# Chapter 7

# Unsupervised Feature Selection

Feature Selection (FS) is a remedy that can reduce the dimensionality of data without degrading the computational efficiency of the problem. Conventional feature subsets are selected based on some evaluation metric that is estimated based on its dependency or relation to decision class value in the data. However, not all the real world problems have label or class attached to them thereby creating the need for FS in unsupervised domain. FS in unsupervised domain, in layman terms, selects those feature subsets that produce best groupings or clusters among individuals [42]. Some of the unsupervised approaches that utilized clustering as a criteria to evaluate feature subsets discussed previously were based on expectation maximisation [41], sequential FS [30], entropy based techniques [7], FS based on genetic algorithm [84], multi cluster feature selection [19]. Similarity based unsupervised FS was employed by Mitra et. al. [109] to avoid redundancy. Unsupervised FS was proposed using ant colony optimization technique in [139]. Adaptive and embedding learning based unsupervised FS adaptively learned the embeddings that preserved manifold structure was successfully used by Wu [161] while Liu et. al. [97] employed neighbourhood embedding for FS. Lim et. al. [93] employed pairwise dependence among features for FS. Principal component and Linear Discriminant Analysis change the semantics of data while performing dimensionality reduction. However, all these approaches ignore the fuzziness existing in the dataset.

Fuzzy rough set theory offers a methodological solution to reduce the dataset dimensionality by selecting the most informative features. It efficiently reduces the dataset size as only set operations are involved, computational overload is also reduced. The entire process is accomplished without requiring any human intervention or any prior information. Most importantly, the underlying semantics of the dataset is preserved. Only some researchers have worked on FS using fuzzy rough set theory in unsupervised learning case, where class labels are unknown or missing. Parthalain et. al. [116] employed fuzzy rough set theory technique for FS on datasets without class information. Wang et. al. [159] combined the idea of fuzzy and sparse learning for unsupervised FS. While embeddings of fuzzy membership was used by Zhang et. al. [172]. The performance of fuzzy rough set theory model is dependent on the quality of feature subsets selected for evaluation while maintaining the computation time to a low value. However, none of the researchers have discussed this issue. Swarm intelligence models the social behaviour of animals by having a population of artificial agents that perform simple task while co-operatively solving hard optimization problem. This chapter proposes the novel method of selecting relevant, non redundant high quality and information rich subset of features employing a metaheuristic earthworm optimization thereby taking into consideration the fuzziness existing in the real world datasets.

## 7.1 Feature Selection based on Fuzzy Rough Set in Unsupervised Domain

In case of supervised approach, a decision class is associated with each instance. Based on the value of degree of dependency, the quality of feature subset can be evaluated. The reduct comprises of the minimal feature subset that maintains the quality of original dataset. Any search technique like greedy hill climbing algorithm can be applied to construct feature subset.

Extending the idea to unsupervised domain, decision class $D$ can be replaced by non-intersecting subset of features $\overline{B}$ i.e $A \bigcap B = \phi$ to compute dependency of $A$ as:

$$\gamma_A(B) = \frac{\sum_{x_i \in U} Pos_A(B)(x_i)}{\mid U \mid} \tag{7.1}$$

The positive region and thereby the lower and upper approximations [116] are defined as:

$$Pos_A(B)(x_i) = sup_{z \in U} R \downarrow_A R_{B_z}(x_i) \tag{7.2}$$

where $R_{B_z}$ is the fuzzy similarity relation for sample $z$ (say).

$$R \downarrow_A R_{B_z}(x_i) = inf_{x_j} I(R_A(x_i, x_j), R_B(x_j, z)) \tag{7.3}$$

$$R \uparrow_A R_{B_z}(x_i) = sup_{x_j} T(R_A(x_i, x_j), R_B(x_j, z)) \tag{7.4}$$

The lower approximation describes the certainty with which an instance belongs to a set. The upper approximation gives the possibility of an object belonging to a set. Usually, lower approximation is used to evaluate the quality of feature subsets, while the information contained is upper approximation is not considered. For two subsets having same lower approximation, the one containing lower upper approximation is more accurate reflection of original content. Therefore, the information contained in upper approximation should be utilized, boundary region (which is given by difference between upper and lower approximation) is such a measure that undertakes both lower and upper approximation into account. The value of boundary region decreases as search progresses towards optimal subset position until the lowest value is reached. The total certainty degree is employed instead of dependency degree to evaluate feature subset quality as the following:

$$\gamma_A(B) = 1 - \frac{\sum_{z\in U}\sum_{x_i\in U} BND_A(B)(x_i)}{\mid U \mid^2} \tag{7.5}$$

This measure can then be used to guide the feature selection task.

### 7.1.1 Feature subset quality evaluation

The quality of feature subset must be determined to guide the search towards optimal feature subset selection. The quality of subset of features is examined using dependency and boundary region measure in this work.

#### 7.1.1.1 Dependency measure

Let $A$ be the subset of attributes. if the non-selected feature subset $B = C - A$ depends on the selected subset $A$, then $B$ can be effectively eliminated as the information content in $B$ would be redundant. Dependency degree can be utilized to compute dependence of $B$ on $A$ using the following formula:

$$\gamma_A(B) = \frac{\sum_{x_i\in U} Pos_A(B)(x_i)}{\mid U \mid} \tag{7.6}$$

The positive region is defined in the same way as equation (7.2) considering that each object belongs to its own class:

$$Pos_A(B)(x_i) = R_A \downarrow R_B(x_i) \tag{7.7}$$

The modified lower and upper approximation are defined as:

$$R_A \downarrow R_B(x_i) = inf_{x_j} I(R_A(x_i, x_j), R_B(x_i, x_j)) \tag{7.8}$$

$$R_A \uparrow R_B(x_i) = sup_{x_j} T(R_A(x_i, x_j), R_B(x_i, x_j)) \tag{7.9}$$

On the similar lines, boundary region measure can also be defined as:

$$BND_A(B)(x_i) = R_A \uparrow R_B(x_i) - R_B \downarrow R_B(x_i) \tag{7.10}$$

The corresponding measure to evaluate feature quality is given as follows:

$$\gamma_A(B) = 1 - \frac{\sum_{x_i \in U} BND_A(B)(x_i)}{\mid U \mid} \tag{7.11}$$

We illustrate the concept via a toy example shown in Table 7.1. Here, standard t-norm and implicator are applied for the computation. The similarity measure given below equation is employed for this example.

$$\begin{aligned} R_a(x_i, x_j) = \max(\min(&\frac{(a(x_j) - (a(x_i) - std_a))}{std_a}, \\ &\frac{((a(x_i) + std_a) - a(x_i))}{std_a}, 0)) \end{aligned}$$

where $std_a$ is the standard deviation of feature $a$. Suppose features $A = a_1, a_5, a_6$ are selected, then the similarity values between each pair of instances are computed and shown in Table 7.2. On similar lines, the similarity values of non-selected features $B = a_2, a_3$ are noted down in Table 7.3. Based on these values, the lower approximation is evaluated as:

$$\begin{aligned} R_A \downarrow R_B(x_1) = inf(&I(1,1), I(0, 0.801175), I(0, 0.801175), \\ &I(0, 0.92047), I(0.242991, 0), I(0, 0.045641), \\ &I(0,0), I(0,0)) = 0.7570 \\ R_A \downarrow R_B(x_2) = inf(&I(0, 0.801175), I(1,1), I(0, 0.602351), \\ &I(0, 0.880705), I(0,0), I(0,0), I(0,0), I(0,0)) = 1 \end{aligned}$$

Similarily, the method is iterated for all the instances and the corresponding positive region is calculated as:

