IJRSET September 2015 Volume 2, Issue 7 Pages: 31-36

ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

MISSING VALUE IMPUTATION IN HIGH DIMENSIONAL MODEL USING CLUSTERING TECHNIQUES

[1] Dr. K. SIVAKUMAR, [2] Dr. G. DALIN
[1] Associate Professor, [2] Assistant Professor,
[1] Department of BCA, [2] PG & Research Department of CS,
[1] Park's College (Autonomous), [2] Hindusthan College of Arts & Science,
[1] Tirupur. [2] Coimbatore.

Abstract: -

Many industrial and research information units incorporate missing values because of quite a lot of explanations. Issues associated with missing values are loss of effectivity, problems in handling and examining the data. Sooner or later missing values challenge will also be handled through missing values imputation. Clustering is normal solutions either fill within the missing values or ignore the lacking knowledge.
This paper work is split into five levels. Selection of enter data from the database is made, performing pre processing on raw knowledge, clustering the pre-processed data making use of hybrid clustering, the outcome of lacking values imputed and customary data is when put next and outcome are interpreted.

Keywords: - Missing, values, Clustering, Data, Imputation, Database.

1. INTRODUCTION

Missing values imputation is an actual yet challenging issue confronted in machine learning and data mining [1, 2]. Missing values may generate bias and affect the quality of the supervised learning process or the performance of classification algorithms [3, 4]. However, most learning algorithms are not well adapted to some application domains due to the difficulty with missing values (for example, Web applications) as most existed algorithms are designed under the assumption that there are no missing values in datasets. Missing values may appear either in conditional attributes or in class attribute (target attribute). There are many approaches to deal with missing values described in [6], for instance: (a) Ignore objects containing missing values; (b) Fill the missing value manually; (c) Substitute the missing values by a global constant or the mean of the objects; (d) Get the most probable value to fill in the missing values. The first approach usually lost too much useful information, whereas the second one is time consuming and expensive in cost, so it is infeasible in many applications. The third approach assumes that all missing values are with the same value, probably leading to considerable distortions in data distribution. Traditional missing value imputation techniques can be roughly classified into parametric imputation (e.g., the linear regression) and non-parametric imputation (e.g., non-parametric kernel-based regression method [20, 21, 22], Nearest Neighbor method [4, 6] (referred to as NN)). The parametric regression imputation is superior if a dataset can be adequately modeled parametrically, or if users can correctly specify the parametric forms for the dataset. For instance, the linear regression methods usually can treat well the continuous target attribute, which is a linear combination of the conditional attributes. However, when we don't know the actual relation between the conditional attributes and the target attribute, the performance of the linear regression for imp uting missing values is very poor. In real application, if the model is misspecified (in fact, it is usually impossible for us to know the distribution of the real dataset), the estimations of

parametric method may be highly biased and the optimal control factor settings may be miscalculated. While nonparametric imputation process is of low-efficiency, the trendy NN method faces two disorders: (1) each instance with missing values requires the calculation of the distances from it to all other occasions in a dataset; and (2) there are only some random possibilities for selecting the closest neighbor. This paper addresses the above issues with the aid of proposing a clustering-centered non-parametric regression procedure for dealing with the hindrance of missing worth in goal attribute (named Clustering headquartered missing worth Imputation, denoted as CMI). In our technique, we refill the lacking values with believable values which can be generated through utilising a kernel-based method. Specifically, we first divide the dataset (together with situations with missing values) into clusters. Then every example with missing-values is assigned to a cluster most just like it.

2. MISSING VALUE IMPUTATION AND CLUSTERING TECHNIQUE

The overall method of this research work is divided into five usual phases as given in determine 2.1. First, determination of enter data from the database is made, performing preprocessing on raw data, clustering the preprocessed information making use of hybrid clustering approach, the results of lacking values imputed and original data is when compared and eventually outcome are interpreted.

