Asian Journal of Information Technology 7 (9): 434-441, 2008
ISSN: 1682-3915

An Analysis on K-Means Algorithm as an Imputation Method to Deal with Missing Values

[1]B. Mehala, [2]K. Vivekanandan and [3]P. Ranjit Jeba Thangaiah
[1]Faculty of G.R., Govindarajulu School of Applied Computer Technology,
P.S.G.R. Krishnammal College for Women, Coimbatore, India
[2]BSMED, Bharathiar University, Coimbatore, India
[3]Department of Computer Science and Engineering, Bharathiar University, Coimbatore, India

Abstract: Imputation is a class of procedures that aims to fill the missing values with estimated ones. This method involves replacing missing values with estimated ones based on some information available in the data set. There are many options varying from naive methods like mean or mode imputation to some learning methods like 4.5°C based on relationships among attributes. In this research the use of K-Means algorithm is analyzed as a new approach to treat missing values. This research is to evaluate the efficiency of K-Means imputation algorithm as an imputation method to treat missing data, comparing its performance with the performance obtained by Mean, Median, Mode and 4.5°C.

Key words: Missing values, imputation, preprocessing, data mining

INTRODUCTION

No quality data, no quality mining results (Jiawei and Kamber, 2006). Data quality is a major concern in Data mining and other correlated area such as Machine learning. Data preparation can be more time consuming than data mining, so it is a challenging task as data mining (Acuna and Rodríguez, 2004). There has been a large increase in the amount of knowledge for dealing with incomplete data on fields such as education (Poirier and Rudd, 1983), economics (Johnson, 1989), psychometrics (Brown, 1983), medicine (Berk, 1987), nursing (Musil *et al*., 2002) and finance (Geoff Morgan, 2002) etc. As most Data mining algorithms induce such as knowledge strictly from data, the quality of knowledge extracted is largely determined by the quality of underlying data. One relevant problem in data quality is the presence of missing data. It is occurred in the phase of data collection.

Missing values are a common occurrence in raw data sets and are problematic to model generation in different fields of study. Development of methods to mediate the trouble these impute values cause, could increase the usefulness of these valuable datasets. There are a number of alternative ways of dealing with missing data (Shichao *et al*., 2005; Chi-Chun and Hahn-Ming, 2004; Mei-Ling *et al*., 2005; Dubes and Jain, 1988) and this document is an attempt to outline some of these approaches.

THE TREATMENT OF MISSING VALUES

Missing data treatment methods can be divided into three categories, as proposed in Little and Rubin (2002).

Ignoring and discarding data: There are two main ways to discard data with missing values. The first one is known as complete case analysis; it is available, in all statistical programs and is the default method in many programs. This method consists of discarding all instances with missing data. The second method is known as discarding instances and/or attributes. This method consists of determining the extent of missing data on each instance and attribute and deleting the instances and/or attributes with high levels of missing data. Before deleting any attribute, it is necessary to evaluate its relevance to the analysis. Unfortunately, relevant attributes should be kept even with high degree of missing values.

Parameter estimation: Maximum likelihood procedures are used to estimate the parameters of a model defined for the complete data. Maximum likelihood procedures that use variants of the Expectation-Maximization algorithm (Dempster *et al*., 1977) can handle parameter estimation in the presence of missing data.

Imputation: Imputation (Daqian and Yang, 2005; Fulufhelo *et al*., 2007) is a class of procedures that aims to fill in the missing values with estimated ones.

Corresponding Author: B. Mehala, Faculty of G.R., Govindarajulu School of Applied Computer Technology,
P.S.G.R. Krishnammal College for Women, Coimbatore, India

The objective is to employ known relationships that can be identified in the valid values of the data set assist in estimating the missing values. This research focuses on imputation of missing data.

IMPUTATION METHODS

Imputation methods (Musil *et al*., 2002) involve replacing missing values with estimated ones based on some information available in the data set. There are many options varying from naive methods like mean or mode imputation (Cristian *et al*., 2005) to some more robust methods based on relationships among attributes. This study surveys some widely used imputation methods, although other forms of imputation are available.

Case deletion: This method consists of discarding all instances with missing values for at least one feature. A variation of this method consists of determining the extent of missing data on each instance and attribute and deletes the instances and/or attributes with high levels of missing data. Before deleting any attribute, it is necessary to evaluate its relevance to the analysis.