$Pos_A(B)(x_3) = R_A \downarrow R_B(x_3) = 0.9147,$

$Pos_A(B)(x_4) = \overline{R_{\overline{C}}} \downarrow R_B(x_4) = 1,$

$Pos_A(B)(x_5) = R_A \downarrow R_B(x_5) = 0.7570,$

$Pos_A(B)(x_6) = \overline{R_{\overline{C}}} \downarrow R_B(x_6) = 0.9147,$

$Pos_A(B)(x_7) = R_A \downarrow R_B(x_7) = 1,$

$Pos_A(B)(x_8) = \overline{R_{\overline{C}}} \downarrow R_B(x_8) = 1$

Table 7.1: Example dataset

| **Instances**<br>**Features** | $a_1$ $a_2$ | | $a_3$ | $a_4$ | $a_5$ |
|---|---|---|---|---|---|
| $x_1$ | 0.08 | 0.08 | 0.1 | 0.24 | 0.9 |
| $x_2$ | 0.06 | 0.06 | 0.05 | 0.25 | 0.33 |
| $x_3$ | 0.1 | 0.1 | 0.15 | 0.65 | 0.3 |
| $x_4$ | 0.08 | 0.08 | 0.08 | 0.98 | 0.24 |
| $x_5$ | 0.09 | 0.15 | 0.4 | 0.1 | 0.66 |
| $x_6$ | 0.15 | 0.02 | 0.34 | 0.4 | 0.01 |
| $x_7$ | 0.24 | 0.75 | 0.32 | 0.18 | 0.86 |
| $x_8$ | 0.276 | 0.225 | 0.81 | 0.27 | 0.33 |

Table 7.2: Fuzzy similarity values for selected features

| $R_A$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ |
|---|---|---|---|---|---|---|---|---|
| $x_1$ | 1 | 0 | 0 | 0 | 0.242991 | 0 | 0 | 0 |
| $x_2$ | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| $x_3$ | 0 | 0 | 1 | 0 | 0 | 0.08528 | 0 | 0 |
| $x_4$ | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| $x_5$ | 0.242991 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| $x_6$ | 0 | 0 | 0.08528 | 0 | 0 | 1 | 0 | 0 |
| $x_7$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| $x_8$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Table 7.3: Fuzzy similarity values for non-selected features

| $R_B$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ |
|---|---|---|---|---|---|---|---|---|
| $x_1$ | 1 | 0.801175 | 0.801175 | 0.92047 | 0 | 0.045641 | 0 | 0 |
| $x_2$ | 0.801175 | 1 | 0.602351 | 0.880705 | 0 | 0 | 0 | 0 |
| $x_3$ | 0.801175 | 0.602351 | 1 | 0.721645 | 0.005876 | 0.244466 | 0 | 0 |
| $x_4$ | 0.92047 | 0.880705 | 0.721645 | 1 | 0 | 0 | 0 | 0 |
| $x_5$ | 0 | 0 | 0.005876 | 0 | 1 | 0.452148 | 0 | 0 |
| $x_6$ | 0.045641 | 0 | 0.244466 | 0 | 0.452148 | 1 | 0 | 0 |
| $x_7$ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| $x_8$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Hence, the value of dependency degree is given as:

$$\gamma_A(B) = \frac{\sum_{x_i \in U} Pos_A(B)(x_i)}{\mid U \mid} = \frac{7.3434}{8} = 0.9179 \tag{7.12}$$

### 7.1.2 Search strategy

The previous section sets the platform for evaluating the quality of feature subset. There is a need for a search strategy to select optimal feature subset that would lead towards good performance. This is advantageous as whole dataset need not be evaluated for achieving good performance.

Binary earthworm optimization proposed in this study, is inspired by reproductory behaviour of earthworms. Each earthworm produces two kinds of offsprings using variants of reproduction namely reproduction 1 and reproduction 2. The child earthworm is of same size as that of parent. Consider the population of $N$ earthworms. Each individual earthworm is represented by $m$ dimensional binary vector representing presence or absence of attributes with $m$ being the total number of attributes in the dataset. Let the $i^{th}$ earthworm in generation $t$ be denoted by $e^{i,t}$. For moving on to next generation, the earthworms reproduce using reproduction 1 or 2 to produce offsprings. The two reproductions are described as follows:

#### 7.1.2.1 Reproduction 1

Earthworms are hermaphrodites allowing a single parent to generate child earthworm. The offspring is generated in the following way as:

$$e_1^{i,t+1} = 1 - \alpha_o e^{i,t} \tag{7.13}$$

where $e^{i,t}$ denotes the $i^{th}$ earthworm in the present generation $t$. Likewise, $e_1^{i,t+1}$ denotes the newly generated $i^{th}$ earthworm in generation $t+1$. To allow the value at any position of earthworm to oscillate in 0 and 1, the parameter $\alpha_o$ is chosen such that $\alpha_o \in [0, 1]$ indicating the similarity between child and parent. If $\alpha_o = 0$, the child earthworm has all the values as 1 while when $\alpha_o = 1$ the newly generated earthworm has all the features that are not possessed by parent earthworm. Therefore, the value of $\alpha_o$ is set to 1 for further experimentation. This leads to global search promoting exploration.

#### 7.1.2.2 Reproduction 2

Reproduction 2 is an improvised version of crossover operator. In this paper, multi-point crossover is employed wherein two earthworms are used to generate a offspring. The two parents $e^{p_1,t}$ and $e^{p_2,t}$ are randomly selected from the population. The $i^{th}$ child earthworm $e_2^{i,t+1}$ is generated using the following equation:

$$e_2^{i,t+1} = [e^{p_1,t}(1:r_1), e^{p_2,t}(r_1+1:r_2), e^{p_1,t}(r_2+1,m)] \tag{7.14}$$

where $r_1$ and $r_2$ are two randomly generated integers lying in range $(1,m)$.

Using these two kinds of reproduction, the $i^{th}$ earthworm is generated as:

$$e^i = \begin{cases} 1, \beta_o e_1^{i,t+1} + (1-\beta_o)e_2^{i,t+1} \geq 0.5 \\ 0, \beta_o e_1^{i,t+1} + (1-\beta_o)e_2^{i,t+1} < 0.5 \end{cases} \tag{7.15}$$

where the parameter $\beta_o$ is used to adjust the contributions of the two reproductions. The value of $\beta_o$ at $t+1$ generation is updated using:

$$\beta_o^{t+1} = \sigma\beta_o^t \tag{7.16}$$

where $\sigma$ is like a cooling factor and has a constant value. For $t = 0$, initial value of $\beta_o$ is set to 1. The value of $\beta_o$ is decreased as the number of generation increases implying that the contribution from reproduction 1 decreases with increment of generations and the impact of reproduction 2 becomes more and more dominant. Thus, local search is more significant when the optimization technique is about to reach the end of generations maintaining the balance between exploration and exploitation. Further, in order to head towards optimal solution the worst fitness earthworm at current generation is replaced with best fitness earthworm of previous generation. The fitness of individual earthworm is formulated using the following equation:

$$Fit_{e^i} = \alpha \times \gamma_A(B) + \beta \times \frac{\mid m \mid - \mid A \mid}{\mid m \mid} \tag{7.17}$$

where $\mid m \mid$ and $\mid A \mid$ denote the number of total and selected features respectively. $\alpha$ and $\beta$ are the constants governing the importance of classification performance and subset length respectively, such that $\alpha = 1 - \beta, \alpha \in [0, 1]$. The entire approach is described in Algorithm 7.1.2.2 and illustrated via Figure 7.1.

