Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.6, No.7, 2015

A Novel Approach for Imputation of Missing Value Analysis using Canopy K-means Clustering

Ramaraj.M[1], Dr.Antony Selvadoss Thanamani[2]
1, Ph.D Research Scholar Department of Computer Science NGM College Pollachi India
2, Associate Professor & Head Department of Computer Science NGM College Pollachi India

Abstract
Multiple imputation provides a useful strategy for dealing with data sets with missing value. Instead of filling in a single value for each missing value. Imputation is a term that denotes a procedure that replaces the missing values in a data set by some possible values. In this work missing values are being inserted completely at random (MCAR). Dataset taken for this work is heart disease dataset that contains some missing values. The main problem of this k means algorithm to don't take the random position of the data point, it's one of the k means for only two way cluster. These multiply imputed data sets are then analyzed by using standard procedures for complete data and combining the results from these analyses. In the proposed method is used for the canopy k-means clustering algorithm using with the predication of higher accuracy calculate the particular data sets.
Keywords: missing value, multiple imputation, k-means algorithm and canopy clustering.

INTRODUCTION

Missing data is one of the problems which are to be solved for real-time application. Traditional and Modern Methods are there for solving this problem [10]. The variables may be of Missing Completely at Random, Missing at Random, Missing not at Random. Each variable should be treated separately. A study on single imputation techniques such as Mean, Median, and Standard Deviation combined with canopy k means algorithm. Training set with their corresponding class groups the data of different sizes [9]. The above techniques are applied in each group and the results are compared.

The interest in dealing with missing values has continued with the statistical applications to new areas such as Data Mining and any datasets [5]. These applications include supervised classification as well as unsupervised classification (clustering). In datasets some people even replace missing values by zero.

MISSING VALUES

Because the value of the missing data the data was collected. When it was ignored because of the privacy concerns of users may not be able to record in a particular case is not relevant [8]. Useful information that can lead to difficulty in obtaining a set of data values. Missing data for some information that may be important is the lack of data hiding.

Methods of Missing Value

There are three methods as following
1. MCAR
In a data set are **missing completely at random (MCAR)** if the events that lead to any particular data-item being missing are independent both of observable variables and of unobservable parameters of interest, and occur entirely at random [2]. When data are MCAR, the analyses performed on the data are unbiased; however, data are rarely MCAR.
2. MCR
Missing at random (MAR) is an alternative, and occurs when the missing ness is related to a particular variable, but it is not related to the value of the variable that has missing data [1].
3. MNAR
Missing not at random (MNAR) is data that is missing for a specific reason (i.e. the value of the variable that's missing is related to the reason it's missing).

IMPUTATION TECHNIQUES

Imputations capture most of the relative efficiency that could be captured with a larger number of imputations [7]. However, a too-small number of imputations can lead to a substantial loss of statistical power, and some scholars now recommend 20 to 100 or more. Any multiply-imputed data analysis must be repeated for each of the imputed data sets and, in some cases, the relevant statistics must be combined in a relatively complicated way [9].
Imputation methods are
1. Partial imputation
2. Partial deletion

3. Full analysis
4. Interpolation

METHODOLOGY

K-means Clustering Algorithm

K means algorithm is one of the most common clustering algorithm in use as our prototype for development of a soft containment of the clustering algorithm. The k-means algorithm iteratively searches for a good division of n objects into k clusters [4]. It seeks to minimize the total variance V of a partition, i.e., the sum of the (squared) distances from each item d to its assigned cluster C.

Canopy k means clustering

The **canopy clustering algorithm** is an unsupervised pre-clustering algorithm introduced by Andrew McCallum, Kamal Nigam and Lyle Ungar in 2000. It is often used as preprocessing step for the K-means algorithm or the Hierarchical clustering algorithm. It is intended to speed up clustering operations on large data sets, where using another algorithm directly may be impractical due to the size of the data set.

The algorithm proceeds as follows, using two thresholds T_1 (the loose distance) and T_2 (the tight distance), where $T_1 > T_2$.[1][2]

1. Begin with the set of data points to be clustered.
2. Remove a point from the set, beginning a new 'canopy'.
3. For each point left in the set, assign it to the new canopy if the distance less than the loose distance T_1.
4. If the distance of the point is additionally less than the tight distance T_2, remove it from the original set.
5. Repeat from step 2 until there are no more data points in the set to cluster.
6. These relatively cheaply clustered canopies can be sub-clustered using a more expensive but accurate algorithm.