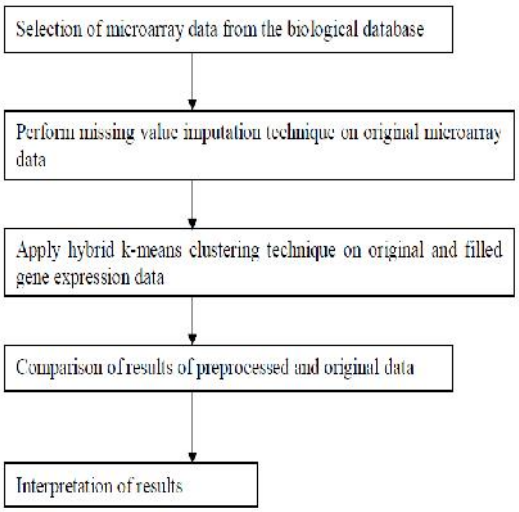

Figure 2.1: - Steps for Proposed system

Imputation methods involve replacing missing values with estimated ones based on some information available in the data set. There are many options varying from naïve methods like mean imputation to some more robust methods based on relationships among attributes.

1. Case substitution

This method is typically used in sample surveys. One instance with missing data (for example, a person that cannot be contacted) is replaced by another non sampled instance;

2. Mean and mode

This method consists of replacing the missing data for a given attribute by the mean (quantitative attribute) or mode (qualitative attribute) of all known values of that attribute;

3. Hot deck and cold deck

In the hot deck method, a missing attribute value is filled in with a value from an estimated distribution for the missing value from the current data. Hot deck is typically implemented into two stages. In the first stage, the data are partitioned into clusters. And, in the second stage, each instance with missing data is associated with one cluster. The complete cases in a cluster are used to fill in the missing values. This can be done by calculating the mean or mode of the attribute within a cluster. Cold deck imputation is similar to hot deck but the data source must be other than the current data source;

4. Prediction model

Prediction models are sophisticated procedures for handling missing data. These methods consist of creating a predictive model to estimate values that will substitute the missing data. The attribute with missing data is used as class-attribute, and the remaining attributes are used as input for the predictive model. An important argument in favor of this approach is that, frequently, attributes have relationships (correlations) among themselves. In this way, those correlations could be used to create a predictive model for classification or regression (depending on the attribute type

with missing data, being, respectively, nominal or continuous). Some
of these relationships among the attributes may be maintained if they were captured by the predictive model.

3. PROPOSES SYSTEM

Information seems to be lacking as a result of a few factors. Researchers listen extra on imputing missing knowledge via dealing with various information Mining algorithms. The most natural missing worth imputation tactics are deleting case, imply worth imputation, highest likelihood and other statistical ways.

In recent years research has explored using laptop studying strategies as a procedure for missing values imputation. Computer studying methods like MLP, SOM, KNN and resolution tree were discovered to perform higher than the natural statistical approaches. [1] in this paper we compare two systems okay-Nearest neighborhood and k-method Clustering mixed with mean substitution. Each the systems workforce the dataset into a few businesses/ clusters. Imply Substitution is applied separately to each and every crew / cluster. When each the results are when put next, ok-NN has an growth in percent of accuracy than k-way Clustering.

3.1. Preprocessing technique

One of the central steps in an information mining method is the preparation and transformation of the initial dataset. But, in most knowledge mining applications, some ingredients of information guidance approach generally, even the entire process can be described independently of an utility and a knowledge mining method. For distributed datasets, probably the most data coaching tasks will also be carried out throughout the design of the data warehouse, however many transformations specialised is also carried out relying on the data mining task. Many transformations could also be wanted to provide features which are extra valuable for selected information mining methods comparable to prediction or classification. Dimensionality Reduction

1. Data cleaning.
2. Data Integration and transformation
3. Data Reduction.

3.2. Missing Value Analysis

Many clustering algorithms for gene expression analysis require a complete matrix of gene expression values as input. For example, methods such as hierarchical clustering and K-means clustering are not robust to missing data, and may lose effectiveness even with a few missing values. In order to minimize the effect of incomplete data set on analysis, missing value imputation methods are needed. It is an important preprocessing step to accurately estimate missing values for gene expression profiles. The reason why these missing values occur is due to human operations, experimental inaccuracy, or unobvious reaction at that time slot of certain genes.

There are many approaches to deal with missing values. For instance:

- Ignore objects containing missing values.
- Fill the missing value manually.
- Substitute the missing values by a global constant.
- Get the most probable value to fill in the missing values.

