Chapter 6

Bireduct Model and its Application

Due to advancement in modern technologies, various sources like network of sensors, interconnected devices, etc generate millions of data every day. This has lead to circumstances where proportion of data to the number of tools to access the same is large. Such ever expansive data is rich both in dimension and size (number of instances). But not all the instances and features may contribute to classification accuracy and may even mitigate the performance. Therefore, there is an increasing need of techniques for data reduction. Feature selection (FS) or instance selection (IS) [34, 70, 151, 152] alone cannot handle the ever increasing size and dimensionality of dataset. Both the aspects of data reduction must be taken into consideration for enhancing classification accuracy. Few works have been done in the field of simultaneous instance and feature selection [9, 10, 48, 100, 101, 103, 138].

Most of the works based on rough, fuzzy rough and intuitionistic fuzzy rough theories focus on finding decision reducts. An extension of reduct, viz rough set bireduct [133, 138] has emerged based on the idea of bi-clustering. It selects features and

instances producing a reduced dataset that increases classification accuracy by removing irrelevant information. It not only removes irrelevant and/or redundant features, but also reduces the data by eliminating outliers. In order to deal with real valued continuous datasets, concept of bireduct was further extended in fuzzy rough framework. Further extensions of the concept was made in [101, 138], using ε-Bireducts [103, 137] and search strategies [35].

This chapter proposes a novel method of generating bireduct in intuitionstic fuzzy rough set framework. It simultaneously reduces dimensionality and data size by employing a robust lower approximation formulation (for calculating dependency degree) and similarity techniques. The maximum similarity of an instance with an outlier of the 'same class' is used for further elimination of outliers. All the existing works have been done in fuzzy rough framework using discernibility matrix approach, which has been extended to intuitionistic fuzzy case by employing dependency degree approach. The proposed work can hence be effectively used as a data reduction pre-processing technique, to learn robust decision rules.

6.0.1 Bireduct formulation

The idea of bireduct was introduced in rough set framework [133], which was further extended to fuzzy case by authors in [103]. The informal definition of bireduct focuses upon selecting minimal subset of features which describes decision class and corresponding subset of instances satisfying such descriptions.

According to [133], for an information system $I = (U, C \cup D)$, a subtable (A, Y) of I such that $A \subseteq C$ and $Y \subseteq U$ is bireduct iff

1. A forms the reduct of the system discerning decision class of Y, where $D(x) \neq D(y)$.

2. Y is maximal subset of U that discerns $x, y \in Y$.

This definition guarantees that no proper subset of A and superset of Y discerns all pairs of $x, y \in Y$. It considers instances $Z \in U \setminus Y$ as outliers that may have resulted because of noise.

This concept of bireduct was extended to fuzzy case in [101], that paved a way to compute bireducts in real-valued continuous domains.

Exploiting the above formulation of bireducts to intuitionistic fuzzy framework would be advantageous for different applications providing two degrees of freedom at the same time, unlike fuzzy case (which gives one degree of freedom). It would reduce data size and dimensionality considerably, and hence the complexity by eliminating both irrelevant and/or redundant features and problematic instances or outliers.

6.1 Intuitionistic Fuzzy Bireducts for Data Reduction

An insight into bireduct is given in Section 6.0.1. Bireducts reduces the complexity of learning algorithms by removing features and instances. Instances perceived as outlier or noisy are ignored in subsequent computation, thereby enhancing prediction performances of the learning algorithms. The proposed model works on this motivation to generate intuitionistic fuzzy bireducts.

6.1.1 Intuitionistic Fuzzy Feature Selection

The previous formulation of intuitionistic fuzzy approximations can be extended as follows:

$$R \downarrow_A X(x) = (\mu_{R\downarrow_A X}(x), \nu_{R\downarrow_A X}(x)) \tag{6.1}$$

$$= \begin{cases} (\inf_{y \notin X} \nu_{R_A}(x,y), \sup_{y \notin X} \mu_{R_A}(x,y)) & x \in X \\ (0,1) & x \notin X \end{cases} \tag{6.2}$$

$$R \uparrow_A X(x) = (\mu_{R\uparrow_A X}(x), \nu_{R\uparrow_A X}(x)) \tag{6.3}$$

$$= \begin{cases} (\sup_{y \in X} \mu_{R_A}(x,y), \inf_{y \in X} \nu_{R_A}(x,y)) & x \in X \\ (0,1) & x \notin X \end{cases} \tag{6.4}$$