RESULT AND DISCUSSION

Algorithm for canopy k means clustering

```
Algorithm LargestNumber
 Input: A list of numbers L.
 Output: The largest number in the list L.
 if L.size = 0 return null
 largest ← L[0]
 for each item in L, do
   if item > largest, then
     largest ← item
 return largest
```

Attribute	cluster_0	cluster_1
sex	0.156	0.377
pain types	3.896	2.208
bp	6.141	4.730
clostrol	73.548	50.465
bs	0.170	0.157
ecpr	0.193	0.239
thalach	30.215	19.277
exang	0.585	0.075
oldpeak	1.076	0.170
result	0.659	0.107

Table 1: it represent the clustering groups
In this table has been describes the overall mean value of dataset and cluster dataset.

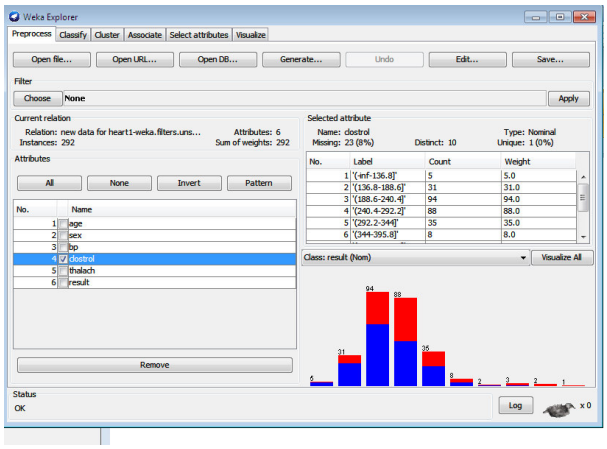


Fig 1: accuracy chart for k-means clustering.

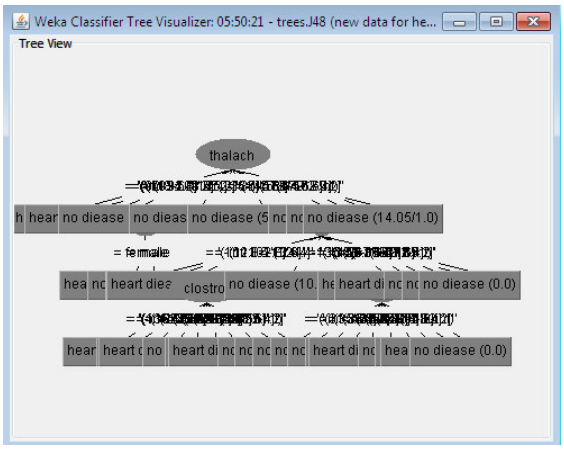
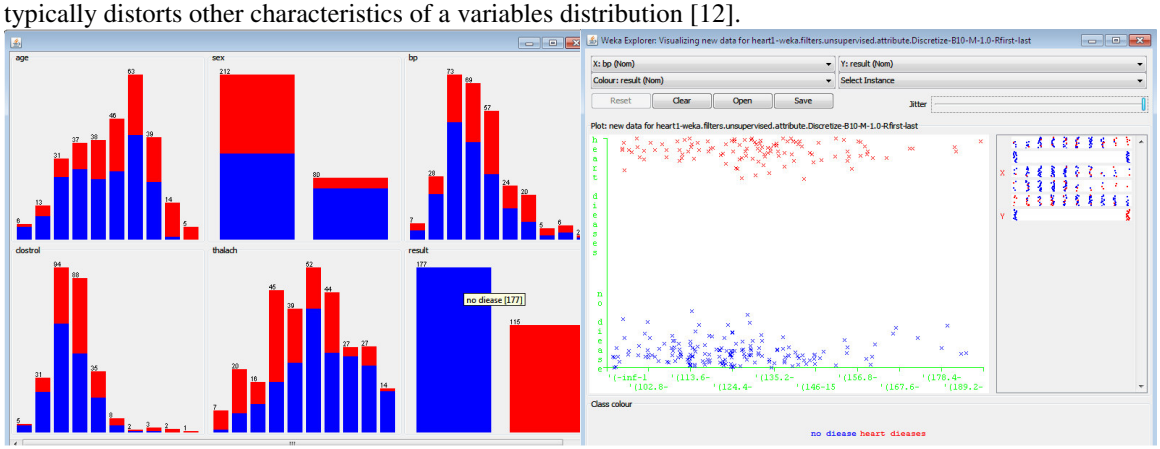


Fig 2: it represent the classification tree viewer of the missing data values

Mean imputation

The most commonly practiced approach is mean substitution— single imputation techniques. Mean substitution replaces missing values on a variable with the mean value of the observed values [11]. The imputed missing values are contingent upon one and only one variable – the between subjects mean for that variable based on the available data. Mean substitution preserves the mean of a variables distribution; however, mean substitution typically distorts other characteristics of a variables distribution [12].



Median Substitution

Mean or median substitution of covariates and outcome variables is still frequently used. This method is slightly improved by first stratifying the data into subgroups and using the subgroup average [13]. Median imputation results in the median of the entire data set being the same as it would be with case deletion, but the variability between individuals' responses is decreased, biasing variances and covariance toward zero.

