Missing Value Imputation Method Based on Clustering and Nearest Neighbours

Satish Gajawada and Durga Toshniwal

***Abstract*—Classification methods have been applied in real life problems. Real world dataset may contain missing values but many classification methods need complete datasets. Hence many missing value imputation methods like clustering based imputation method were proposed in literature. But when objects with missing values are high then the available complete objects in the dataset are less. Imputing missing values with limited amount of complete objects may not give good results when imputation is performed using complete objects only. The number of complete objects can be increased by treating the imputed object as complete object and using the imputed object for further imputations along with the available complete objects. In this paper we propose a missing value imputation method based on K-Means and nearest neighbors. This method uses the imputed objects for further imputations. The proposed method has been applied on clinical datasets from UCI Machine Learning Repository.**

***Index Terms*—Missing value imputation, using imputed values, K-means, Nearest neighbors.**

I. Introduction

Many real world datasets contain missing values which can make it very difficult to use data analysis methods which require complete datasets. This problem can be solved by imputing the missing values. Several missing value imputation methods were proposed in literature and there exists no universally best imputation method [1]. The goal of missing value imputation methods is to fill the missing values of the object using the available information in the object. Clustering methods were used in literature to impute missing values. K-means clustering based imputation consists of 2 steps. In first step K-means clustering is applied to get clusters. In the next step the cluster information is used to impute the missing values [2]. In KNN imputation method the K nearest neighbours of the object with missing values are used to impute the missing values in the object [3]. Mehala et al. [4] applied K-means based imputation on clinical datasets from UCI Machine Learning Repository. Patil et al. [5] proposed K-means based imputation method. Ki Yeol Kim et al. [6] proposed Sequential KNN and it was observed that the performance of this method is better than the other methods. In Sequential KNN method, the information from the imputed values is used for further imputations. In this paper, we propose a missing value imputation method based on K-means and nearest neighbors which uses the information present in imputed objects as well for all further imputations.

II. Proposed Method

Section 2.1 shows proposed method. Section 2.2 explains proposed missing value imputation method.

A. Description of Proposed Method

a) K-means Clustering

The dataset is divided into two sets where one set contains complete instances that do not contain any missing values and the other set contains incomplete instances which contains missing values. K-means clustering is applied on complete instances set to obtain clusters of complete instances.

B. Proposed Method

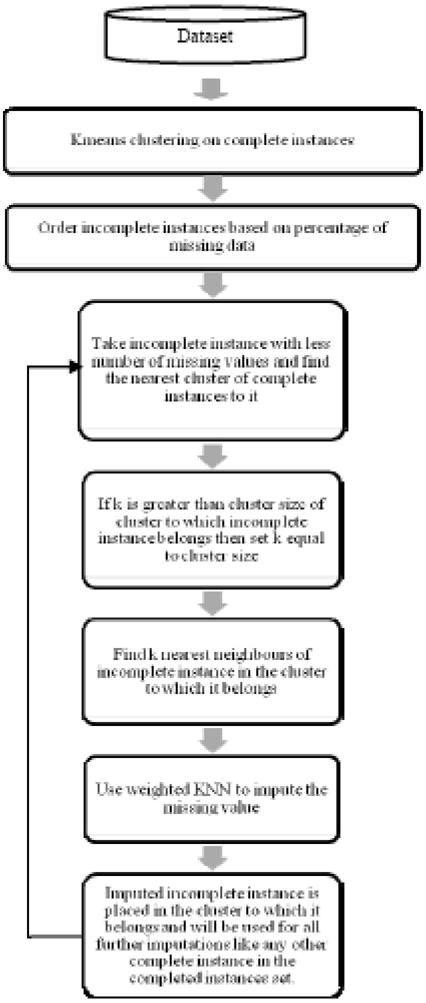
Dataset
Kmeans clustering on complete instances
Order incomplete instances based on percentage of missing data
Take incomplete instance with less number of missing values and find the nearest cluster of complete instances to it
If k is greater than cluster size of cluster to which incomplete instance belongs then set k equal to cluster size
Find k nearest neighbours of incomplete instance in the cluster to which it belongs
Use weighted KNN to impute the missing value
Imputed incomplete instance is placed in the cluster to which it belongs and will be used for all further imputations like any other complete instance in the completed instances set.