The first approach usually leads to loss of useful information, whereas the second one is time consuming and expensive, so it is infeasible in many applications. The third approach assumes that all missing values have the same value, probably leading to considerable distortions in data distribution.

There are four missing value imputation methods used during preprocessing.

1. AVG Imputation Method
2. Max Imputation method
3. Minimum Imputation Method
4. Discrete Mean Method.

Among the various data mining clustering techniques, hierarchical clustering and k-means clustering are widely used by researchers. These methods group objects (genes) having many attributes into clusters. Each cluster is similar with respect to certain characteristics; that is, genes in each cluster are similar function to each other. On the other hand, each group should be different from other groups with respect to different characteristics (Bernard Chen et al 2009).However each of these traditional clustering methods has its limitations. Hierarchical clustering method is not efficient if the data set is large. Also, a set of

clusters determined by the hierarchical clustering is not unique as it is based on local decisions without evaluating the clustering result. Therefore, a gene cannot be reassigned once it is assigned to a certain cluster and this may give incorrect clustering results (Sungwoo Kwon and Chonghun Han 2002).The traditional hierarchical clustering method is an agglomerative approach, which organizes similar branch points into a cluster based on the choice of the distance measure and, therefore, results in a tree-like dendrogram.

Usually this method does not guarantee that dendrogram similarity is maximized because each cluster may consist of several different sub-clusters (Hugh Chipman and Robert Tibshirani 2006). In k-means clustering, the method selects initial predetermined k cluster centroids and calculates the proximities from each point to all k centroids. When each datum is assigned to the k cluster members, the data are reallocated to one of the new clusters. The main problem is that the user has to define number of clusters in advance. Different starting points may result in different clustering partitions.

3.3. Imputation with K-means clustering algorithm

K-Means is an algorithm to classify or to group your objects based on attributes/features into K number of group. K is positive integer number.

The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid. Thus, the purpose of K-mean clustering is to classify the data.

Step 1: - Determine the centroid coordinates.

Step 2:- Determine the distance of each object to the centroids.

Step 3: - Group the object based on minimum distance.

Nearest Neighbour Analysis is a method for classifying cases based on their similarity to other cases. In machine learning, it was developed as a way to recognize patterns of data without requiring an exact match to any stored patterns, or cases. Similar cases are near each other and dissimilar cases are distant from each other. Thus, the distance between two cases is a measure of their dissimilarity. Cases that are near each other are said to be “neighbours.” When a new case (holdout) is presented, its distance from each of the cases in the model is computed. The classifications of the most similar cases – the nearest neighbours – are tallied and the new case is placed into the category that contains the greatest number of nearest neighbours.

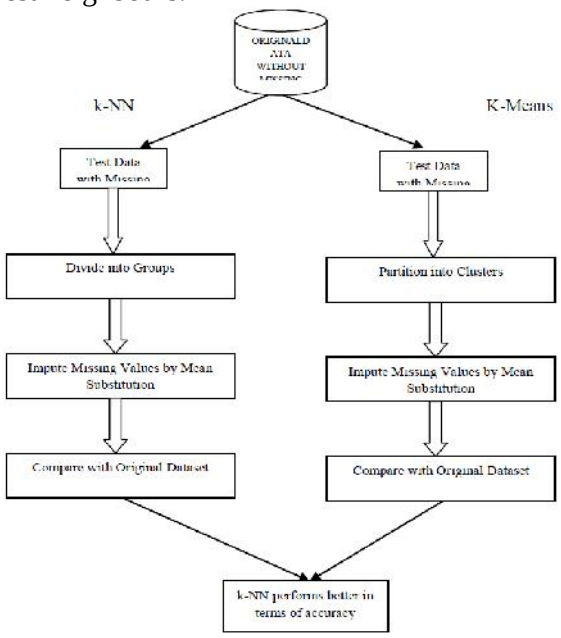

Figure 3.1: - K-Means & K-NN algorithm Process

As a result, clustering results generated by this algorithm suffers from the drawback of producing inconsistent clusters. This issue may be solved by combining k-means clustering with some other clustering technique which is referred as hybrid k-means clustering technique.

