CHAPTER 7

EXTENDED RELIEF-DISC TECHNIQUE

7.1 PROBLEM STATEMENT

The most predominant method in preprocessing technique is feature selection method. Some of the RELIEF based algorithms are considered as the most successful algorithms for estimating the quality of features. RELIEF-DISC which was shown to be efficient in estimating features, cannot handle incomplete data for continuous features. A new algorithm Extended RELIEF-DISC is proposed and implemented to deal with the noisy and incomplete datasets. This chapter investigates the performance of the decision tree classifier and Naïve Bayes classifier by imputing the missing values. The datasets are retrieved from the UCI ML Repository.

The methods for filling the unknown values are classified into three categories as stated by Mehala et al. (2009); the simplest method of neglecting or removing the data, Parameter estimation where parameters are estimated using maximum likelihood procedures and Imputation methods, where unknown values are filled in with predicted value. The main aim of this chapter is to fill the unknown values and to choose the significant features for improving the classifiers performance. The rest of the chapter is structured as follows. Section 7.2 gives the proposed concept. Experimental analysis and the

comparative result analysis and discussion are described in section 7.3. Conclusion is described in section 7.4.

7.2 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 continuous feature be F={F_1,F_2,F_3,..F_m}where *m* is the number of features. For each feature, the data are sorted. The cut point is made between each pair of values in the features if they belong to two different classes. These cut points forms the boundary intervals. Next step is to find the mean value within each interval. In each class, the lowest mean value is used to fill the unknown values. Then the relevant features are selected by applying the concept of RELIEF-DISC. This is depicted in figure 7.2.1.

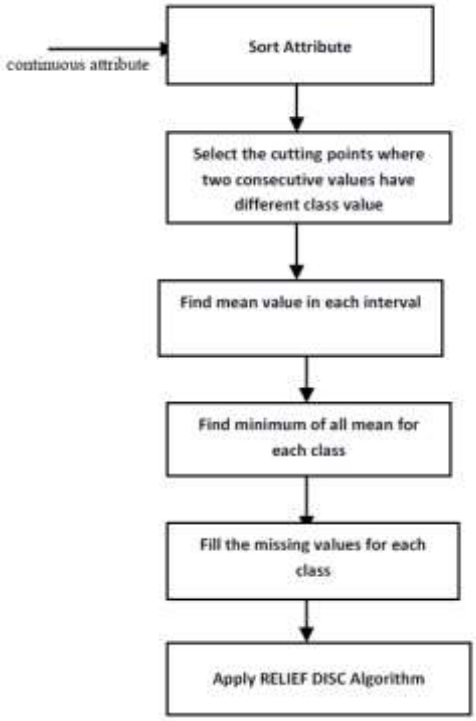

Figure 7.2.1 : Flowchart of Extended RELIEF_DISC algorithm

Algorithm 7.1 : Extended RELIEF-DISC algorithm

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

For every F_i do:

Step 1

1.1 Find lowest (d_n) and highest (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 value 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 lower bound and j is the upper bound, find the mean value corresponding to a single class value

$$\widehat{x_{ij}} = \sum_{i:\, x_{kij} \in C_{km}} \frac{x_{kij}}{n_{km}} \tag{7.2.1}$$

2.2 Find the lowest mean in each class C_k.

$$x_{ij} = min_{x_{ij} \in C_k} \widehat{x_{ij}} \tag{7.2.2}$$

2.3 Choose the lowest mean value to fill the unknown values in each class C_k.

Step 3

3.1 Select the relevant features using RELIEF-DISC algorithm.

End.

7.3 EXPERIMENTAL ANALYSIS

Four datasets are taken from UCI Machine Learning Database repository to carry out the experiment. The number of values and the number of features used are depicted in table 7.3.1. The main aim of the experiments conducted in this work is to analyze the efficiency of the C4.5 classification algorithm and Naïve Bayes classifier after selecting the relevant features. In this experiment, unknown values are synthetically assigned in different rates in different attributes. Datasets without unknown values are taken and some values are removed from it randomly. The rates of the unknown values removed are from 2% to 4%.

Table 7.3.1 : Datasets used for analysis

Datasets	**No. of Instances**	**No. of Attributes**
Diabetes	768	9
Iris	150	5
Breast Cancer	699	10
Lung Cancer	32	57

7.3.1 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 and Naïve Bayes classifier are shown in the table 7.3.2 given below.