Fig. 1. Proposed method

a) Imputing Incomplete Instance

Incomplete instances in the incomplete set are arranged in the order of their missing values such that the instance with

Manuscript received April 12, 2012; revised May 12, 2012

The authors are with the Department of Electronics and Computer Engineering, Indian Institute of Technology Roorkee, Roorkee, India (e-mail: gajawadasatish@gmail.com).

less number of missing values comes first in the list and the instance with more number of missing values comes later in the list. The incomplete instance which is first in the list is taken and cluster which is nearest to this instance is found. K nearest neighbours of the incomplete instance in the cluster nearest to it are found where K is set to cluster size if K is greater than cluster size. The missing values in the incomplete instance are imputed using weighted KNN.

b) Moving to Complete Instance Set

The imputed incomplete instance is moved to complete instance set and assigned to the cluster used for imputing that instance. The cluster center is updated to the centroid of all points in the cluster after assigning imputed instance to the cluster. The imputed instance will be used for all further imputations of other incomplete instances like any other complete instances in the complete instance set.

c) Imputing All Incomplete Instances

Steps described in 2.2.2 and 2.2.3 are repeated until all incomplete instances in the incomplete set are imputed.

III. Experimental Results

We have applied proposed method on various clinical datasets from Machine Learning Repository [7].

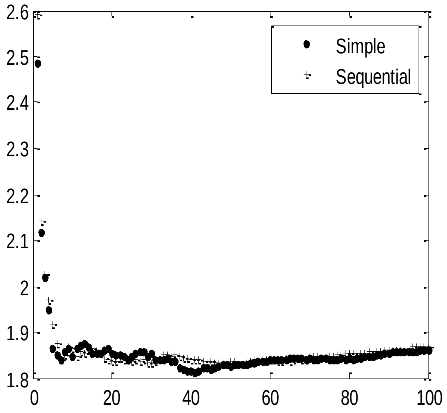

Fig. 2. K neighbours versus RMSE for 5% missing data in WBC data

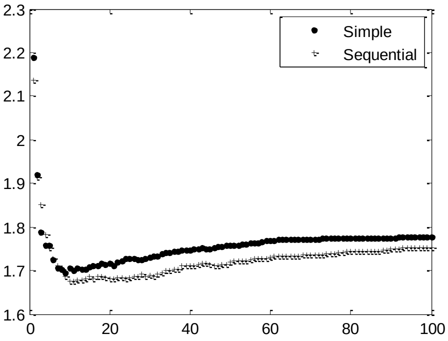

Fig. 3. K neighbours versus RMSE for 15% missing data in WBC data

Fig 2 to Fig 4 shows the results obtained for Wisconsin breast cancer data for different percentage of missing values. Fig 5 to Fig 7 shows the results obtained for Bupa liver disorder dataset. In Fig 2 to Fig 7 the method 'Sequential' refers to the results obtained by using proposed method and the method 'Simple' refers to results obtained without using imputed values for further imputations. We created several datasets with missing values using datasets from UCI Machine Learning repository.