Statistical imputation: This is one of the most frequently used methods. It consists of replacing the missing data for a given feature by the mean or mode or median of all known values of that attribute in the class where the instance with missing attribute belongs (Laird and Rubin, 1987; Little and Rubin, 2002).

Hot deck and cold deck imputation: In 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. In Random Hot deck, a missing value of an attribute is replaced by an observed value of the attribute chosen randomly. Some cold deck imputation methods are similar to hot deck method, but in this case the data source to choose the imputed value must be different from the current data source (Acuna and Rodríguez, 2004).

Imputation using a predicate 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 the response attributes 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 among themselves. In this way, those correlations could be used to create a predictive model for classification or regression (Mundfrom and Whitcomb, 1998).

KNN imputation: In KNN imputation, the missing values of an instance are imputed considering a given number of instances that are most similar to the instance of interest. The similarity of two instances is determined using a distance function (Cover and Hart, 1967).

Imputation using decision tree algorithms: The decision tree building algorithm deals with the problem of missing values. Ian H. Witten and Eibe Frank (2005) outlined a solution that involves notationally splitting the instances into pieces, using a numeric weighting method and sending part of it down each branch. Eventually, the various parts of the instances will reach the leaf node and the decisions at these leaf nodes must be applied to partial instances. Instead having integer counts, the weights are used. The same weight procedure is used to partition the training set once a splitting attribute has been chosen, to allow recursive application of the decision tree formulation procedure on each daughter nodes. Instances for which the relevant value is missing are notationally splitting the instances into pieces, using a numeric weighting method and sending part of it down the various braches. Pieces of the instance contribute to decisions at lower nodes in the usual way through the information gain calculation. They may be further split at lower nodes, if the values of other attributes are unknown as well.

The CN2 algorithm (Peter and Niblett, 1988) uses rather simple imputation method to treat missing data. CN2 combines the efficiency and ability to cope with noisy data of ID3 with if-then rule from and flexible search strategy. The representation for rules output by CN2 is an ordered set of if-then rules, also known as a 'decision list'. CN2 uses a heuristic function to estimate search during rule construction, based on an estimate of noise present in the data. Every missing value filled in with its attribute most common known values, before calculating the entropy measure. This results in rules that do not necessarily classify all the training examples correctly.

All the decision trees classifiers handle missing values by using built in approaches. For instance, CART replaces a missing value of a given attribute using the corresponding value of a surrogate attribute, which has the highest correlation with the original attribute. About 4.5°C (Quinlan, 1993) uses a probabilistic approach to handle missing data in both the training and the test sample. About 4.5°C also contains a mechanism to re-express decision trees as ordered lists of if-then rules. Each path from the root of the tree to a leaf gives the

conditions that must be satisfied if a case is to be classified by that leaf. About 4.5°C generalizes this prototype rule by dropping any conditions that are irrelevant to the class, guided again by the heuristic for estimating true error rates.

K-Means imputation: K-Means imputation an extension of basic K-Means (Dubes and Jain, 1988; Kaufman and Rousseeuw, 1990) that accounts for unavailable values that is K-Means imputation is similar to K-Means algorithm but it handles the missing data in the data set. Define K centroids, one for each cluster. These centroids should be placed in a cunning way because different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point need to re-calculate K new centroids as barycenters of the clusters resulting from the previous step. After these K new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop may notice that the K centroids change their location step by step until no more changes are done. Calculate mean of all point in the cluster. Let, us consider k is number of cluster. Let us consider that K-number of clusters the value x_{ij} of the m-th class, C_{km} is missing in the K^{th} cluster then it will be replaced by

$$x_{kj} = \sum_{i:\, x_{ij} \in C_{km}} \frac{x_{kj}}{n_{kjm}}$$

where, n_{kjm} represents the number of non-missing values in the j-th feature of the k-th class.

EXPERIMENT ANALYSIS

Four numerical data sets are taken for this research. The tables given below are used to obtain the missing data treatment methods. This research aims to determine the performance of getting missing values of different methods, based on the result tables, the best method can be used for treating the missing values. The following section show the experimental results for the Bupa, CMC, Pima and Breast data sets. Result is tabulated based on the actual value and ± of actual value, which is predicated by each method. For better understanding the values are converted into percentage.

