International Journal of Recent Trends in Engineering, Vol. 1, No. 2, May 2009

Selecting Scalable Algorithms to Deal With Missing Values

B. Mehala[1], P. Ranjit Jeba Thangaiah[2], and K. Vivekanandan[3]
[1]G R Govindarajulu School of Applied Computer Technology, P S G R Krishnammal College for women, Coimbatore, India
[2]Research Scholar, Department of Computer Science and Engineering, Bharathiar University, Coimbatore, India
[3]Reader, BSMED, Bharathiar University, Coimbatore, India
Email: {b.mehala, ranjitjebathangaiah, vivekbsmed}@gmail.com

***Abstract* - Missing data is a common feature for large data sets in general. 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. One advantage of this approach is that the imputation phase is separated from the analysis phase, allowing different data mining algorithms to be applied to complete data sets. There are many options varying from naive methods like mean or mode imputation to some learning methods, based on relationships among attributes. This work analyses the behavior of C4.5 to handle missing data in classification based mining algorithm and K-Means to handle missing data in cluster based mining algorithm.**

***Index terms* - Missing values, imputation, preprocessing, data mining, K-Means, C4.5**

I. Introduction

In many applications of data mining considerable part of data seems to be missing. Despite the frequency occurrence of missing data, most data mining algorithms handle missing data in a rather informal way, or simply ignore the problem. Missing data can be divided into three classes as proposed by Laird, R.J. et al.[3]. Missing Completely At Random (MCAR), Missing At Random (MAR) and Not Missing At Random (NMAR). In this work the case of MAR-type missing data is considered (i.e.) the probability of an instance having a missing value for an attribute may depend on the known values, but not on the value of missing data itself. There are a number of alternative ways of dealing with missing data [11, 8, 10, 14]. This document is an attempt to outline some of those approaches.

II. The Treatment Of Missing Values

There are several methods for treating missing data, some methods are described below. Missing data treatment methods can be divided into three categories, as proposed in [3].

A. Ignoring and discarding data.

There are two main ways to discard data with missing values. The first method 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.

B. 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 can handle parameter estimation in the presence of missing data [1].

C. Imputation.

Imputation method is a class of procedures that aims to fill in the missing values with estimated ones. 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 [7, 4]. This work focuses on imputation of missing data. More details about this class of methods are described in the next session.

III. Review Of Previous Work

Development of suitable methods to impute missing values can increase the value of the datasets. There are many options varying from naive methods like mean or mode imputation [5] to some more robust methods based on relationships among attributes. This section surveys some commonly and widely used imputation methods, although other forms of imputation are available.

A. Statistical Imputation

This is one of the most frequently used methods. It consists of replacing the missing data for a given feature by mean or mode or median of all known values of that attribute in the class where the instance with missing attribute belongs [3].

B. *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 [15]. The similarity of two instances is determined using a distance function.

C. *Imputation using decision trees algorithms*

All the decision trees classifiers handle missing values by using built in approaches. CN2 combines the efficiency and ability to cope with noisy data of ID3 with if-then rule from and flexible search strategy [12]. 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. C4.5 uses a probabilistic approach to handle missing data in both the training and the test sample [9].

D. *K-means Imputation*

The principle is to perform a clustering of data as a whole using the K-Means clustering method whilst taking into account the missing values in calculation of the distances via an appropriate metric [13, 2]. Each individual is assigned to a unique cluster and the missing value for the variable x is then replaced by the mean of x calculated from all the individuals in the 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

$$\hat{x}_{kij} = \sum_{i: x_{kij} \in c_{km}} \frac{x_{kij}}{n_{km}}$$

Where $n_{km,}$ represents the number of non-missing values in the j-th, feature of the k-th class.

IV. EXPERIMENTAL PROCEDURE

The main objective of the experiments conducted in this work is to compare the efficiency of K-Means imputation algorithm as an imputation method to treat missing data and C4.5. In these experiments, missing values are artificially imputed, in different rates and attributes, into the data sets. In particular, the behavior of these treatments is analyzed when the amount of missing data is high. Each graph compares the performance of methods, induced from data with different levels of missing values on a set of attributes.

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 methods, namely C4.5 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 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. *Performance Comparison for the Bupa Database*

In the case of few number of missing the performance of k-means cluster 2 and cluster 3 is better when compare with other methods. When the number of missing values is high the performance of C4.5 algorithm obtained good results.

TABLE I

Comparative results for the Bupa dataset

Missing Value	C4.5	K –Means			
		Cluster 2	Cluster 3	Cluster 4	Cluster 5
2%	69.857143	70.428571	75.285714	67.142857	66.142857
4%	73.928571	74.714286	79.857143	75.071429	73.785714
6%	71.190476	71.809524	72.952381	64.952380	75.952381
8%	74.285714	69.107142	69.392857	64.571428	69.392857
10%	70.285714	65.171428	69.085714	66.085714	66.857143
12%	68.975610	64.690478	67.128919	65.690477	66.640534

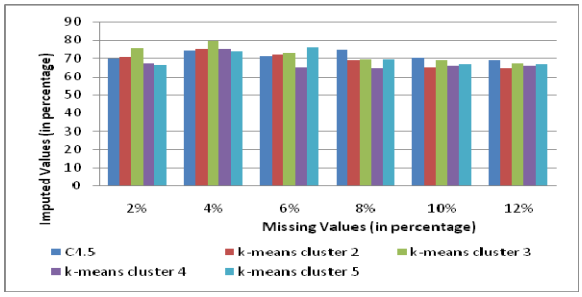

Figure 1Comparative Result of C4.5 and K-means for Bupa Dataset. Numerical value for the graph is given in Table 1.

B. *Performance Comparison for the Breast Cancer Database*

Missing data imputation with C4.5 method provides good result for all the cases.

TABLE II

Comparative results for the Breast Cancer dataset

Missing Value	**C4.5**	**K – Means**			
		Cluster 2	**Cluster 3**	**Cluster 4**	**Cluster 5**
2%	77.285714	60.500000	63.156651	65.750000	62.000000
4%	81.964286	62.892857	71.142857	72.812500	65.999999
6%	85.690476	74.547619	73.500000	74.214285	70.119047
8%	86.372549	73.137255	74.431372	74.372549	70.431372
10%	86.250000	71.814286	70.328571	69.152380	64.904762
12%	85.785714	72.261905	70.840733	70.517053	67.197876

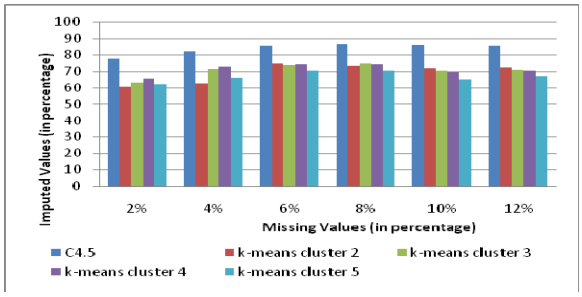

Figure 2. Comparative Result of C4.5 and K-Means for Breast Cancer Dataset. Numerical value for the graph is given in Table 2

C. *Performance Comparison for the Pima Database*

In the case of few number of missing the performance of C4.5 is better when compare with other methods. When the number of missing values is high the performance of k-means cluster 5 is superior.

TABLE III
Comparative results for the Pima dataset

Missing Value	C4.5	K – Means			
		Cluster 2	Cluster 3	Cluster 4	Cluster 5
2%	54.333333	46.666667	44.533333	40.933333	47.866667
4%	48.225806	56.290323	57.612903	53.806452	48.419355
6%	52.543478	51.452445	56.323370	58.282609	54.739130
8%	54.327869	54.537705	57.703005	58.983607	56.836066
10%	51.407895	54.842983	55.939474	58.013158	55.107380
12%	52.097826	56.641304	58.858695	56.336957	54.684783

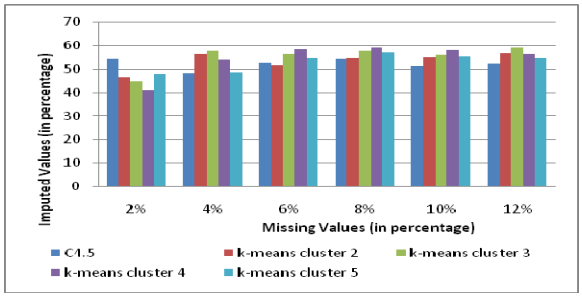

Figure 3. Comparison of C4.5, and K-means for Pima Dataset. Numerical value for the graph is given in Table 3.

D. *Performance Comparison for the CMC Database*

The performance of C4.5 is in most cases 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 IV
Comparative results for the CMC dataset

Missing Value	C4.5	K – Means			
		Cluster 2	Cluster 3	Cluster 4	Cluster 5
2%	85.111111	37.518519	37.185185	37.555556	37.111111
4%	83.313725	39.322034	37.576271	39.372881	37.881356
6%	78.579545	39.784091	38.977273	40.045455	40.443182
8%	78.288136	39.042373	38.262712	39.347458	39.288136
10%	78.442177	38.700680	38.183673	38.530612	38.993197
12%	54.333333	38.412429	38.435028	38.186441	38.649718

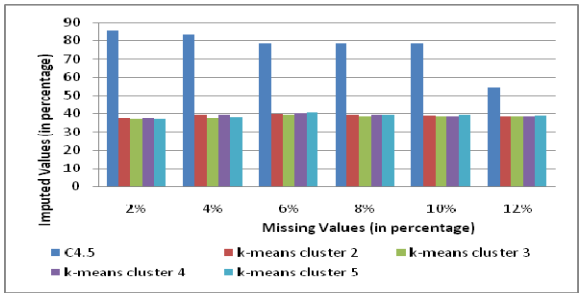

Figure 4. Comparative Result of C4.5 and K-means for CMC Dataset. Numerical value for the graph is given in Table 4.

V. Conclusion And Limitations

Missing values are traditionally regarded as a tough problem and must be imputed before the dataset is used. Missing data imputation can be harmful because even most of the advanced imputation method existing can only be able to approximate the actual value. The predicated values are usually better-behaved, since they conform to other attribute values. In this work, as more attributes with missing values are inserted and the amount of missing data increased, simpler are induced models. This work analyses the behavior and efficiency for missing data treatment: C4.5 algorithm to treat missing data and K-means for missing data imputation. These methods are analyzed by inserting different percentage of missing data into different attributes of the four commonly used data sets, showing promising results. For the data sets Bupa, Breast Cancer and Pima the K-means imputation provides good results in most cases.

The proposed approach uses only numerical attributes to impute the missing values. In further it can be extended to handle categorical attributes. Same methods can be used to compare with other factors like time, space, cost etc. As a future work, the behavior methods can be analyzed when missing values are not randomly distributed. In this case, there is a possibility of creating invalid knowledge. For an effective analysis, not only the error rate has to be inspected, but also the quality of knowledge induced by the learning system should be considered.

REFERENCES

[1] A. P. Dempster, R. J. Laird, D. B. Rubin,"Maximum Likelyhood from Incomplete Data via the EM Algoritm (with Discussion)". *Journal of Royal Statistical Society* vol. B39, pp. 1-38, 1977.

[2] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley & Sons, 1990.

[3] R. J. Laird, D. B. Rubin, *Statistical Analysis with Missing Data*. Second Edition. John Wiley and Sons, New York, 2002

[4] C. M. Musil, C. B. Warner, P. K. Yobas, S. L. Jones. “A Comparison of Imputation Techniques for handling Missing Data”, *Western Journal of Nursing Research*, 24 (5), pp. 815-829, 2002

[5] C. Preda, A. Duhamel, M. Picavet, T. Kechadi. "Tools for Statistical Analysis with Missing Data: Application to a Large Medical Database", *ENMI*, pp. 181 – 186, 2005.

[6] D. J. Mundfrom, A. Whitcomb, “Imputing missing values: The effect on the accuracy of classification”. *Multiple Linear Regression Viewpoints*. 25(1), pp. 13-19, 1998.

[7] F. Vincent, Nelwamondo and T. Marwala, “Rough Sets Computations to Impute Missing Data” Private Bag 3, Wits, 2050, South Africa, 2007

[8] G. Kalton, D. Kasprzyk. *"The treatment of missing survey data"*. Survey Methodology, vol. 12, pp. 1-16, 1986.
[9] J. R. Quinlan *C4.5: Programs for Machine Learning*, Morgan Kaufmann, Los Altos, California, 1993.
[10] M. Shyu, I. P .Kuruppu-Appuhamilage, S.Chen and L.Chang, "Handling Missing Values via Decomposition of the Conditioned Set", *IEEE IRI conference*,2005
[11] P. Chan and O. J. Dunn," The treatment of missing values in discriminant analysis". *Journal of the American Statistical Association*, vol. 6, pp. 473-477, 1972.
[[13] R. C. Dubes and A. K. Jain. *Algorithms for Clustering* ,Data.Prentice Hall, 1988.
[14] S. Zhang, Z. Qin, C. X. Ling, S. Sheng "Missing is Useful": Missing Values in Cost-Sensitive Decision Trees, *IEEE Transactions On Knowledge And Data Engineering*, 17(12), 2005.
[15] T. M. Cover, P. E. Hart, "Nearest neighbor pattern classification", *IEEE Transactions on Information Theory*, vol. 13, pp. 21–27, 1967.