Predictive Performance of k-NN Classifier and K-Means Clustering in Imputation of Missing Values

Priya.S, Dr Antony Selvadoss Thanamani

Abstract—The presence of missing data in a datasets can affect the performance of classifier which leads to difficulty of extracting useful information from datasets. Missing Data is a widespread problem that can affect the ability to use data to construct effective predictions systems. We analyze the predictive performance by comparing K-Means Clustering with kNN Classifier for imputing missing value. For investigation, we simulate with 5 missing data percentages; we found that k-NN performs better than K-Means Clustering, in terms of accuracy.

keywords—Accuracy,Dataset, K-Means clustering, k-NN Classifier, Missing Data, Percentage, Predictive Performance

1 Introduction

Data Mining is the process of extracting hidden knowledge from large volumes of raw data. Data mining has been defined as "the nontrivial extraction of previously unknown, implicit and potentially useful information from data. Data Mining is used to discover knowledge out of data and presenting it in a form that is easily understood to humans [1]. Data Mining is the notion of all methods and techniques, which allow analyzing very large data sets to extract and discover previously unknown structures and relations out of such huge heaps of details. This information is filtered, prepared and classified so that it will be a valuable aid for decisions and strategies [3].

Data seems to be missing due to several reasons. Researchers concentrate more on imputing missing data by handling various Data Mining algorithms. The most traditional missing value imputation techniques are deleting case, mean value imputation, maximum likelihood and other statistical methods. In recent years research has explored the use of machine learning techniques as a method for missing values imputation. Machine learning methods like MLP, SOM, KNN and decision tree have been found to perform better than the traditional statistical methods. [1]

This paper compares two techniques K-Nearest Neighborhood and K-Means Clustering combined with mean substitution. Both the techniques group the dataset into several groups/ clusters. Mean Substitution is applied separately to each group cluster. When both the results are compared, k-NN has an improvement in percentage of accuracy than K-Means Clustering.

- *Priya.S is currently pursuing Doctrate of philosophy in Computer Science , Bharathiar University, Coimbatore ,India and working as Assistant Professor of computer Science,Govt. First Grade College,K.G.F,. E-mail: priyapunith@yahoo.com*
- *Dr Antony Selvadoss Thanamani is currently working as Professor and Head, Department of Computer Science, NGM College, Coimbatore, India (affiliated to Bharathiar University,Coimbatore) and the Principal Investigator of UGC – MAJOR Research Project in Computer science*

2 Missing Data Mechanism

Missing data or missing values occur when no data value is stored for an instance in the current record. Missing data might occur because value is not relevant to a particular case, could not be recorded when data was collected or ignored by users because of privacy concerns [14]. Most information system usually has some missing values due to unavailability of data. Sometimes data is not presented or get corrupted due to inconsistency of data files. Missing data is a common problem that has a significant effect on the conclusion that can be drawn from the data. Missing data is absence of data items that hide some information that may be important [1].

Types of missing data:

There are basically three types of missing data, these are:

1. MCAR- It is probability of missing data on any attribute does not depend on any value of attribute [7]. The term "Missing Completely at Random" refers to data where the missingness mechanism does not depend on the variable of interest, or any other variable, which is observed in the dataset [2]

2. MAR- The probability of missing data on any attributes does not depends on its own value but value of other attribute [7]. Sometimes data might not be missing at random but may be 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 [2].

3. MNAR- Missing data depends on the values that are missing [7]. Sometimes data might not be missing at random but may be 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 [2].

3 .Missing Data Imputation Methods

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 alue 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 favour 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.

4. 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.

4.1 ALGORITHM:

1. Determine the centroid coordinate
2. Determine the distance of each object to the centroids
3. Group the object based on minimum distance (find the closest centroid)

5. Imputation with K-nearest Neighbor Algorithm

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[17]. 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.

6. Framework And Experimental Analysis

The following framework explains about the comparison made in this paper. The experimental results show the improvement in performance

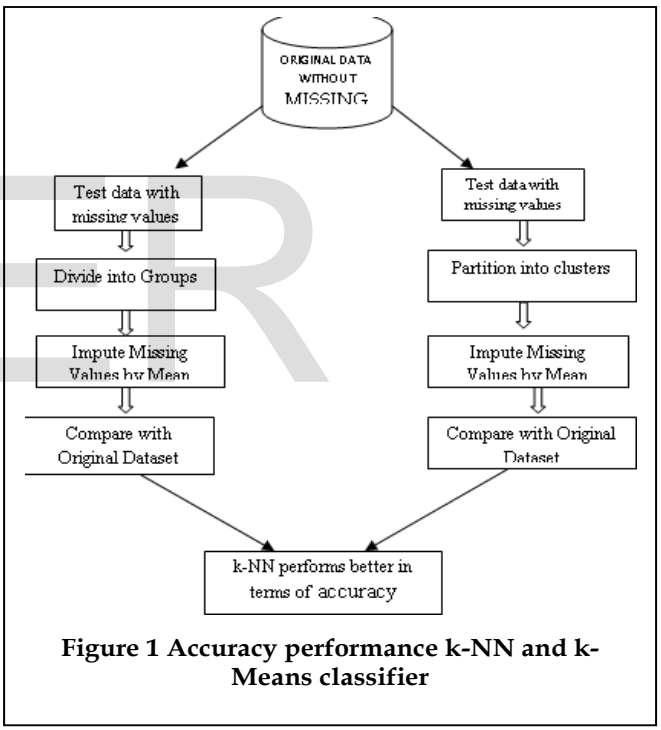

Figure 1 Accuracy performance k-NN and k-Means classifier

Initially test dataset is made by replacing some original values with missing value(NaN-Not a Number). Now the original dataset and test data set is partitioned into clusters in case of K-Means and groups in case of k-NN. Missing value in each group/cluster is filled with mean value. Now the test dataset is compared with the original dataset for finding the accuracy of performance. This process is repeated for various missing percentages 3,4,12,14 and 20.

Table 1
Comparison of clustering techniques in terms of Accuracy

Missing Percentage	Mean Substitution	K-Means Clustering	k-Nearest Neighbour
3%	67	62	69
4%	63	65	69
12%	59	63	68
14%	59	64	67
20%	52	56	60
Percentage	60%	62%	67%

Thus the results are compared. It shows some improvement in percentage of accuracy.

7. Conclusion And Future Enhancement

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. Hence it is recommend KNN –based method for imputation of missing values. It is also analyzed that when the missing percentage is high, whatever the method is the accuracy decreases. This proposed method can be enhanced by comparing various machine learning techniques like SOM, MLP. Mean Substitution can be replaced by mode, median, standard deviation or by applying Expectation – Maximization, regression based methods

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 Data Clustering", 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" .

[17] R. Malarvizhi, Dr. Antony Selvadoss Thanamani, "K-Nearest Neighbor in Missing Data Imputation", International Journal of Engineering Research and Development, Volume 5 Issue 1-November-2012,