**Algorithm 7.1.2.2 Earthworm search strategy**

```
Input: $generation$: number of generation; $N$: number of earthworms; $\alpha_o$: similarity parameter of reproduction 1; $\beta_o$: adjustment parameter used for varying the impact of two reproductions; $\sigma = 0.9$: cooling factor used for computing $\beta$;
  $e^i = m$ bit vector generated randomly; $i = 1, 2, \ldots, N$
  Evaluate the fitness of each earthworm
  $t = 0$;
  while $t < generation$ do
    for $\forall$ earthworm $e^{i,t}$ do
      # Implement Reproduction 1
      Generate $e_1^{i,t+1}$ using Reproduction 1 using Equation (7.13)
      # Implement Reproduction 2
      Randomly select two parent earthworms $e^{p_1,t}$ and $e^{p_2,t}$
      Generate offspring $e_2^{i,t+1}$ using Equation (7.14)
      Update the $i^{th}$ earthworm at $t+1$ generation using Equation (7.15)
    end for
    Replace the worst earthworm at $t+1$ generation with the best fitness earthworm at $t$ generation
    $t \leftarrow t + 1$;
  end while
```

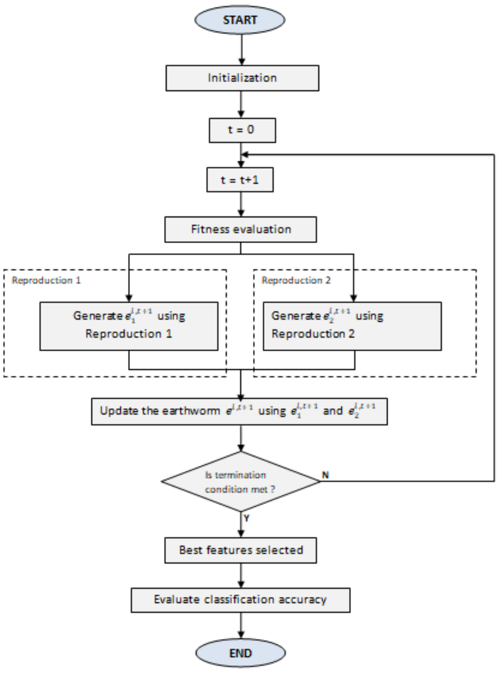
START
Initialization
t = 0
t = t+1
Fitness evaluation
Reproduction 1
Generate $e_1^{i,t+1}$ using Reproduction 1
Reproduction 2
Generate $e_2^{i,t+1}$ using Reproduction 2
Update the earthworm $e^{i,t+1}$ using $e_1^{i,t+1}$ and $e_2^{i,t+1}$
Is termination condition met ?
N
Y
Best features selected
Evaluate classification accuracy
END

FIGURE 7.1: The flowchart of entire methodology for obtaining a reduced representation of the dataset

```
**return** Best fitness earthworm
```

## 7.2 Experimentation

The effectiveness of the proposed approach can be illustrated by conducting proper experimentation for the same. The various parameter settings employed for experimentation are as follows: The parameters $\alpha$ and $\beta$ used in fitness calculation are set to 0.9 and 0.1 respectively. The similarity factor $\alpha_o = 1$, the initial proportional factor $\beta_o$ is set to 1 and cooling factor $\sigma$ as 0.9. The value of generation and the number of earthworms are set to 50. These default values as given in [156] are used for further experimentation.

### 7.2.1 Results

Sixteen different benchmark datasets used for experimental evaluation are taken from UCI repository [3] (as shown in Table 7.4). Performance of proposed approach and its variants is illustrated by splitting the experimentation into four sections namely:

1. The proposed approach (UFRESO) and its variant based on boundary measure (UFRBESO).
2. Comparative analysis of the proposed work with existing dependency based approach.
3. Comparison with previous non dependency based approaches.
4. Applying feature selection taking the class labels into consideration i.e. supervised approach based on fuzzy rough sets and performing the comparative study of the same. It might be unreasonable to compare with supervised approaches as class labels brings in additional information. However, that could help in verifying the fact that the valuable information are retained in case of unsupervised approach also.

The comparison is done in aspects of reduct size, classification accuracy and classification error. Ten fold cross validation is used for performance computation. So, the results are averages of 10 folds. The significance are established using statistical tests namely two tailed studentś t test by calculating difference between average accuracy of two models statistically. The level of significance is set to 0.05 for the experiment. The probability p-val associated with t test is also depicted in the tables. The small p-val values (less than 0.05) illustrates the significant differences amongst the algorithms. Various symbols like +, -, o are used to show corresponding statistical win,

Table 7.4: Benchmark dataset

| Dataset | Instance | Feature | Class | Classification accuracy | | | |
|---|---|---|---|---|---|---|---|
| | | | | KNN | | SVM | |
| | | | | Acc | Std | Acc | Std |
| **Auto-univ-au1_1000** | 1000 | 20 | 2 | 71.70 | 4.21 | 74.10 | 3.57 |
| **German** | 1000 | 24 | 2 | 70.80 | 6.35 | 76.60 | 3.53 |
| **Cardiotocography-3class** | 2126 | 36 | 3 | 98.86 | 0.71 | 92.07 | 10.37 |
| **Diabetes** | 768 | 8 | 2 | 72.50 | 5.70 | 64.86 | 6.35 |
| **Leaf** | 340 | 14 | 30 | 66.47 | 5.03 | 48.23 | 9.32 |
| **Abalone-11class** | 3842 | 9 | 11 | 23.20 | 2.10 | 27.11 | 1.98 |
| **Heart-cleveland** | 303 | 14 | 5 | 55.12 | 6.74 | 57.00 | 6.13 |
| **Hepatitis** | 155 | 19 | 2 | 81.33 | 7.56 | 84.00 | 7.16 |
| **Ionosphere** | 351 | 34 | 2 | 84.28 | 4.71 | 88.00 | 5.00 |
| **Fertility-diagnosis** | 100 | 9 | 2 | 88.00 | 1.35 | 88.00 | 12.29 |
| **Flags-religion** | 194 | 28 | 8 | 48.42 | 10.46 | 44.21 | 9.98 |
| **Lung-cancer** | 32 | 57 | 3 | 43.33 | 22.49 | 46.67 | 32.20 |
| **Lymphography** | 148 | 18 | 4 | 85.00 | 10.35 | 80.71 | 12.16 |
| **Trains** | 10 | 26 | 2 | 50.00 | 52.70 | 70.00 | 48.30 |
| **Wine** | 178 | 13 | 3 | 95.29 | 6.67 | 95.29 | 5.40 |
| **Semeion** | 1593 | 257 | 10 | 90.94 | 2.26 | 93.63 | 1.91 |

loss and tie respectively of the respective approach at 5% level of significance. Two classifiers namely KNN (with 3 nearest neighbours) [91] and SVM [81] are used for learning.

#### 7.2.1.1 Using variants of proposed approach

The various performance measures i.e. reduct size, accuracy along with standard deviation and error are shown in Tables 7.5, 7.6 and 7.7 for both the dependency and boundary based measures. It could be clearly seen that there is difference in number of features selected by UFRESO and UFRBESO with latter selecting fewer features than former for most of the datasets. The accuracy measures values are nearly same as that obtained from original unreduced dataset except for leaf and lymphography for which low value is obtained which can be justified by unsupervised

Table 7.5: Number of features selected by employing variants of proposed approach

| Dataset | UFRESO | UFRBESO |
|---|---|---|
| **Auto-univ-au1_1000** | 15.1 | 6.3 |
| **German** | 17.3 | 8.5 |
| **Cardiotocography-3class** | 22.1 | 11.2 |
| **Diabetes** | 6.6 | 1.5 |
| **Leaf** | 10.9 | 3.2 |
| **Abalone-11class** | 7.1 | 1.7 |
| **Heart-cleveland** | 10.5 | 3.1 |
| **Hepatitis** | 14.3 | 5.3 |
| **Ionosphere** | 16.0 | 17.4 |
| **Fertility-diagnosis** | 8.2 | 2.1 |
| **Flags-religion** | 19.6 | 8.6 |
| **Lung-cancer** | 19.1 | 33.6 |
| **Lymphography** | 13.0 | 4.9 |
| **Trains** | 8.9 | 18.8 |
| **Wine** | 10.9 | 2.6 |
| **Semeion** | 103.9 | 149.9 |

nature of dataset. However, hepatitis, fertility-diagnosis and trains yielded higher performance than benchmark ones. On comparing UFRESO and UFRBESO, the classification accuracy of UFRESO is more or comparable with UFRBESO for all the datasets except lung-cancer. The bit lower performance may be attributed to the use of boundary (or possible instances in calculation of approximations). A similar trend could be observed for classification error also.