4. THE PERFORMANCE EVALUATION

At first scan dataset is made via exchanging some long-established values with missing value(NaN-now not a quantity). Now the customary dataset and scan data set is partitioned into clusters in case of okay-method and corporations in case of k-NN. Lacking worth in each and every crew/cluster is filled with mean value. Now the experiment dataset is in comparison with the normal dataset for finding the accuracy of efficiency. This procedure is repeated for more than a few missing percentages 2,5,10,15 and 20.

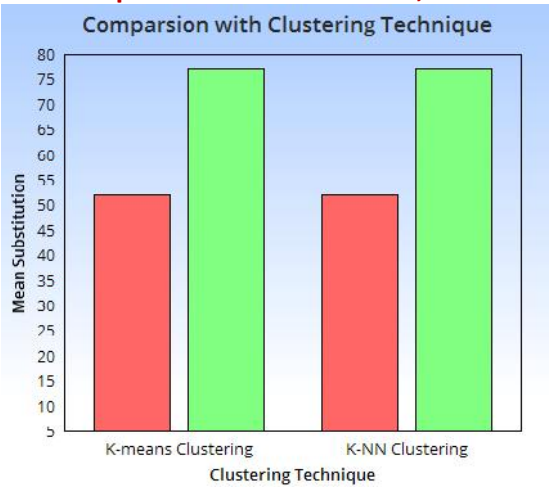

Figure 4.1: - Mean Substitution

5. CONCLUSION

On this paper, we presented a brand new technique to the missing worth hindrance for clustering algorithms. We now have discussed, and demonstrated, the difficulties of information imputation approaches that process imputed values as if they had been as nontoxic as the precise observations. Okay-means and KNN methods furnish rapid and correct methods of estimating missing values.KNN –founded imputations presents for a robust and sensitive strategy to estimating lacking information. Consequently it's recommend. There's scope for a number of new missing value imputation methods centered on utilizing imputed values for later imputations. KNN –situated process for imputation of lacking values. It is also analyzed that once the missing percent is high, whatever the method is the accuracy decreases. This proposed method can be improved by evaluating various desktop finding out techniques like SOM, MLP. Imply Substitution may also be replaced by means of mode, median, common deviation or through applying Expectation – Maximization, regression situated ways.

REFERENCES

[1] J.L Peugh, and C.K. Enders, "Missing data in Educational Research: A review of reporting practices and suggestions for improvement, "Review of Educational Research vol 74, pp 525-556, 2004.

[2] S-R. R. Ester-Lydia , Pino – Mejias Manuel, Lopez Coello Maria-Dolores , Cubiles – de – la- Vega, "Missing value imputation on Missing completely at Random data using multilayer perceptrons, "Neural Networks, no 1, 2011.

[3] B.Mehala, P.Ranjit Jeba Thangaiah and K.Vivekanandan , " Selecting Scalable Algorithms to Deal with Missing Values",International Journal of Recent Trends in engineering, vol.1. No 2, May 2009.

[4] Gustavo E.A.P.A. Batista and Maria Carolina Monard , "A Study of K-Nearest Neighbour as an Imputation method".

[5] Allison, P.D-"Missing Data", Thousand Oaks, CA: Sage -2001.

[6] 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.

[7] Kin Wagstaff ,"Clustering with Missing Values : No Imputation Required" - NSF grant IIS-0325329,pp.1-10.

[8] S.Hichao Zhang , Jilian Zhang, Xiaofeng Zhu, Yongsong Qin,chengqi Zhang , "Missing Value Imputation Based on DataClustering", Springer-Verlag Berlin, Heidelberg ,2008.

[9] 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.

[10] 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

[11] 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.

[12] 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.

[13] 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.

[14] R.S.Somasundaram,R.Nedunchezhian , "Evaluation of Three Simple Imputation Methods for Enhancing Preprocessing of

Data with Missing Values", International Journal of Computer Applications (0975 8887) Volume 21 – No.10 ,pp.14-19 ,May 2011.
[15] K.Raja , G.Tholkappia Arasu , Chitra. S.Nair , "Imputation Framework for Missing Value" , International Journal of Computer Trends and Technology – volume3Issue2 2012.
[16] BOB L.Wall , Jeff K.Elser "Imputation of Missing Data for Input to Support Vector Machines" ,