Table 7.3.2: Accuracy performance using Naïve Bayes and C4.5 Classifiers

Dataset / Methods	RELIEF		RELIEF-DISC		Extended RELIEF-DISC	
	NB	C4.5	NB	C4.5	NB	C4.5
Breast Cancer	95.42%	94.42%	96.14%	95.42%	96.42%	96.34%
Lung Cancer	83.31%	85%	81.25%	90.63%	84.23%	92.35%
Diabetes	76.17%	74.12%	76.43%	74.35%	76.43%	74.42%
Iris	96%	96%	92.67%	95.33%	92%	95.37%

Result and Discussion

From the table above, the integrated proposed method Extended RELIEF-DISC works well for both Naïve Bayes and C4.5 classifiers. Naïve Bayes classifier outperforms C4.5 for breast cancer dataset and Diabetes dataset. C4.5 works well for multi class dataset. The experiment shows that the C4.5 classifier does not perform well for large dataset while the Naïve Bayes classifier can perform well for huge volume of dataset.

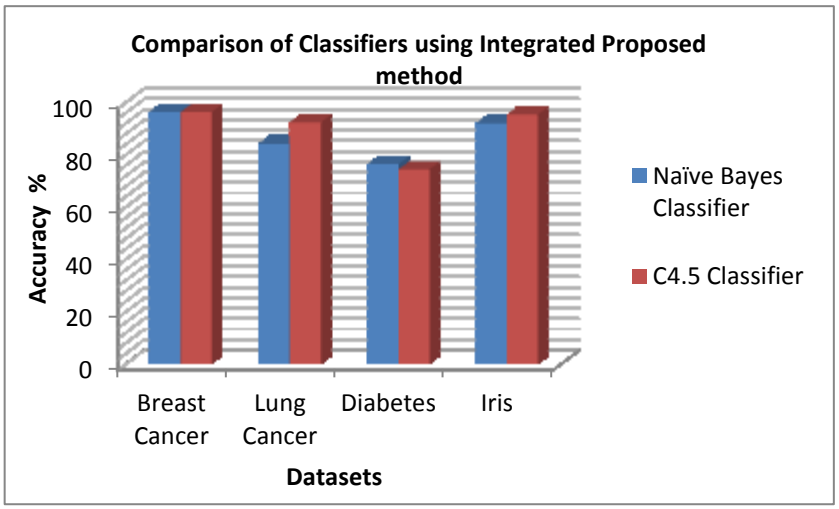

Figure 7.3.1 : Comparison of C4.5 and Naïve Bayes using Integrated method

Figure 7.3.1 depicts the accuracy performance of C4.5 classifier and Naïve Bayes classifier for four datasets by implementing using integrated proposed method.

B. Computational Time

For the proposed method Extended RELIEF-DISC, the computational time of step1 is same as in phase2. For RELIEF-DISC, the time of finding the feature weight-age is O(*i*) where '*i*' is the number of intervals. The time of existing RELIEF method is O(*m*) where '*m*' is the number of instances in the sample.

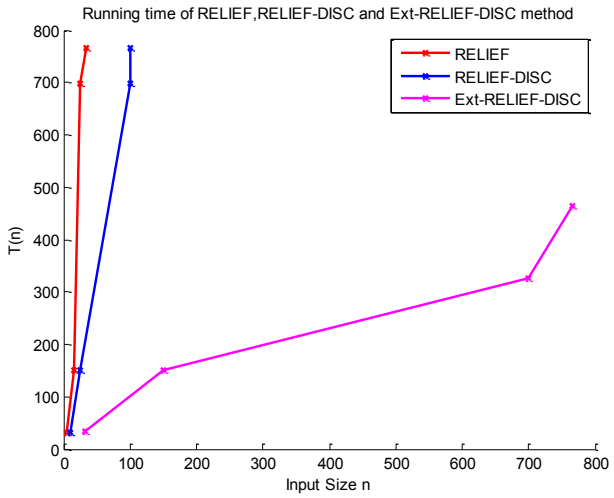

Figure 7.3.2: Running time (in ms) of RELIEF,RELIEF-DISC and Extended RELIEF-DISC method

7.4 SUMMARY

RELIEF-DISC selects the features which are relevant, but does not handle the incomplete data. But the extended RELIEF-DISC algorithm fills the missing values in the initial stage of discretization and then selects the relevant features. The general method uses the mean of the entire non missing values. But in extended RELIEFDISC, since the missing values are filled using the lowest mean value in each class, the classification rate is proved to be more and efficient.

The experiment for extended RELIEF-DISC was performed using MatLab 7.0.1 and the classifier performance was examined using Weka 3.6. Unknown values in the dataset may affect the classification rate. It must be filled in before using these datasets for classification. This work analyses the classification performance of the C4.5 classifier and Naïve Bayes classifier. The proposed concept uses only the numerical features to fill the unknown values. Handling of categorical attributes can be extended as the future work.