A dataset without missing value is taken; randomly few values in each row are removed. The rates of the value taken out are 2, 4, 6, 8, 10 and 12%, respectively. All the five methods, namely Mean, Median, Mode, 4.5°C and K-Means with number of clusters are applied to the datasets with missing values in order to obtain a non-missing value data set. The table values for each dataset are restricted for the results obtained, when the missing values are at 2%. The other results at 4, 6, 8, 10 and 12% are given in the graphs.

This study shows the performance of mean, median, mode, 4.5°C and K-Means imputation method. Each graph compares the performance of all methods, induced from data with different levels of missing values on a set of attributes.

Bupa data set: In this dataset, when the missing value is less, the performance of K-Means cluster 2 and cluster 3 is better when compare to other methods. When the number of missing values is high the performance of 4.5°C algorithm obtained good results. Table 1 gives the comparative results for the Bupa dataset with missing values at 2%. Figure 1-6 gives the graph obtained with missing value at 2, 4, 6, 8, 10 and 12%, respectively.

Breast cancer dataset: Missing data imputation with 4.5°C method provides good result for all the cases in this

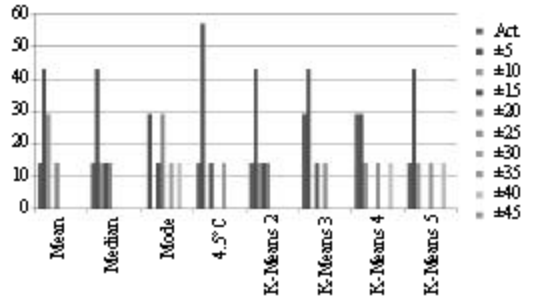

Fig. 1: Missing values at 2%

Table 1: Comparative results for the Bupa dataset-missing values at 2%

Methods	Act	±5	±10	±15	±20	±25	±30	±35	±40
Mean	14	43	29	0	14	0	0	0	0
Median	14	43	14	14	14	0	0	0	0
Mode	0	29	0	14	29	0	14	0	14
4.5°C	14	57	0	14	0	0	14	0	0
K-means cluster 2	14	43	14	14	14	0	0	0	0
K-Means cluster 3	29	43	0	14	0	14	0	0	0
K-Means cluster 4	29	29	14	0	0	14	0	0	14
K-Means cluster 5	14	43	14	0	0	14	0	0	14

