Kabra R. S. Bichkar," Performance Prediction of Engineering Students using "Decision Trees", International Journal of Computer Applications (0975 – 8887) Volume 36– No.11, December 2011

[24] Anshul Goyal and Rajni Mehta,"Performance Comparison of Naïve Bayes and J48 Classification Algorithms", international Journal of Applied Engineering Research, ISSN 0973-4562 Vol.7 No.11 (2012)