#### 7.2.1.2 Comparison with state of art dependency based approach

A comparative analysis of the proposed approach with existing state of art algorithm is undertaken in this section. Unsupervised fuzzy rough set based dimensionality reduction (UFRFS) [102] is used for comparison (as shown in Tables 7.8, 7.12 and 7.9). There is increase in classification accuracy for UFRESO for all datasets except diabetes and lung-cancer for which the decrease in insignificant. Less number of

Table 7.6: Classification accuracy by employing variants of proposed approach

| **Dataset** | | **UFRESO** | | **UFRBESO** | |
|---|---|---|---|---|---|
| | | **Acc** | **Std** | **Acc** | **Std** |
| **Auto-univ-au1_1000** | **KNN** | 70.73 | 2.21 | 67.86 | 1.18 |
| | **SVM** | 71.65 | 2.50 | 72.81 | 0.98 |
| **German** | **KNN** | 73.27 | 2.61 | 64.21 | 4.18 |
| | **SVM** | 77.44 | 1.48 | 71.81 | 1.08 |
| **Cardiotocography-3class** | **KNN** | 96.21 | 1.38 | 93.78 | 1.53 |
| | **SVM** | 83.18 | 3.29 | 76.87 | 7.35 |
| **Diabetes** | **KNN** | 70.14 | 2.58 | 67.48 | 0.84 |
| | **SVM** | 63.32 | 3.43 | 66.64 | 0.99 |
| **Leaf** | **KNN** | 67.74 | 4.90 | 32.49 | 3.61 |
| | **SVM** | 39.14 | 5.64 | 11.25 | 5.94 |
| **Abalone-11class** | **KNN** | 22.78 | 0.46 | 21.81 | 0.54 |
| | **SVM** | 28.74 | 1.02 | 23.46 | 1.31 |
| **Heart-cleveland** | **KNN** | 56.20 | 2.55 | 53.70 | 5.17 |
| | **SVM** | 60.28 | 2.98 | 55.58 | 1.81 |
| **Hepatitis** | **KNN** | 88.21 | 3.05 | 78.39 | 4.8 |
| | **SVM** | 91.26 | 2.38 | 76.95 | 4.11 |
| **Ionosphere** | **KNN** | 87.86 | 1.24 | 87.57 | 2.51 |
| | **SVM** | 85.23 | 0.99 | 86.63 | 2.91 |
| **Fertility-diagnosis** | **KNN** | 88.29 | 1.61 | 88.78 | 4.18 |
| | **SVM** | 93.36 | 2.54 | 91.14 | 3.67 |
| **Flags-religion** | **KNN** | 51.26 | 2.71 | 38.03 | 4.09 |
| | **SVM** | 39.90 | 1.89 | 38.02 | 4.53 |
| **Lung-cancer** | **KNN** | 37.33 | 3.69 | 57.29 | 5.42 |
| | **SVM** | 44.51 | 9.28 | 49.71 | 6.56 |
| **Lymphography** | **KNN** | 78.98 | 4.01 | 61.97 | 2.83 |
| | **SVM** | 83.89 | 1.46 | 68.13 | 3.14 |
| **Trains** | **KNN** | 72.32 | 22.12 | 74.13 | 19.11 |
| | **SVM** | 90.18 | 10.73 | 90.18 | 10.73 |
| **Wine** | **KNN** | 94.54 | 2.34 | 76.48 | 2.54 |
| | **SVM** | 93.74 | 2.66 | 82.17 | 5.14 |
| **Semeion** | **KNN** | 89.20 | 0.06 | 88.72 | 0.29 |
| | **SVM** | 89.81 | 0.05 | 92.30 | 0.79 |

Table 7.7: Classification error results for variants of proposed approach

| **Dataset** | | **Classification error** | |
|---|---|---|---|
| | | **UFRESO** | **UFRBESO** |
| **Auto-univ-au1_1000** | **KNN** | 0.29 | 0.33 |
| | **SVM** | 0.26 | 0.27 |
| **German** | **KNN** | 0.29 | 0.33 |
| | **SVM** | 0.24 | 0.28 |
| **Cardiotocography-3class** | **KNN** | 0.03 | 0.07 |
| | **SVM** | 0.17 | 0.36 |
| **Diabetes** | **KNN** | 0.28 | 0.32 |
| | **SVM** | 0.34 | 0.34 |
| **Leaf** | **KNN** | 4.07 | 7.96 |
| | **SVM** | 6.47 | 11.25 |
| **Abalone-11class** | **KNN** | 1.59 | 1.83 |
| | **SVM** | 1.39 | 1.54 |
| **Heart-cleveland** | **KNN** | 0.71 | 0.86 |
| | **SVM** | 0.63 | 0.75 |
| **Hepatitis** | **KNN** | 0.14 | 0.18 |
| | **SVM** | 0.10 | 0.20 |
| **Ionosphere** | **KNN** | 0.11 | 0.12 |
| | **SVM** | 0.14 | 0.13 |
| **Fertility-diagnosis** | **KNN** | 0.10 | 0.15 |
| | **SVM** | 0.07 | 0.12 |
| **Flags-religion** | **KNN** | 1.47 | 1.55 |
| | **SVM** | 1.79 | 1.70 |
| **Lung-cancer** | **KNN** | 0.70 | 0.50 |
| | **SVM** | 0.73 | 0.53 |
| **Lymphography** | **KNN** | 0.21 | 0.40 |
| | **SVM** | 0.17 | 0.32 |
| **Trains** | **KNN** | 0.50 | 0.30 |
| | **SVM** | 0.20 | 0.20 |
| **Wine** | **KNN** | 0.05 | 0.25 |
| | **SVM** | 0.05 | 0.24 |
| **Semeion** | **KNN** | 0.50 | 0.50 |
| | **SVM** | 0.41 | 0.27 |

Table 7.8: Number of features selected on comparison with other state of art feature selection algorithm

| **Dataset** | **UFRFS** | **UFRESO** |
|---|---|---|
| **Auto-univ-au1_1000** | 19.4 | **15.1** |
| **German** | **17.3** | **17.3** |
| **Cardiotocography-3class** | **20.6** | 22.1 |
| **Diabetes** | 8.0 | **6.6** |
| **Leaf** | **7.2** | 10.9 |
| **Abalone-11class** | 8.0 | **7.1** |
| **Heart-cleveland** | 11.0 | **10.5** |
| **Hepatitis** | 15.1 | **14.3** |
| **Ionosphere** | 19.4 | **16.0** |
| **Fertility-diagnosis** | 8.9 | **8.2** |
| **Flags-religion** | **1.0** | 19.6 |
| **Lung-cancer** | 32.1 | **19.1** |
| **Lymphography** | 15.2 | **13.0** |
| **Trains** | 10.1 | **8.9** |
| **Wine** | **6.0** | 10.9 |
| **Semeion** | **50.9** | 103.9 |

features are selected by UFRESO for most of the datasets along wtih high performance. A lower classification error is observed by UFRESO by most of the datasets. Statistical results illustrates that UFRESO has achieved better or comparable performance and it losses only for one dataset namely diabetes for KNN classifier while for SVM, UFRESO losses for three and wins for seven. The statistical results lays down the superiority of UFRESO.

#### 7.2.1.3 Comparison with state of art non dependency based approach

Four state of art algorithms namely Spectral feature selection for supervised and unsupervised learning (USFS) [175], Unsupervised feature selection for multi cluster data (UFSMC) [19], Laplacian score for feature selection (LSFS) [59] and Unsupervised Feature Selection using Feature Similarity (FSFS) [109] has been employed for comparison and the results are noted down in Tables 7.13, 7.14 and 7.15. Cai et. al.

Table 7.9: Classification error comparison with other state of art feature selection algorithm

| **Dataset** | | **Classification error** | |
|---|---|---|---|
| | | **UFRFS** | **UFRESO** |
| **Auto-univ-au1_1000** | **KNN** | 0.35 | 0.29 |
| | **SVM** | 0.26 | 0.26 |
| **German** | **KNN** | 0.34 | 0.29 |
| | **SVM** | 0.23 | 0.24 |
| **Cardiotocography-3class** | **KNN** | 0.03 | 0.03 |
| | **SVM** | 0.29 | 0.17 |
| **Diabetes** | **KNN** | 0.30 | 0.28 |
| | **SVM** | 0.33 | 0.34 |
| **Leaf** | **KNN** | 0.04 | 4.07 |
| | **SVM** | 0.06 | 6.47 |
| **Abalone-11class** | **KNN** | 0.14 | 1.59 |
| | **SVM** | 0.16 | 1.39 |
| **Heart-cleveland** | **KNN** | 0.18 | 0.71 |
| | **SVM** | 0.27 | 0.63 |
| **Hepatitis** | **KNN** | 0.26 | 0.14 |
| | **SVM** | 0.15 | 0.10 |
| **Ionosphere** | **KNN** | 0.11 | 0.11 |
| | **SVM** | 0.13 | 0.14 |
| **Fertility-diagnosis** | **KNN** | 0.17 | 0.10 |
| | **SVM** | 0.12 | 0.07 |
| **Flags-religion** | **KNN** | 0.20 | 1.47 |
| | **SVM** | 0.20 | 1.79 |
| **Lung-cancer** | **KNN** | 0.38 | 0.70 |
| | **SVM** | 0.34 | 0.73 |
| **Lymphography** | **KNN** | 0.16 | 0.21 |
| | **SVM** | 0.27 | 0.17 |
| **Trains** | **KNN** | 0.51 | 0.50 |
| | **SVM** | 0.20 | 0.20 |
| **Wine** | **KNN** | 0.12 | 0.05 |
| | **SVM** | 0.26 | 0.05 |
| **Semeion** | **KNN** | 0.08 | 0.50 |
| | **SVM** | 0.16 | 0.41 |

[19] showing UFSMC to be effective in preserving cluster structure. All these works are parameter driven while the proposed approach is data driven. The number of features selected by UFRESO are significantly less than obtained by UFEMC, USFS, LSFS and FSFS. The significant increase in accuracy for all the datasets except a minute decrease for auto-univ-au1_1000, cardiotocography-3class and diabetes can also be observed from paired t-test results. But regardless of a bit increase for three datasets, the need for supplying parameter values beforehand makes the USFS, UFSMC, LSFS and FSFS approaches user dependent. Further, the number of wins and losses clearly indicate that UFRESO has won the match. Exactly same pattern is observed for classification error also.

#### 7.2.1.4 Comparison with supervised approach

In this section, fuzzy rough set based feature selection [72] using ant colony (FRASO) and particle swarm (FRPSO) search heuristic is used for comparison. Although the comparison of supervised and unsupervised approach is unreasonable, however that would be effective in laying the supremacy of the proposed work. Tables 7.10, 7.16 and 7.11 depicts the recorded results. The number of features selected by supervised approaches is comparatively less than that obtained by UFRESO. The presence of class labels adds some information that is utilized by supervised approaches, which can not be taken as an advantage in case of unsupervised methods. Considering classification accuracy, there is similar trend in accuracy for both supervised and unsupervised approaches for most of datasets, again because of discrimination information supplied by decision class. However, datasets namely heart-cleveland, fertility-diagnosis, flags-religion, trains and semeion have produced high performance with UFRESO that may have resulted because of super set of features selected by

Table 7.10: Number of features selected on comparison with other state of art supervised feature selection algorithms

| Dataset | FRPSO | FRASO | UFRESO |
|---|---|---|---|
| **Auto-univ-au1_1000** | 16.3 | **1.0** | 15.1 |
| **German** | **13** | 16.5 | 17.3 |
| **Cardiotocography-3class** | **10.8** | 19.4 | 22.1 |
| **Diabetes** | 8.0 | 8.0 | **6.6** |
| **Leaf** | 11.7 | 13.1 | **10.9** |
| **Abalone-11class** | 8.0 | 8.0 | **7.1** |
| **Heart-cleveland** | 8.1 | **7.7** | 10.5 |
| **Hepatitis** | 7.1 | **6.9** | 14.3 |
| **Ionosphere** | **7.0** | 7.8 | 16.0 |
| **Fertility-diagnosis** | **6.9** | **6.9** | 8.2 |
| **Flags-religion** | **1.0** | 2.0 | 19.6 |
| **Lung-cancer** | **5.4** | 5.5 | 19.1 |
| **Lymphography** | **7.3** | 7.3 | 13.0 |
| **Trains** | **1.0** | 2.0 | 8.9 |
| **Wine** | **4.8** | **4.8** | 10.9 |
| **Semeion** | 22.5 | **18.4** | 103.9 |

UFRESO. Statistical results demonstrates the supremacy of UFRESO. Classification error further reflects the better performing behaviour of UFRESO.

## 7.3 Summary

Two fuzzy rough set based approaches employing dependency degree and boundary measure are proposed in this study for feature selection in unsupervised domain. The two phases of feature selection namely feature subset selection and subset quality evaluation has been dealt in this work. The candidate feature subsets are selected using proposed earthworm search strategy. The use of search strategies have further enhanced the performance. The feature subset quality is evaluated considering fuzziness arising in real world applications.

Table 7.11: Classification error comparison with other state of art supervised feature selection algorithms

| **Dataset** | | **Classification error** | | |
|---|---|---|---|---|
| | | **FRPSO** | **FRASO** | **UFRESO** |
| **Auto-univ-au1_1000** | **KNN** | 0.36 | 0.38 | 0.29 |
| | **SVM** | 0.26 | 0.26 | 0.26 |
| **German** | **KNN** | 0.32 | 0.33 | 0.29 |
| | **SVM** | 0.33 | 0.29 | 0.24 |
| **Cardiotocography-3class** | **KNN** | 0.04 | 0.03 | 0.03 |
| | **SVM** | 0.23 | 0.23 | 0.17 |
| **Diabetes** | **KNN** | 0.30 | 0.30 | 0.28 |
| | **SVM** | 0.35 | 0.35 | 0.34 |
| **Leaf** | **KNN** | 0.03 | 0.03 | 4.07 |
| | **SVM** | 0.06 | 0.06 | 6.47 |
| **Abalone-11class** | **KNN** | 0.14 | 0.14 | 1.59 |
| | **SVM** | 0.16 | 0.16 | 1.39 |
| **Heart-cleveland** | **KNN** | 0.19 | 0.20 | 0.71 |
| | **SVM** | 0.27 | 0.27 | 0.63 |
| **Hepatitis** | **KNN** | 0.22 | 0.22 | 0.14 |
| | **SVM** | 0.15 | 0.15 | 0.10 |
| **Ionosphere** | **KNN** | 0.12 | 0.15 | 0.11 |
| | **SVM** | 0.17 | 0.18 | 0.14 |
| **Fertility-diagnosis** | **KNN** | 0.18 | 0.20 | 0.10 |
| | **SVM** | 0.12 | 0.12 | 0.07 |
| **Flags-religion** | **KNN** | 0.20 | 0.20 | 1.47 |
| | **SVM** | 0.20 | 0.20 | 1.79 |
| **Lung-cancer** | **KNN** | 0.40 | 0.41 | 0.70 |
| | **SVM** | 0.36 | 0.33 | 0.73 |
| **Lymphography** | **KNN** | 0.13 | 0.13 | 0.21 |
| | **SVM** | 0.27 | 0.27 | 0.17 |
| **Trains** | **KNN** | 0.60 | 0.54 | 0.50 |
| | **SVM** | 0.20 | 0.10 | 0.20 |
| **Wine** | **KNN** | 0.05 | 0.06 | 0.05 |
| | **SVM** | 0.24 | 0.25 | 0.05 |
| **Semeion** | **KNN** | 0.08 | 0.10 | 0.50 |
| | **SVM** | 0.16 | 0.16 | 0.41 |

Table 7.12: Classification accuracy comparison with other state of art feature selection algorithm

| **Dataset** | | **UFRFS** | | | **UFRESO** | |
|---|---|---|---|---|---|---|
| | | **Acc** | **Std** | **p-val** | **Acc** | **Std** |
| **Auto-univ-au1_1000** | **KNN** | 72.6 | 3.53 | 0.17 o | 70.73 | 2.21 |
| | **SVM** | 74.1 | 0.53 | 0.01 + | 71.65 | 2.5 |
| **German** | **KNN** | 69.3 | 3.95 | 0.02 - | 73.27 | 2.61 |
| | **SVM** | 76.9 | 4.09 | 0.70 o | 77.44 | 1.48 |
| **Cardiotocography-3class** | **KNN** | 96.66 | 1.53 | 0.50 o | 96.21 | 1.38 |
| | **SVM** | 77.85 | 0.14 | 0.00 - | 83.18 | 3.29 |
| **Diabetes** | **KNN** | 73.69 | 3.03 | 0.01 + | 70.14 | 2.58 |
| | **SVM** | 67.31 | 2.97 | 0.01 + | 63.32 | 3.43 |
| **Leaf** | **KNN** | 51.47 | 7.63 | 0.00 - | 67.74 | 4.9 |
| | **SVM** | 26.76 | 6.71 | 0.00 - | 39.14 | 5.64 |
| **Abalone-11class** | **KNN** | 23.2 | 2.1 | 0.54 o | 22.78 | 0.46 |
| | **SVM** | 26.92 | 2.18 | 0.03 - | 28.74 | 1.02 |
| **Heart-cleveland** | **KNN** | 56.46 | 6.51 | 0.91 o | 56.2 | 2.55 |
| | **SVM** | 58.14 | 5.32 | 0.28 o | 60.28 | 2.98 |
| **Hepatitis** | **KNN** | 78.71 | 8.93 | 0.01 - | 88.21 | 3.05 |
| | **SVM** | 85.21 | 9.32 | 0.06 o | 91.26 | 2.38 |
| **Ionosphere** | **KNN** | 89.46 | 3.31 | 0.17 o | 87.86 | 1.24 |
| | **SVM** | 86.9 | 5.07 | 0.32 o | 85.23 | 0.99 |
| **Fertility-diagnosis** | **KNN** | 89 | 7.38 | 0.77 o | 88.29 | 1.61 |
| | **SVM** | 88 | 4.22 | 0.00 - | 93.36 | 2.54 |
| **Flags-religion** | **KNN** | 30.95 | 0.82 | 0.00 - | 51.26 | 2.71 |
| | **SVM** | 30.95 | 0.82 | 0.00 - | 39.9 | 1.89 |
| **Lung-cancer** | **KNN** | 48.33 | 22.84 | 0.15 o | 37.33 | 3.69 |
| | **SVM** | 57.5 | 15.91 | 0.04 + | 44.51 | 9.28 |
| **Lymphography** | **KNN** | 72.81 | 9.6 | 0.08 o | 78.98 | 4.01 |
| | **SVM** | 83.05 | 10.26 | 0.80 o | 83.89 | 1.46 |
| **Trains** | **KNN** | 40 | 51.64 | 0.09 o | 72.32 | 22.12 |
| | **SVM** | 80 | 42.16 | 0.47 o | 90.18 | 10.73 |
| **Wine** | **KNN** | 84.77 | 10.37 | 0.01 - | 94.54 | 2.34 |
| | **SVM** | 87.03 | 9.6 | 0.05 - | 93.74 | 2.66 |
| **Semeion** | **KNN** | 64.15 | 3.85 | 0.00 - | 89.2 | 0.06 |
| | **SVM** | 63.58 | 4.31 | 0.00 - | 89.81 | 0.05 |
| **Loss/Win/Tie** | **KNN** | 6/1/9 | | | | |
| | **SVM** | 7/3/6 | | | | |

Table 7.13: Number of features selected on comparison with other state of art non dependency based feature selection algorithms

| Dataset | UFSMC | USFS | LSFS | FSFS | UFRESO |
|---|---|---|---|---|---|
| **Auto-univ-au1_1000** | 17.7 | 20.0 | 16.8 | 18.0 | **15.1** |
| **German** | 21.3 | 23.0 | 17.9 | 19.5 | **17.3** |
| **Cardiotocography-3class** | 30.0 | **19.4** | 35.0 | 31.0 | 22.1 |
| **Diabetes** | 8.0 | 8.0 | 8.0 | 7.0 | **6.6** |
| **Leaf** | 14.0 | 14.0 | 13.7 | 13.0 | **10.9** |
| **Abalone-11class** | 8.0 | 8.0 | 8.0 | **7.0** | 7.1 |
| **Heart-cleveland** | 12.9 | 12,9 | 11.2 | 12.0 | **10.5** |
| **Hepatitis** | 19.0 | 17.7 | 19.0 | 16.3 | **14.3** |
| **Ionosphere** | 28.4 | 28.1 | 32.0 | 30.0 | **16.0** |
| **Fertility-diagnosis** | 8.9 | 9.0 | 9.0 | **8.0** | 8.2 |
| **Flags-religion** | 28.0 | 27.6 | 28.0 | 25.0 | **19.6** |
| **Lung-cancer** | 54.3 | 41.9 | 36.3 | 46.9 | **19.1** |
| **Lymphography** | 17.9 | 13.8 | 15.9 | 15.9 | **13.0** |
| **Trains** | 13.3 | 21.7 | 10.3 | 20.6 | **8.9** |
| **Wine** | 11.3 | 11.8 | 13.0 | 12.0 | **10.9** |
| **Semeion** | 171.9 | 166.9 | 197.7 | 195.4 | **103.9** |

Table 7.14: Classification accuracy comparison with other state of art non dependency based feature selection algorithms

| Dataset | | UFSMC | | | USFS | | | LSFS | | | FSFS | | | UFRESO | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Acc | Std | p-val | Acc | Std | p-val | Acc | Std | p-val | Acc | Std | p-val | Acc | Std |
| **Auto-univ-au1_1000** | **KNN** | 70 | 4.47 | 0.65 o | 70.89 | 4.29 | 0.52 o | 70.3 | 4.89 | 0.80 o | 69.14 | 1 | 0.05 - | 70.73 | 2.21 |
| | **SVM** | 73.5 | 2.41 | 0.11 o | 74.3 | 2.53 | 0.03 + | 74 | 2.68 | 0.06 o | 71.64 | 1.52 | 0.99 o | 71.65 | 2.5 |
| **German** | **KNN** | 71.5 | 3.93 | 0.25 o | 67.1 | 2.73 | 0.00 - | 67.9 | 3.36 | 0.00 - | 67.67 | 1.14 | 0.00 - | 73.27 | 2.61 |
| | **SVM** | 70.8 | 4.77 | 0.00 - | 68.7 | 7.45 | 0.00 - | 66 | 13.94 | 0.02 - | 74.73 | 0.8 | 0.00 - | 77.44 | 1.48 |
| **Cardiotocography-3class** | **KNN** | 83.06 | 10.71 | 0.00 - | 80.57 | 9.68 | 0.00 - | 79.3 | 11.38 | 0.00 - | 98.39 | 0.34 | 0.00 + | 96.21 | 1.38 |
| | **SVM** | 89.73 | 12.42 | 0.12 o | 73.83 | 20.24 | 0.17 o | 70.67 | 19.73 | 0.06 o | 82.71 | 8.79 | 0.88 o | 83.18 | 3.29 |
| **Diabetes** | **KNN** | 69 | 4.64 | 0.51 o | 69 | 4.64 | 0.51 o | 69 | 4.64 | 0.51 o | 68.14 | 3.79 | 0.18 o | 70.14 | 2.58 |
| | **SVM** | 67.19 | 3.29 | 0.02 + | 67.19 | 3.29 | 0.02 + | 67.19 | 3.29 | 0.02 + | 62.55 | 2.23 | 0.56 o | 63.32 | 3.43 |
| **Leaf** | **KNN** | 5.88 | 6.16 | 0.00 - | 5.88 | 6.16 | 0.00 - | 4.7 | 4.4 | 0.00 - | 58.91 | 1.74 | 0.00 - | 67.74 | 4.9 |
| | **SVM** | 2.64 | 3.07 | 0.00 - | 2.64 | 3.07 | 0.00 - | 2.64 | 3.07 | 0.00 - | 37.68 | 2.11 | 0.45 o | 39.14 | 5.64 |
| **Abalone-11class** | **KNN** | 23.89 | 6.2 | 0.58 o | 23.89 | 6.2 | 0.58 o | 23.89 | 6.2 | 0.58 o | 21.79 | 0.49 | 0.00 - | 22.78 | 0.46 |
| | **SVM** | 27.38 | 7.32 | 0.57 o | 27.38 | 7.32 | 0.57 o | 27.38 | 7.32 | 0.57 o | 28.01 | 0.35 | 0.05 - | 28.74 | 1.02 |
| **Heart-cleveland** | **KNN** | 47.82 | 8.02 | 0.01 - | 49.16 | 9.74 | 0.04 - | 48.5 | 8.4 | 0.01 - | 56.85 | 2.03 | 0.54 o | 56.2 | 2.55 |
| | **SVM** | 47.69 | 21.83 | 0.09 o | 53.4 | 7.02 | 0.01 - | 37.06 | 17.41 | 0.00 - | 54.47 | 1.22 | 0.00 - | 60.28 | 2.98 |
| **Hepatitis** | **KNN** | 73.33 | 11.63 | 0.00 - | 73.33 | 11.63 | 0.00 - | 73.33 | 11.63 | 0.00 - | 85.92 | 4.02 | 0.17 o | 88.21 | 3.05 |
| | **SVM** | 78.7 | 19.4 | 0.06 o | 77 | 16.07 | 0.01 - | 79.7 | 11.55 | 0.01 - | 88.69 | 6.16 | 0.23 o | 91.26 | 2.38 |
| **Ionosphere** | **KNN** | 84.92 | 9.99 | 0.37 o | 84.36 | 10.76 | 0.32 o | 81.5 | 11.4 | 0.10 o | 81.43 | 4.65 | 0.00 - | 87.86 | 1.24 |
| | **SVM** | 85.23 | 9.22 | 1.00 o | 82.36 | 9.1 | 0.33 o | 81.51 | 8.56 | 0.19 o | 83.35 | 4.6 | 0.22 o | 85.23 | 0.99 |
| **Fertility-diagnosis** | **KNN** | 86 | 11.13 | 0.53 o | 86 | 11.13 | 0.53 o | 86 | 11.13 | 0.53 o | 81.13 | 2.65 | 0.00 - | 88.29 | 1.61 |
| | **SVM** | 88 | 11.66 | 0.17 o | 88 | 11.66 | 0.17 o | 88 | 11.66 | 0.17 o | 81.47 | 2.49 | 0.00 - | 93.36 | 2.54 |
| **Flags-religion** | **KNN** | 38.1 | 8.13 | 0.00 - | 38.63 | 8.85 | 0.00 - | 38.1 | 8.13 | 0.00 - | 37.07 | 6.22 | 0.00 - | 51.26 | 2.71 |
| | **SVM** | 28.34 | 8.3 | 0.00 - | 36.05 | 14.87 | 0.43 o | 33.1 | 9.83 | 0.05 - | 37.2 | 10.19 | 0.42 o | 39.9 | 1.89 |
| **Lung-cancer** | **KNN** | 35.83 | 22.98 | 0.84 o | 39.16 | 32.5 | 0.86 o | 33.33 | 27.13 | 0.65 o | 26.95 | 11.85 | 0.02 - | 37.33 | 3.69 |
| | **SVM** | 35 | 21.66 | 0.22 o | 40 | 21.66 | 0.55 o | 39.16 | 32.92 | 0.63 o | 24.88 | 11.16 | 0.00 - | 44.51 | 9.28 |
| **Lymphography** | **KNN** | 79.66 | 6.2 | 0.77 o | 77.71 | 5.93 | 0.58 o | 78.33 | 8.96 | 0.84 o | 73.83 | 3.18 | 0.01 - | 78.98 | 4.01 |
| | **SVM** | 81.04 | 13.63 | 0.52 o | 79 | 10.64 | 0.17 o | 81 | 13.07 | 0.50 o | 77.93 | 0.98 | 0.00 - | 83.89 | 1.46 |
| **Trains** | **KNN** | 60 | 48.98 | 0.48 o | 60 | 48.98 | 0.48 o | 70 | 45.82 | 0.89 o | 68.75 | 14.5 | 0.67 o | 72.32 | 22.12 |
| | **SVM** | 50 | 50 | 0.02 - | 50 | 50 | 0.02 - | 80 | 40 | 0.45 o | 79.71 | 13.66 | 0.07 o | 90.18 | 10.73 |
| **Wine** | **KNN** | 87.05 | 14.88 | 0.13 o | 85.35 | 13.79 | 0.05 - | 71.96 | 13.84 | 0.00 - | 91.03 | 2.41 | 0.00 - | 94.54 | 2.34 |
| | **SVM** | 91.01 | 7.54 | 0.29 o | 87.51 | 9.99 | 0.07 o | 86.99 | 12 | 0.00 - | 89.87 | 2.08 | 0.00 - | 93.74 | 2.66 |
| **Semeion** | **KNN** | 89.76 | 1.92 | 0.37 o | 88.25 | 4.18 | 0.48 o | 89.13 | 3.45 | 0.95 o | 87.44 | 0.65 | 0.00 - | 89.2 | 0.06 |
| | **SVM** | 87.38 | 2.27 | 0.00 - | 87.25 | 1.66 | 0.00 - | 85.56 | 3.19 | 0.00 - | 92.37 | 0.87 | 0.00 + | 89.81 | 0.05 |
| **Loss/Win/Tie** | **KNN** | 5/0/11 | | | 7/0/9 | | | 7/0/9 | | | 11/1/4 | | | | |
| | **SVM** | 5/1/10 | | | 6/2/8 | | | 7/1/8 | | | 7/1/8 | | | | |

Table 7.15: Classification error comparison with other state of art non dependency based feature selection algorithms

| Dataset | | **Classification error** | | | | |
|---|---|---|---|---|---|---|
| | | **UFSMC** | **USFS** | **LSFS** | **FSFS** | **UFRESO** |
| **Auto-univ-au1_1000** | **KNN** | 0.30 | 0.29 | 0.29 | 0.30 | 0.29 |
| | **SVM** | 0.26 | 0.25 | 0.26 | 0.27 | 0.26 |
| **German** | **KNN** | 0.28 | 0.32 | 0.32 | 0.31 | 0.29 |
| | **SVM** | 0.29 | 0.31 | 0.33 | 0.25 | 0.24 |
| **Cardiotocography-3class** | **KNN** | 0.20 | 0.26 | 0.27 | 0.01 | 0.03 |
| | **SVM** | 0.11 | 0.40 | 0.36 | 0.10 | 0.17 |
| **Diabetes** | **KNN** | 0.30 | 0.30 | 0.30 | 0.29 | 0.28 |
| | **SVM** | 0.32 | 0.32 | 0.32 | 0.36 | 0.34 |
| **Leaf** | **KNN** | 11.19 | 11.19 | 11.87 | 4.52 | 4.07 |
| | **SVM** | 11.72 | 11.72 | 12.21 | 7.05 | 6.47 |
| **Abalone-11class** | **KNN** | 1.56 | 1.56 | 1.56 | 1.62 | 1.59 |
| | **SVM** | 1.37 | 1.37 | 1.37 | 1.35 | 1.39 |
| **Heart-cleveland** | **KNN** | 0.91 | 0.90 | 0.97 | 0.72 | 0.71 |
| | **SVM** | 0.86 | 0.78 | 1.05 | 0.66 | 0.63 |
| **Hepatitis** | **KNN** | 0.26 | 0.26 | 0.26 | 0.15 | 0.14 |
| | **SVM** | 0.21 | 0.22 | 0.20 | 0.16 | 0.10 |
| **Ionosphere** | **KNN** | 0.15 | 0.15 | 0.16 | 0.16 | 0.11 |
| | **SVM** | 0.14 | 0.17 | 0.18 | 0.14 | 0.14 |
| **Fertility-diagnosis** | **KNN** | 0.14 | 0.14 | 0.14 | 0.20 | 0.10 |
| | **SVM** | 0.12 | 0.12 | 0.12 | 0.19 | 0.07 |
| **Flags-religion** | **KNN** | 1.70 | 1.72 | 1.70 | 1.70 | 1.47 |
| | **SVM** | 1.88 | 1.88 | 1.80 | 1.63 | 1.79 |
| **Lung-cancer** | **KNN** | 0.67 | 0.64 | 0.70 | 0.70 | 0.70 |
| | **SVM** | 0.68 | 0.63 | 0.64 | 0.66 | 0.73 |
| **Lymphography** | **KNN** | 0.20 | 0.24 | 0.22 | 0.25 | 0.21 |
| | **SVM** | 0.20 | 0.23 | 0.20 | 0.25 | 0.17 |
| **Trains** | **KNN** | 0.40 | 0.40 | 0.30 | 0.30 | 0.50 |
| | **SVM** | 0.50 | 0.50 | 0.20 | 0.20 | 0.20 |
| **Wine** | **KNN** | 0.15 | 0.17 | 0.35 | 0.07 | 0.05 |
| | **SVM** | 0.08 | 0.12 | 0.13 | 0.07 | 0.05 |
| **Semeion** | **KNN** | 0.43 | 0.52 | 0.46 | 0.50 | 0.50 |
| | **SVM** | 0.51 | 0.51 | 0.59 | 0.35 | 0.41 |

TABLE 7.16: Classification accuracy comparison with other state of art supervised feature selection algorithms

| **Dataset** | | **FRPSO** | | | **FRASO** | | | **UFRESO** | |
|---|---|---|---|---|---|---|---|---|---|
| | | **Acc** | **Std** | **p-val** | **Acc** | **Std** | **p-val** | **Acc** | **Std** |
| **Auto-univ-au1_1000** | **KNN** | 71.1 | 2.73 | 0.74 o | 74.1 | 0.32 | 0.00 - | 70.73 | 2.21 |
| | **SVM** | 74.1 | 0.32 | 0.01 + | 74.1 | 0.32 | 0.01 + | 71.65 | 2.5 |
| **German** | **KNN** | 72.2 | 5.22 | 0.57 o | 71.9 | 4.07 | 0.38 o | 73.27 | 2.61 |
| | **SVM** | 66.7 | 12.03 | 0.01 - | 71.2 | 10.48 | 0.08 o | 77.44 | 1.48 |
| **Cardiotocography-3class** | **KNN** | 95.58 | 1.35 | 0.32 o | 96.47 | 1.76 | 0.72 o | 96.21 | 1.38 |
| | **SVM** | 95.81 | 0.81 | 0.00 + | 96.61 | 1.58 | 0.00 + | 83.18 | 3.29 |
| **Diabetes** | **KNN** | 73.69 | 3.03 | 0.01 + | 73.69 | 3.03 | 0.01 + | 70.14 | 2.58 |
| | **SVM** | 65.11 | 0.36 | 0.12 o | 65.11 | 0.36 | 0.12 o | 63.32 | 3.43 |
| **Leaf** | **KNN** | 55.88 | 8.43 | 0.00 - | 57.65 | 9.53 | 0.01 - | 67.74 | 4.9 |
| | **SVM** | 47.94 | 6.8 | 0.01 + | 47.35 | 8.02 | 0.02 + | 39.14 | 5.64 |
| **Abalone-11class** | **KNN** | 23.63 | 1.71 | 0.15 o | 23.63 | 1.71 | 0.15 o | 22.78 | 0.46 |
| | **SVM** | 27.2 | 1.4 | 0.01 - | 27.2 | 1.4 | 0.01 - | 28.74 | 1.02 |
| **Heart-cleveland** | **KNN** | 54.15 | 5.9 | 0.33 o | 55.19 | 8.62 | 0.73 o | 56.2 | 2.55 |
| | **SVM** | 59.8 | 6.08 | 0.83 o | 58.14 | 5.55 | 0.30 o | 60.28 | 2.98 |
| **Hepatitis** | **KNN** | 81.29 | 8.26 | 0.02 - | 83.04 | 10 | 0.14 o | 88.21 | 3.05 |
| | **SVM** | 85.13 | 6.26 | 0.01 - | 84.96 | 10.49 | 0.08 o | 91.26 | 2.38 |
| **Ionosphere** | **KNN** | 88.89 | 4.14 | 0.46 o | 86.33 | 4.39 | 0.30 o | 87.86 | 1.24 |
| | **SVM** | 82.62 | 4.76 | 0.11 o | 82.35 | 5.29 | 0.11 o | 85.23 | 0.99 |
| **Fertility-diagnosis** | **KNN** | 88 | 4.22 | 0.84 o | 86 | 8.43 | 0.41 o | 88.29 | 1.61 |
| | **SVM** | 88 | 4.23 | 0.00 - | 88 | 4.22 | 0.00 - | 93.36 | 2.54 |
| **Flags-religion** | **KNN** | 30.95 | 0.82 | 0.00 - | 30.87 | 9.28 | 0.00 - | 51.26 | 2.71 |
| | **SVM** | 30.95 | 0.82 | 0.00 - | 31.37 | 9.16 | 0.01 - | 39.9 | 1.89 |
| **Lung-cancer** | **KNN** | 43.33 | 32.82 | 0.57 + | 34.17 | 23.06 | 0.67 o | 37.33 | 3.69 |
| | **SVM** | 54.17 | 28.12 | 0.32 + | 66.67 | 23.9 | 0.01 + | 44.51 | 9.28 |
| **Lymphography** | **KNN** | 78.38 | 8.75 | 0.85 - | 80.33 | 12.11 | 0.74 o | 78.98 | 4.01 |
| | **SVM** | 84.43 | 10.97 | 0.88 + | 83.1 | 5.72 | 0.68 o | 83.89 | 1.46 |
| **Trains** | **KNN** | 0 | 0 | 0.00 - | 40 | 51.64 | 0.09 o | 72.32 | 22.12 |
| | **SVM** | 80 | 42.16 | 0.47 o | 90 | 31.62 | 0.99 o | 90.18 | 10.73 |
| **Wine** | **KNN** | 94.41 | 5.24 | 0.94 o | 92.71 | 5.33 | 0.33 o | 94.54 | 2.34 |
| | **SVM** | 92.78 | 9.82 | 0.77 o | 88.69 | 8.49 | 0.09 o | 93.74 | 2.66 |
| **Semeion** | **KNN** | 70.44 | 2.97 | 0.00 - | 62.27 | 4.09 | 0.00 - | 89.2 | 0.06 |
| | **SVM** | 71.25 | 2.77 | 0.00 - | 61.57 | 4.03 | 0.00 - | 89.81 | 0.05 |
| **Loss/Win/Tie** | **KNN** | 6/2/8 | | | 4/1/11 | | | | |
| | **SVM** | 6/5/5 | | | 4/4/8 | | | | |

***********