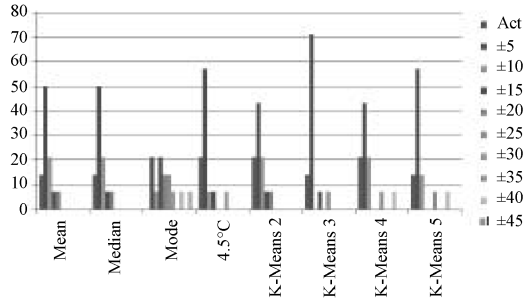

Fig. 2: Missing values at 4%

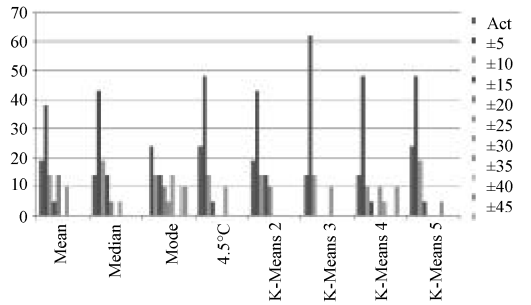

Fig. 3: Missing values at 6%

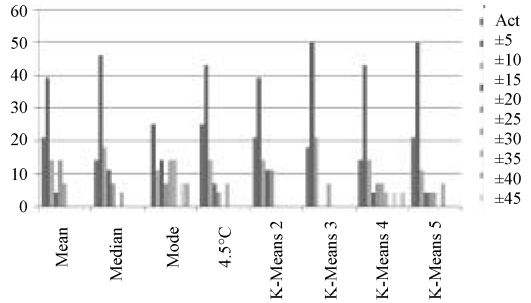

Fig. 4: Missing values at 8%

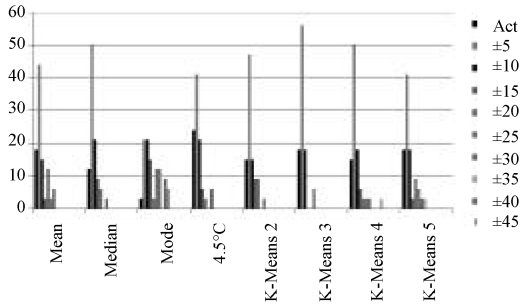

Fig. 5: Missing values at 10%

dataset. Table 2 gives the comparative results for the Breast Cancer dataset with missing values at 2%. Figure 7-12 gives the graph obtained with missing value at 2, 4, 6, 8, 10 and 12%, respectively.

Pima dataset: In this dataset, when the missing value is less, the performance of 4.5°C is better when compare with

Table 2: Comparative results for the Breast Cancer dataset-missing values at 2%

Methods	Act	±2	±4	±6	±8	±10
Mean	14	57	21	0	7	0
Median	50	7	29	7	0	7
Mode	50	21	14	7	0	7
4.5°C	71	29	0	0	0	0
K-Means cluster 2	43	43	14	0	0	0
K-Means cluster 3	50	36	7	7	0	0
K-Means cluster 4	43	50	0	7	0	0
K-Means cluster 5	50	29	7	14	0	0

Table 3: Comparative results for the Pima dataset-missing values at 2%

Methods	Act	±2	±4	±6	±8	±10	±12	±15	±20	±25
Mean	0	27	0	7	7	7	7	7	7	13
Median	7	20	0	7	13	0	7	0	7	13
Mode	13	0	7	0	0	0	0	0	0	7
4.5°C	13	0	0	7	7	0	7	0	7	13
K-Means cluster 2	0	13	0	7	13	13	0	0	7	0
K-Means cluster 3	0	13	0	7	7	7	13	0	7	0
K-Means cluster 4	0	13	0	7	7	7	13	0	0	0
K-Means cluster 5	7	0	0	13	0	13	13	0	0	0