Standard Deviation

The standard deviation measures the spread of the data about the mean value. It is useful in comparing sets of data which may have the same mean but a different range [14].

Sno	No of instance	Classification accuracy
K means	292	68%
EM clustering	292	68%
MDB clustering	292	69%
Canopy clustering	292	80%

Table 2: it represent the classification accuracy of the each clustering methods.

In this table 2 is represent the classification accuracy of the each cluerting method is descriped with them. When the data set take the 292 instance use and implementing the clustering method is k means clustering and EM clustering accuracy is same for them as 68% and MDB clustering accuracy is 69% and finally to implement the canopy clustering to distribute the higher and better accuracy and expect them.

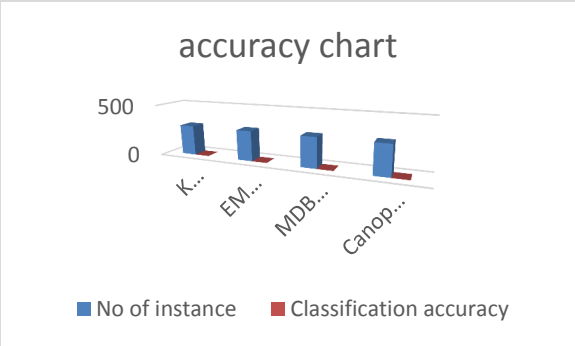


Fig 3: it represent the classification accuracy chart.

CONCLUSION

In this paper described the imputation of missing value using canopy clustering. Canopy clustering algorithm is one of the recent and fast growning on the unsupervised learning for grouping up of data. Traditional methods such as mean median and standard deviation is used to improve the preformance of accuracy in the missing data imputation.

REFERRENCE

[1]. Allison, P.D-—Missing Data‖, Thousand Oaks, CA: Sage -2001.
[2]. Bennett, D.A. —How can I deal with missing data in my study? Australian and New Zealand Journal of Public Health‖, 25, pp.464 – 469, 2001.
[3]. Graham, J.W. —Adding missing-data-relevant variables to FIML- based structural equation models. Structural Equation Modeling‖, 10, pp.80 – 100, 2003.
[4]. Graham, J.W, —Missing Data Analysis: Making it work in the real world. Annual Review of Psychology‖, 60, 549 – 576 , 2009.
[5]. Gabriel L.Schlomer, Sheri Bauman, and Noel A. Card : — Best Practices for Missing Data Management in Counseling Psychology‖ , Journal of Counseling Psychology 2010, Vol.57.No 1,1 – 10.
[6]. Jeffrey C.Wayman , —Multiple Imputation For Missing Data : What Is It And How Can I Use It?‖ , Paper presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago, IL ,pp . 2 - 16, 2003.
[7]. A.Rogier T.Donders, Geert J.M.G Vander Heljden, Theo Stijnen, Kernel G.M Moons, —Review: A gentle introduction to imputation of missing values‖ , Journal of Clinical Epidemiology 59 , pp.1087 – 1091, 2006.
[8]. Kin Wagstaff ,‖Clustering with Missing Values : No Imputation Required‖ -NSF grant IIS-0325329,pp.1-10.
[9]. S.Hichao Zhang , Jilian Zhang, Xiaofeng Zhu, Yongsong Qin,chengqi Zhang , —Missing Value Imputation Based on Data Clustering‖, Springer-Verlag Berlin, Heidelberg ,2008.
[10]. Richard J.Hathuway , James C.Bezex, Jacalyn M.Huband , —Scalable Visual Assessment of Cluster Tendency for Large Data Sets‖, Pattern Recognition ,Volume 39, Issue 7,pp,1315-1324- Feb 2006.
[11]. Qinbao Song, Martin Shepperd ,‖A New Imputation Method for Small Software Project Data set‖, The Journal of Systems and Software 80 ,pp,51–62, 2007.
[12]. Gabriel L.Scholmer, Sheri Bauman and Noel A.card —Best practices for Missing Data Management in Counseling Psychology‖, Journal of Counseling Psychology, Vol. 57, No. 1,pp. 1–10,2010.
[13]. R.Kavitha Kumar, Dr.R.M Chandrasekar,—Missing Data Imputation in Cardiac Data Set‖ ,International Journal on Computer Science and Engineering , Vol.02 , No.05,pp-1836 – 1840 , 2010.
[14]. Jinhai Ma, Noori Aichar –Danesh , Lisa Dolovich, Lahana Thabane , —Imputation Strategies for Missing Binary Outcomes in Cluster Randomized Trials‖- BMC Med Res Methodol. 2011; pp- 11: 18. – 2011.

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Academic conference: http://www.iiste.org/conference/upcoming-conferences-call-for-paper/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar