CHAPTER 5

DISCEND – AN IMPUTATION METHOD

5.1 PROBLEM STATEMENT

The classifier performance will degrade if there are unknown values in the dataset. Some estimation methods are used to replace the unknown values with the predicted values. Several methods have been proposed to deal with the missing values. In this chapter, discretization based imputation method (DISCEND) is proposed which can increase the relevancy between the instances and attributes. Evaluating the classifiers' performance is done using four datasets. Accuracy of the classifier and computational time are the parameters taken for examining the classifiers' performance. The datasets are taken from the UCI ML repository.

5.2 INTRODUCTION

In Mean imputation method, mean value is calculated by using all known values of the attribute in each class and used to replace the unknown values. Let us consider that the value x_{ij} of the k-th class, C_k, is missing, then it will be replaced by

$$\widehat{x_{ij}} = \sum_{i:x_{ij} \in C_k} \frac{x_{ij}}{n_k} \quad (5.2.1)$$

where n_k represents the number of non-missing values in the j-th feature of the k-th class.

In k-mean imputation, mean value is calculated by using all the values in each cluster and used to replace the unknown values. Let the value x_{ij} of the m-th class, C_{km}, be missing in the k-th cluster, then it will be substituted by

$$\widehat{x_{kij}} = \sum_{i:x_{kij} \in C_{km}} \frac{x_{kij}}{n_{km}} \quad (5.2.2)$$

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

The technique used in the existing methods does not yield the accurate value, so in this chapter the missing values are imputed while discretizing the continuous attributes.

Drawbacks of the existing methods

- Some information related to class will be lost while deleting the missing values (Little and Rubin, 2002).
- Using single mean value to replace all missing values will affect the variance. (Mehala et al. 2009, Xiaofeng Zhu et al. 2011).

In section 5.3, the proposed concept based on discretization method is explained. Experimental evaluation and the comparison results are described in section 5.4. Conclusion is described in section 5.5.

5.3. PROPOSED SYSTEM

Imputation using Discretization

Let D={d_1,d_2,d_3,...,d_n} be the dataset where *n* is the number of data and let the attributes be A={A_1,A_2,A_3,…,A_m}where *m* is the number of attributes.

The proposed system consists of two phases. In the first phase, for each attribute, the data are sorted. Initial cutting points were found out between each pair of the instances in the attribute where the two consecutive values have different class value. The second phase is to fill the missing values by taking the mean value within each interval for each class instead of finding the mean value of the entire non missing values in the dataset. Mean values are calculated in each interval for each class. Then the minimum mean value from each class is used to fill the missing values corresponding to that class. This will increase the relevancy between the instances and attributes.

Let the missing values be x_{ij} of the k-th class, C_k in the m-th interval. The mean value in each interval is calculated by

$$\widetilde{x_{ij}} = \sum_{i:x_{ij} \in C_{km}} \frac{x_{ij}}{n_m} \quad (5.3.1)$$

Then x_{ij} is given by minimum value of all the mean values

$$x_{ij} = \min \{ \widetilde{x_{ij}} \} \quad (5.3.2)$$

where n_m denotes the total number of filled in values in the j-th feature of the m-th interval for each k-th class.

The filled in dataset is used in the second phase to classify the data using C4.5 and Naïve Bayes classifiers and the performance of the classifiers are analyzed using two metrics; accuracy and the computational time.

In DISCEND algorithm, the training dataset *D* is sorted for each feature in feature set F_i. The maximum and minimum value is found out and the boundary points are initialized. If the interval has missing value, then the mean value is calculated with the data in the interval for each class. The minimum mean value is used to fill the missing value corresponding to each class.

Algorithm 5.1 : DISCEND algorithm

Let D be the training dataset with continuous features F_i; S classes.

For every F_i do:

Phase 1

Step 1

1.1 Find highest (d_n) and lowest (d_o) values

1.2 sort all distinct values of F_i in ascending order

1.3 Find the cut points where the values have different class values and initialize all possible interval boundaries $B=\{[d_0,d_1][d_1,d_2],....,[d_{n-1},d_n]\}$

Step 2

2.1 For every interval $[d_i,d_j]$ where "i" is the minimum value and "j" is the maximum value,

2.2 Find the mean value corresponding to a single class value

2.3 Find the minimum mean value corresponding for each class C_k.

2.4 Fill the missing values of each class C_k with the minimum mean value of the same class C_k.

Phase 2 Step 3

3.1 Calculate the misclassification rate, accuracy and computational time.

End

5.4. EXPERIMENTAL ANALYSIS

Four datasets are taken from the Machine Learning Database UCI ML Repository to carry out the experiments. The datasets are Diabetes, Breast Cancer, Lung Cancer and Iris datasets. Table 5.4.1 describes the information such as number of instances and the number of attributes in the datasets taken for experiment. The main aim of the experiments carried out in this work is to examine whether C4.5 classification algorithm and Naïve Bayes classifier are efficient. In these experiments, unknown values are artificially imputed in different rates in different attributes. Few values are removed from the datasets randomly at the rate of 2% to 4%.

Table 5.4.1 : Datasets used for analysis

Datasets	**Instances**	**Attributes**
Diabetes	768	9
Iris	150	5
Breast Cancer	699	10
Lung Cancer	32	57

5.4.1 Result Analysis and Discussion

A. Accuracy Performance

The experimental analysis is done based on the accuracy performance and the results obtained using C4.5 classifier are shown in the table 5.4.1.1 given below.

Table 5.4.1.1 : Accuracy performance of four imputation methods using C4.5 Classifier

Dataset/ Methods	Ignore Method	Most Often Method	Mean Method	Proposed (DISCEND)
Breast Cancer	94.09	93.99	93.99	94.71
Lung Cancer	85	90.63	88.77	90
Diabetes	72.43	72.92	72.65	74.22
Iris	93.07	95.33	95.33	96.33

The results obtained using Naïve Bayes classifier is shown in table 5.4.1.2 given below

Table 5.4.1.2 : Accuracy performance of four imputation methods using NB Classifier

Dataset/ Methods	Ignore Method	Most Often Method	Mean Method	Proposed (DISCEND)
Breast Cancer	95.09	95.42	94.22	95.42
Lung Cancer	83.31	81.25	79.53	80
Diabetes	73.99	74.65	74.3	75.52
Iris	94.06	92.67	92.67	92

If the missing value rate increases, the prediction accuracies of all the classifiers obviously will trend to decrease. The tables show that on an average, both the Naïve Bayes classifier and C4.5 classifier works well for the proposed method (Discend).

From the tables, it is also clear that C4.5 classifier gives good result for multiclass dataset and also shows improved accuracy for small dataset. Naïve Bayes classifier is suitable for huge volume of data with high rate of unknown values. Naïve Bayes classifier is found to be less sensitive to missing values when compared with the C4.5 classifier. The proposed method works well for the breast cancer dataset and Iris dataset which is a multi class dataset.

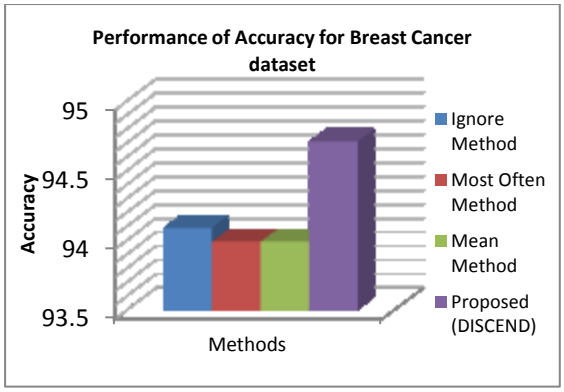

Figure 5.4.1.1a

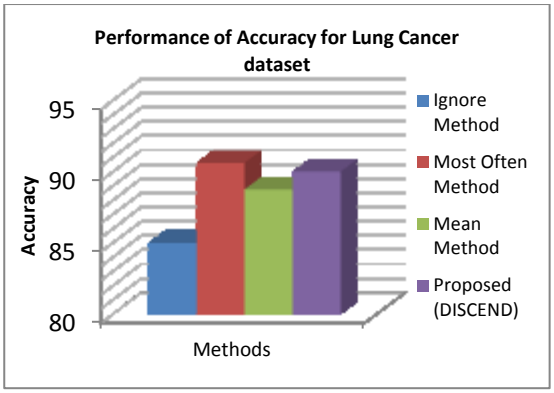

Figure 5.4.1.1b

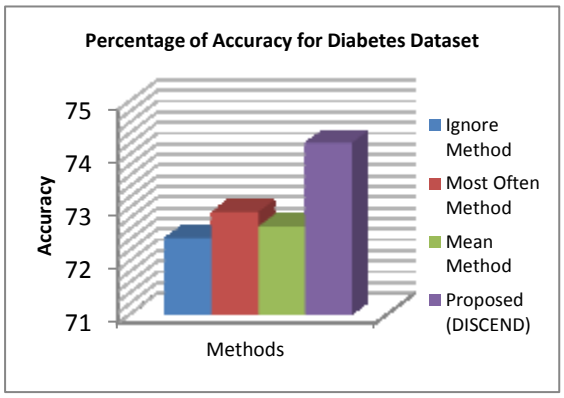

Figure 5.4.1.1c

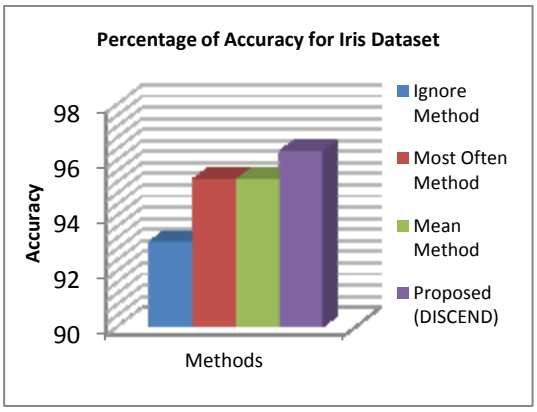

Figure 5.4.1.1d

Figure 5.4.1.1 (a-d) : Comparison result of C4.5 using imputation method

Figure 5.4.1.1 (a-d) shows the accuracy comparison of the four methods using four datasets. The accuracy for Diabetics, Breast and Lung cancer datasets seems to be slightly improved.

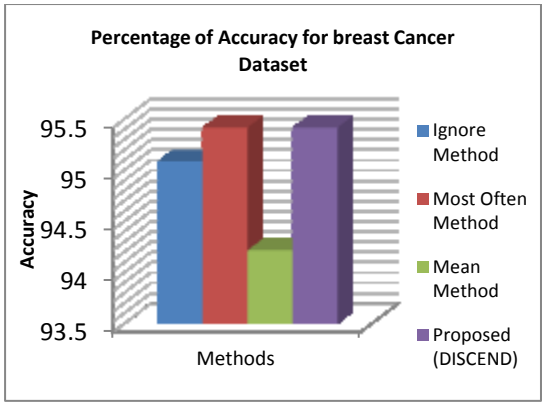

Figure 5.4.1.2a

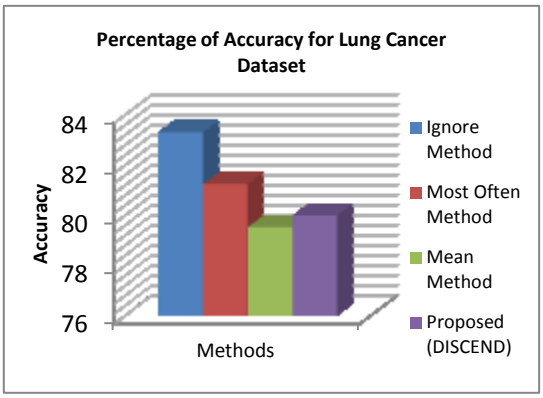

Figure 5.4.1.2b

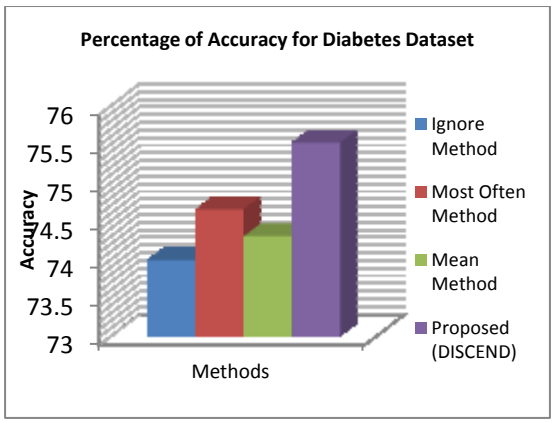

Figure 5.4.1.2c

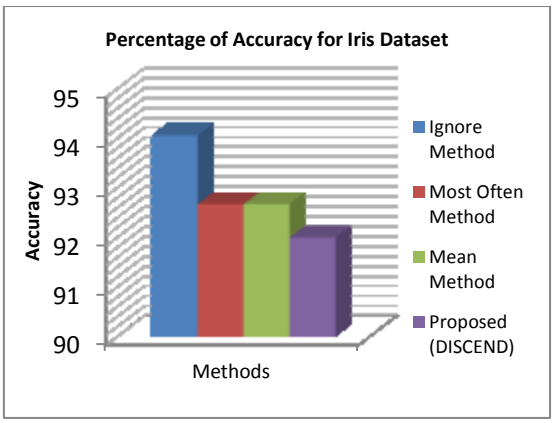

Figure 5.4.1.2d

Figure 5.4.1.2 (a-d) : Comparison result of NB using imputation method

Figure 5.4.1.2 (a-d) shows the accuracy comparison of the four methods using four datasets. The accuracy for Breast cancer and Diabetes datasets seems to be slightly improved.

B. Computational time

Let '*n*' be the set of instances and '*m*' be the set of instances in each interval '*i*'. Let '*s*' be the number of classes. Then the time taken to find the lowest and highest values is O(*n*). Time taken to do sorting is O(log(*n*)). Time

taken to find the mean value in each interval is O(i*m). Finally, time taken to find the minimum mean value is O(s*$log(n)$). So, the total computational time of the proposed algorithm (DISCEND) (excluding the single unit of time for some statements) is O(s*i*$log(n)$). Mean method fills the missing value by taking mean of all instances in each class. The computational time of the mean method is O(s*n).

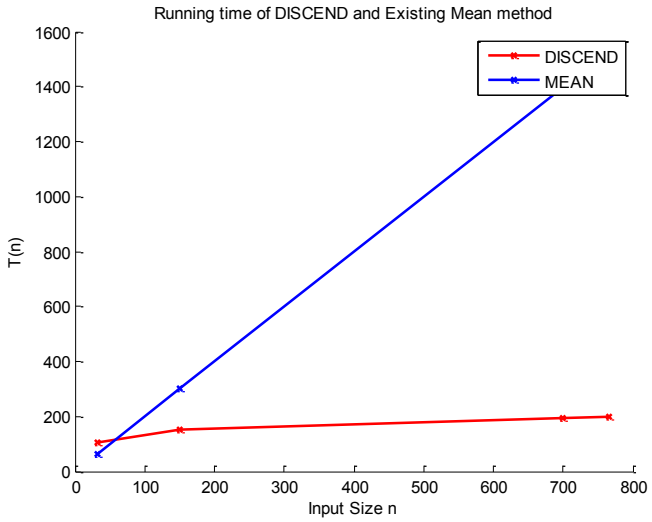

Figure 5.4.1.3: running time (in ms) of DISCEND and Mean method

5.5. SUMMARY

The proposed method appears to yield increased classification rate than the existing methods from the comparison done. MatLab 7.0.1 tool is used to conduct the experiment for filling the unknown values and the classifiers performance was examined using Weka 3.6. Incomplete data results in high misclassification rate, so this must be resolved before mining process. This work examines the classification performance of the C4.5 and Naïve Bayes classifiers.

Only numerical attributes are taken to impute the unknown values. As a future work, the same concept may be made appropriate for the categorical attributes. The proposed and existing methods are compared on the basis of accuracy and computational time. Filling the unknown values within the same class increases the relevancy between the instances and the attributes.