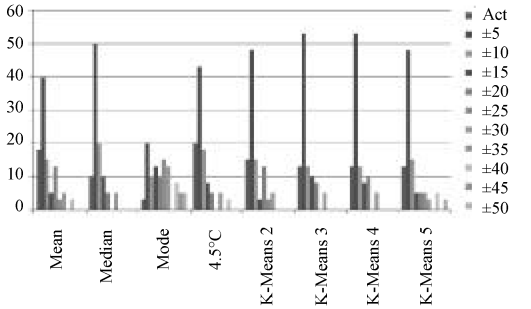

Fig. 6: Missing values at 12%

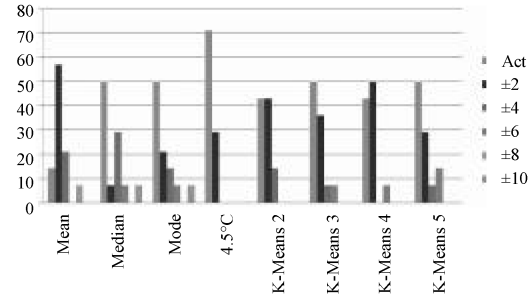

Fig. 7: Missing values at 2%

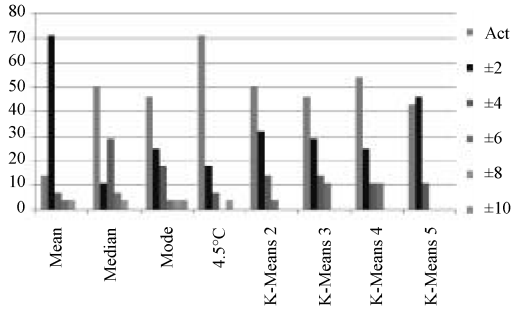

Fig. 8: Missing values at 4%

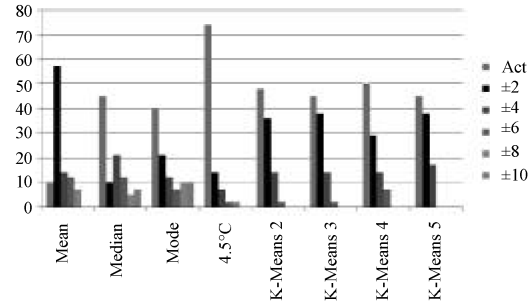

Fig. 9: Missing values at 6%

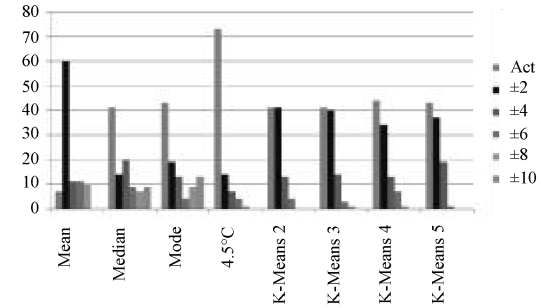

Fig. 10: Missing values at 8%

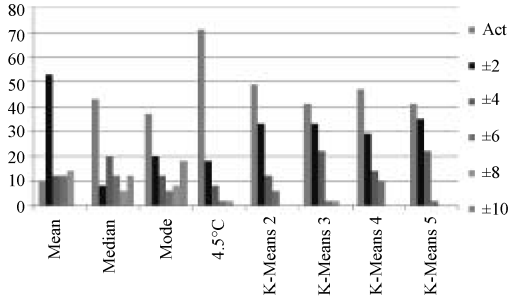

Fig. 11: Missing values at 10%

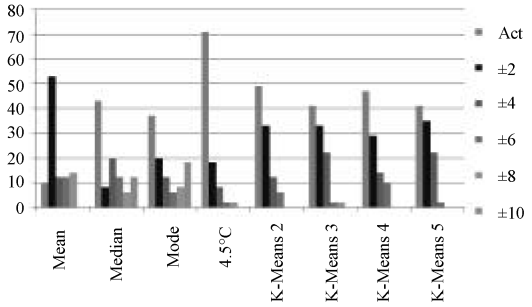

Fig. 12: Missing values at 12%

other methods. When the number of missing values is high the performance of K-Means cluster 5 is superior. Table 3 gives the comparative results for the Pima dataset with missing values at 2%. Figure 13-18 gives the graph obtained with missing value at 2, 4, 6, 8, 10 and 12%, respectively.

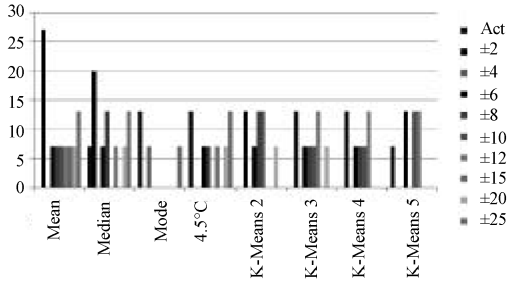

Fig. 13: Missing values at 2%

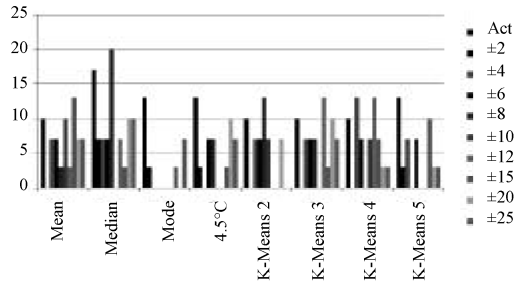

Fig. 14: Missing values at 4%

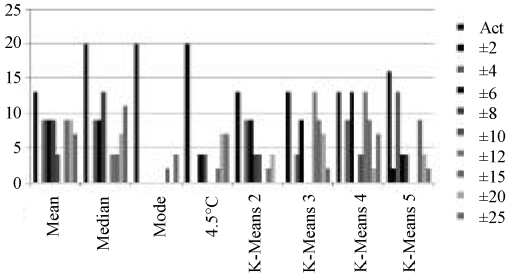

Fig. 15: Missing values at 6%

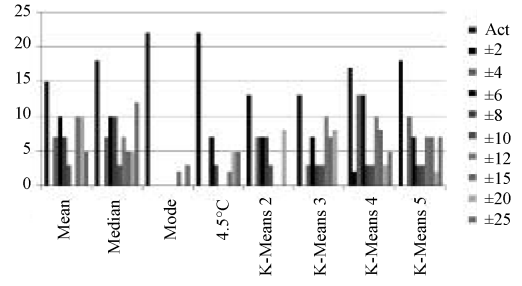

Fig. 16: Missing values at 8%

CMC dataset: In this dataset, the performance of C4.5 is superior to the performance of other methods for the CMC dataset. Median also obtained good result. K-Means gives more number of near by values for all cases. Table 4 gives the comparative results for the CMC dataset with missing values at 2%. Figure 19-24 gives the graph obtained with missing value at 2, 4, 6, 8, 10 and 12%, respectively.