The above defined approximations avoids misclassification of data [154]. Further, these values are affected by the presence of noise. Employing k-mean structure [63] to increase robustness. Arranging values corresponding to infimum/supremum in increasing/ decreasing order of magnitude and computing mean of first k samples gives the reformulated definition of lower and upper approximations as:

$$R \downarrow_A X(x) = \begin{cases} (\frac{1}{k} \sum_{y \notin X}^{k} \nu_{R_A}(x,y), \frac{1}{k} \sum_{y \notin X}^{k} \mu_{R_A}(x,y)) & x \in X \\ (0,1) & x \notin X \end{cases} \tag{6.5}$$

$$R \uparrow_A X(x) = \begin{cases} (\frac{1}{k} \sum_{y \in X}^{k} \mu_{R_A}(x,y), \frac{1}{k} \sum_{y \in X}^{k} \nu_{RA}(x,y)) & x \in X \\ (0,1) & x \notin X \end{cases} \tag{6.6}$$

Having framed the approximations in this way, dependency degree can be conveniently computed for the feature subset.

6.1.2 Intuitionistic Fuzzy Instance Selection

Once the definition of positive region is formulated, it can be used for elimination of outliers. Let $Pos_A(x)$ be the value of positive region for an instance $x \in U$. The proposed work introduces the following methods for instance selection in intuitionistic fuzzy case:

6.1.2.1 Method I

A simple approach is to eliminate all instances whose positive region value is below certain threshold parameter τ_o, an extension of fuzzy case [70]. Ostensibly, when positive region value is below certain threshold parameter τ_o, then it is uncertain as to which decision class an instance truly belongs. Such instances can be removed without any hassle, as shown in Algorithm 6.1.2.1 and flowchart 6.1.

Algorithm 6.1.2.1

Input: U: Set of all instances, τ_o: Threshold

for $\forall x \in U$ **do**

if $\mid Pos_A(x) \mid < \tau_o$ **then**

$U \leftarrow U - \{x\}$

end if

end for

return U

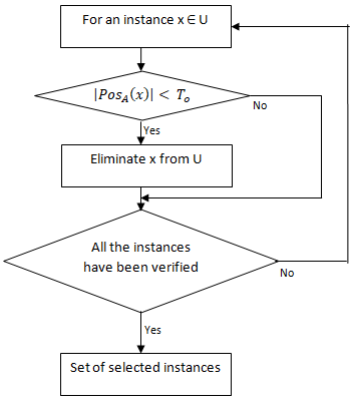


FIGURE 6.1: The flowchart of IFBRPSO-1

6.1.2.2 Method II

The above algorithm removes more instances than absolutely necessary. The value of positive region for an instance is affected by removal of an instance, which is not been considered in Section 6.1.2.1. Since the value of positive region basically gives the distance of 'nearest different class sample', so its value is increased on the removal of an instance.

A better approach would be to select an instance x for removal with minimum value of positive region, ρ_{min}. This instance is effectively removed as it is in the proximity of different class. Further, a nearest similar sample of an outlier or noisy sample x will be an outlier or noisy instance if its distance to x is less than ρ_{min}. So, the instance z belonging to same decision class as x with $\rho_{x,z} = min_i(1-R_A(x,i))$, $i \in U$ (has same decision class as x) is eliminated if $\rho_{x,z} < \rho_{min}$. This process eliminates at most two instances at a time. After removal of these problematic instances, value of positive region changes. The values of positive region of the instances are recalculated and the whole process is then repeated until all the problematic instances are eliminated. The Algorithm 6.1.2.2 and flowchart 6.2 depicts the entire procedure.