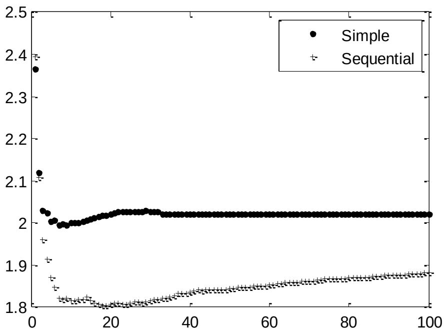

Fig. 4. K neighbours versus RMSE for 25% missing data in WBC data

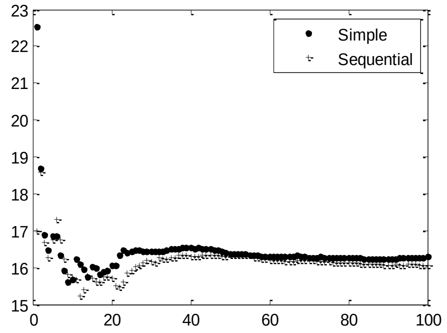

Fig. 5. K neighbours versus RMSE for 5% missing data in BLD data

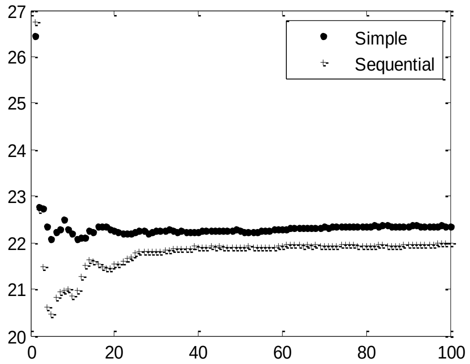

Fig. 6. K neighbours versus RMSE for 15% missing data in BLD data

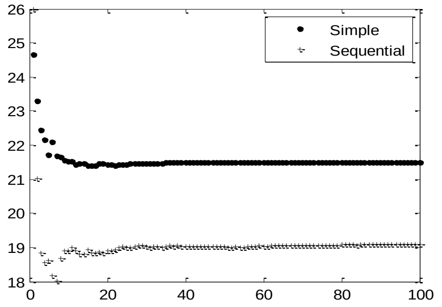

Fig. 7. K neighbours versus RMSE for 25% missing data in BLD data

Although from Figure 2 to Figure 7 we can observe that Root Mean Square Error of imputed values with proposed method is less or almost equal to imputation method without using imputed values but the proposed method may give more error for other cases. This is because if values are wrongly imputed at the beginning then this error propagates to all further imputations and hence proposed method may give more error in those cases. Although many imputation methods are proposed in literature but there are not many methods where the information from imputed values is used in further imputations. Hence there is scope to create several new missing value imputation methods in this direction.

IV. Conclusion

In this paper, we proposed a missing value imputation method based on clustering and nearest neighbors. The proposed method has been applied on clinical datasets from Machine Learning Repository. Although results reported show that proposed method performed better than simple method (without using imputed values for further imputations) but it is not the case for all the datasets as error in earlier imputation may propagate to further imputations. Hence this point should be kept in mind while applying methods similar to proposed methods on real world datasets. There is scope for several new missing value imputation methods based on using imputed values for later imputations.

Acknowledgements

This research work has been partially funded by a research grant received from IBM Corporation, as part of IBM SUR Award.

References

[1] A. Farhangfar, L. Kurgan, and J. Dy, "Impact of imputation of missing values on classification error for discrete data," *Pattern Recognition*, vol. 41, pp. 3692-3705, 2008.

[2] D. Li and J. Deogun, "William Spaulding, Bill Shuart," *Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method*, RSCTC. pp. 573-579, 2004.

[3] E. Acuna and C. Rodriguez, "The treatment of missing values and its effect in the classifier accuracy," *Classification, clustering and data mining applications*, pp. 639-648, 2004.

[4] B. Mehala, K. Vivekanandan, and P. R. J. Thangaiah, "An Analysis on K-Means Algorithm as an Imputation Method to Deal with Missing Values," *Asian Journal of Information Technology*, vol. 9, pp. 434-441, 2008.

[5] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Missing Value Imputation Based on K-Mean Clustering with Weighted Distance," *Communications in Computer and Information Science, 1, Contemporary Computing*, Part 11. 94, pp. 600-609.

[6] Ki-Yeol Kim, Byoung-Jin Kim, and Gwan-Su Yi, "Reuse of imputed data in microarray analysis increases imputation efficiency," *BMC Bioinformatics*, vol. 5, pp. 160-168, 2004.

[7] A. Frank and A. Asuncion, "UCI Machine Learning Repository. Irvine," CA: University of California, School of Information and Computer Science, 2010.