Table 4: Comparative results for the CMC dataset-missing values at 2%

Methods	Act	±2	±4	±6	±8	±10	±12
Mean	33	63	0	4	0	0	0
Median	41	56	0	4	0	0	0
Mode	22	37	26	4	4	0	0
4.5°C	59	33	4	0	0	4	0
K-Means cluster 2	15	67	19	0	0	0	0
K-Means cluster 3	15	67	19	0	0	0	0
K-Means cluster 4	15	67	19	0	0	0	0
K-Means cluster 5	15	67	15	0	0	0	0

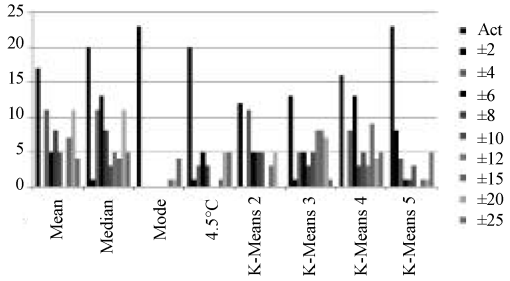

Fig. 17: Missing values at 10%

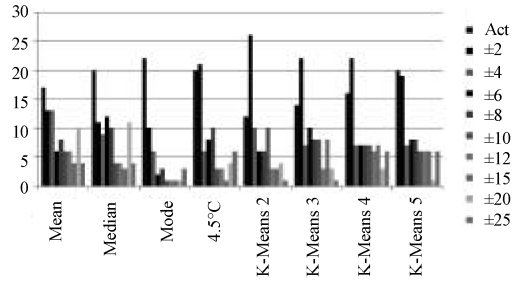

Fig. 18: Missing values at 12%

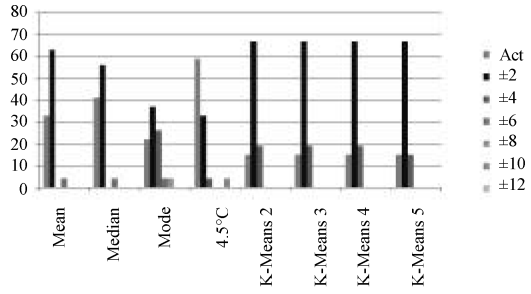

Fig. 19: Missing values at 2%

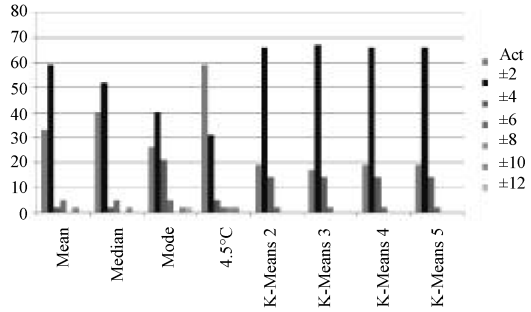

Fig. 20: Missing values at 4%

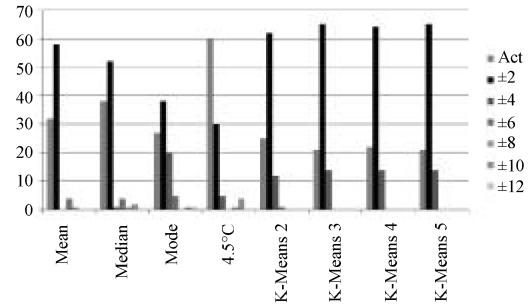

Fig. 21: Missing values at 6%

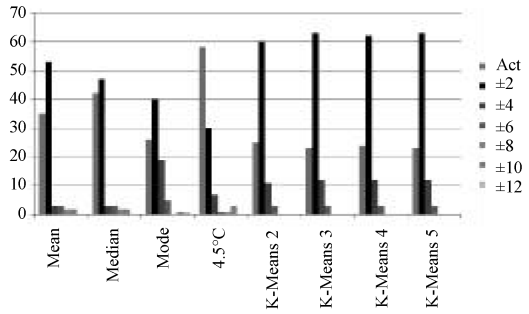

Fig. 22: Missing values at 8%

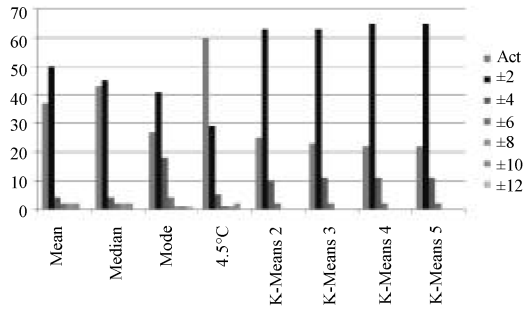

Fig. 23: Missing values at 10%

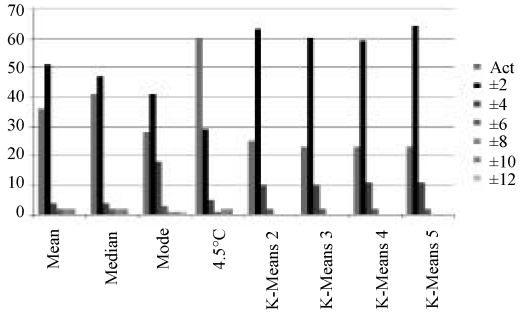

Fig. 24: Missing values at 12%

The complete results with several values of K-Means imputation are given. The performance of mean, median, mode, 4.5°C and K-Means imputation methods are shown. Each graph compares the

performance of all methods, induced from data with different levels of missing values on a set of attributes.

CONCLUSION AND LIMITATIONS

Missing data imputation can be harmful because even most advanced imputation method is only able to approximate the actual value. The predicated values are usually better-behaved, since they conform to other attribute values. This research analyses the behavior of five methods for missing data treatment: Mean, Median, Mode, 4.5°C algorithm to treat missing data and K-Means for missing data imputation. These methods are analyzed inserting different percentage of missing data into different attributes of four data sets, showing promising results. For the data sets Bupa, Breast Cancer and Pima the K-Means imputation provides good result in most cases.

The proposed approach is analyzed and checked with only numerical attributes. In future, it can be extended to handle categorical attributes. In addition, studies can be conducted with additional datasets to see, if the results carry over to those dataset as well. The same work can be extended for large datasets as well as increasing the number of clusters. The methods in this research can be compared based on other factors like time, space, cost etc. The behavior methods can be analyzed when missing values are not randomly distributed. For an effective analysis, not only the error rate has to be inspected, but also the quality of knowledge induced by learning system should be considered.

REFERENCES

Acuna, E. and C. Rodriguez, 2004. The treatment of missing values and its effect in the classifier accuracy. Classification, Clustering and Data Mining Applications. Springer-Verlag Berlin-Heidelberg, pp: 639-648. http://academic.uprm.edu/~eacuna/IFCS04r.pdf.

Berk, K., 1987. Computing incomplete repeated measures. Biometrics, 43: 269-291. PMID: 2440484.

Brown, C.H., 1983. Asymptotic comparison of missing data procedures for estimating factor loadings. Psychometrics, Springer New York, 48 (2): 269-291. DOI: 10.1007/BF02294022.

Chi-Chun, H. and L. Hahn-Ming, 2004. A Grey-Based Nearest Neighbor Approach for Missing Attribute Value Prediction. J. Applied Intelligence (JAI), Kluwer Academic Publishers, Manufactured in The United States, 20: 239-252. DOI: 10.1023/B:APIN.0000021416.41043.0f.

Cover, T.M. and P.E. Hart, 1967. Nearest neighbor pattern classification. IEEE Trans. Inform. Theor., 13: 21-27. http://www.stanford.edu/~cover/papers/transIT/0021cove.pdf.

Cristian P., D. Alain, P. Monique and K. Tahar, 2005. Tools for statistical analysis with missing data: Application to a large medical database. ENMI, pp: 181-186. http://www.magic5.unile.it/PapDoc/Article/MIE2005/TOC%20Scientific%20Contributions/Decision%20Support%20and%20Clinical%20Guidelines/165.pdf.

Daqian, G. and G. Yang, 2005. Incremental gradient descent imputation method for missing data in learning classifier systems. GECCO, ACM, Washington, DC, USA, pp: 72-73. DOI: http://doi.acm.org/10.1145/1102256.1102270, http://portal.acm.org/citation.cfm?id=1102270&CFID=6188267&CFTOKEN=42900381.

Dempster, A.P., R.J. Laird and D.B. Rubin, 1977. Maximum likelyhood from incomplete data via the EM algoritm (with Discussion). J. R. Stat. Soc., B39: 1-38. http://www.jstor.org/pss/2984875.

Dubes, R.C. and A.K. Jain, 1988. Algorithms for Clustering Data. Prentice Hall College, ISBN-10: 013022278X, 13: 978-0130222787.

Fulufhelo, V., Nelwamondo and M. Tshilidzi, 2007. Rough sets computations to impute missing data. Comput. Vision and Pattern Recog., 1: 1-19. DOI: 0704.3635. http://arxiv.org/PS_cache/arxiv/pdf/0704/0704.3635v1.pdf.

Geoff, M., 2002. Cleaning Financial Data. The Numerical Algorithms Group, published by Financial Engineering News. http://www.nag.co.uk/IndustryArticles/Cleaning_Financial_Data.pdf.

Ian, H.W. and E. Frank, 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco. 2nd Edn. ISBN: 0-12-088407-0.

Jiawei, H. and K. Micheline, 2006. Data mining Concept and Techniques. 2nd Edn. Morgon Kaufmaan Publishers. ISBN: 1-55860-901-6.

Johnson, E.G., 1989. Considerations and techniques for the analysis of NAEP data. J. Edu. Stist., 14: 03-334.

Kaufman, L. and P. Rousseeuw, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons. ISBN-10: 0471878766, 13: 978-0471878766.

Laird, R.J. and D.B. Rubin, 1987. Statistical Analysis with missing Data. New York: John Wiley and Sons. ISBN-10: 0471802549, 13: 978-0471802549.

Little, R.J. and D.B. Rubin, 2002. Statistical Analysis with Missing Data. 2nd Edn. John Wiley and Sons, New York. ISBN-10: 0471183865, 13: 978-0471183860.

Mei-Ling, S., I.P. Kuruppu-Appuhamilage, S.C. Chen and L.W. Chang, 2005. Handling missing values via decomposition of the conditioned set. IEEE-IRI Int. Conf. Inform. Reuse and Integration, 15-17: 199-204. DOI : 10.1109/IRI-05.2005.1506473. http://ieeexplore.ieee.org/Xplore/login.jsp?url=/iel5/10065/32280/01506473.pdf?arnumber=1506473.

Mundfrom, D.J. and A. Whitcomb, 1998. Imputing missing values: The effect on the accuracy of classification. Multiple Linear Regression Viewpoints, 25 (1): 13-19. http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/15/7f/96.pdf.

Musil, C.M., C.B. Warner, P.K. Yobas and S.L. Jones, 2002. A comparison of imputation techniques for handling missing data. Western J. Nurs. Res., 24 (5): 815-829. DOI: 10.1177/019394502762477004.

Peter C. and T. Niblett, 1988. The CN2 Induction Algorithm. Machine Learn. J., 3 (4): 261-283. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.53.9180.

Poirier, D.J. and P.A. Rudd, 1983. Diagnostic testing in missing data models. Int. Econ. Rev., 24: 537-546. http://www.jstor.org/pss/2648784.

Quinlan, J.R., 1993. C4.5: Programs for machine learning. Morgan Kaufmann, Los Altos, California. ISBN: 1-55860-238-0.

Shichao, Z., Q. Zhenxing, X.L. Charles, S. Shengli, 2005. Missing is Useful: Missing Values in Cost-Sensitive Decision Trees. IEEE Trans. Knowledge and Data Eng., 17 (12): 1689-1693. DOI: http://doi.ieeecomputersociety.org/10.1109/TKDE.2005.188.