Algorithm 6.1.2.2

Input: U: Set of all instances

$\rho \leftarrow 1, \rho_{min} \leftarrow 1$

for $\forall y \in U$ **do**

if $\mid Pos_A(y) \mid < \rho_{min}$ **then**

$x \leftarrow y$

$\rho_{min} \leftarrow Pos_A(y)$

end if

end for

if $\rho_{min} < \rho$ **then**

$U \leftarrow U - \{x\}$

$\rho_{x,z} = 1$

for $\forall y \in U$ **do**

if x and y belong to same decision class **then**

if $\mid 1 - R_A(x, y) \mid < \rho_{x,z}$ **then**

$z \leftarrow y$

$\rho_{x,z} \leftarrow \mid 1 - R_A(x, y) \mid$

end if

end if

end for

if $\rho_{x,z} < \rho_{min}$ **then**

$U \leftarrow U - \{z\}$

end if

else

return U

end if

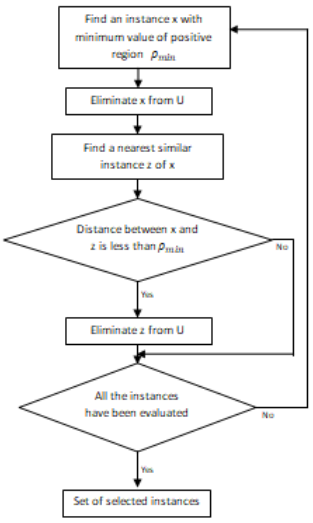


FIGURE 6.2: The flowchart of IFBRPSO-2

6.1.3 Simultaneous Intuitionistic Fuzzy Instance and Feature Selection

In conventional dependency based intuitionistic fuzzy rough set based approach, dataset is reduced by only selecting the subset of features that preserve dependency of unreduced dataset. However, data can be reduced by removing either (or both) of the instances or features for the intuitionistic fuzzy bireduct formulation. The rationale behind this, is that dataset may contain noisy samples and/or outliers. So, it is profitable to eliminate such problematic instances. Simultaneous intuitionistic fuzzy instance and feature selection is based on this very concept.

The two approaches (feature selection and instance selection) described in previous sections are combined to generate intuitionistic fuzzy bireducts. A toy example is illustrated in Table 6.1 for better understanding of the process. Using the intuitionistic fuzzy similarity measure in equations (6.7), (6.8) and (6.9), the intuitionistic

TABLE 6.1: Example Dataset

Instances / Features	a_1	a_2	a_3	a_4	**D**
x_1	0	1	1	10	1
x_2	4	2	0	15	1
x_3	0	0	4	20	2
x_4	0	3	2	15	2
x_5	3	0	1.5	25	1
x_6	1	1	2.5	20	1

$$R_{a_1} = \begin{bmatrix} (1,0) & (0,1) & (1,0) & (1,0) & (0.250,0.600) & (0.750,0.142) \\ (0,1) & (1,0) & (0,1) & (0,1) & (0.7500.142) & (0.250,0.600) \\ (1,0) & (0,1) & (1,0) & (1,0) & (0.250,0.600) & (0.750,0.142) \\ (1,0) & (0,1) & (1,0) & (1,0) & (0.250,0.600) & (0.750,0.142) \\ (0.250,.600) & (0.750,0.142) & (0.250,0.600) & (0.250,0.600) & (1,0) & (0.500,0.333) \\ (0.750,0.142) & (0.250,0.600) & (0.750,0.142) & (0.750,0.142) & (0.500,0.333) & (1,0) \end{bmatrix} \tag{6.11}$$

fuzzy relation for a feature a_1 is given by R_{a_1} in equation (6.11).

$$\mu_{R_A}(x,y) = T_{a \in A} \mu_{R_a}(x,y) \tag{6.7}$$

$$\nu_{R_A}(x,y) = \frac{1 - \mu_{R_A}(x,y)}{1 + \mu_{R_A}(x,y)} \tag{6.8}$$

R may be defined for a feature a by:

$$\mu_{R_a}(x,y) = 1 - \frac{a(x) - a(y)}{a_{\max} - a_{\min}} \tag{6.9}$$

$$\mu_{R_a}(x,y) = \max(\min(\frac{a(y) - (a(x) - std_a)}{std_a}, \frac{(a(x) + std_a) - a(x)}{std_a}), 0) \tag{6.10}$$

where $a(x)$ is the value of feature a for the instance x, $a_{\max}, a_{\min}$ is the maximum and minimum value that feature a has and standard deviation for feature a is given by std_a.

Using this, lower approximation $R \downarrow_{a_1} D$ is computed for the two decision classes $U \setminus D = \{D_1, D_2\}$, where $D_1 = \{x_1, x_2, x_5, x_6\}$ and $D_2 = \{x_3, x_4\}$ as shown in table

TABLE 6.2: Lower Approximation of feature a_1

$U \setminus D$	x_1	x_2	x_3	x_4	x_5	x_6
D_1	$(0,1)$	$(1,0)$	$(0,1)$	$(0,1)$	$(0.600, 0.250)$	$(0.142, 0.750)$
D_2	$(0,1)$	$(0,1)$	$(0.071, 0.875)$	$(0.071, 0.875)$	$(0,1)$	$(0,1)$

6.2.

The value of positive region is therefore computed using equation (6.12).

$$Pos_A(D)(x) = (\mu_{Pos_A(D)}(x), \nu_{Pos_A(D)}(x)) \tag{6.12}$$

$$= (\sup_{X \in U \setminus D} \mu_{R\downarrow_A X}(x), \inf_{X \in U \setminus D} \nu_{R\downarrow_A X}(x)) \tag{6.13}$$

$Pos_{a_1}(x_1) = (0,1)$

$Pos_{a_1}(x_2) = (1,0)$

$Pos_{a_1}(x_3) = (0.071, 0.875)$

$Pos_{a_1}(x_4) = (0.071, 0.875)$

$Pos_{a_1}(x_5) = (0.600, 0.250)$

$Pos_{a_1}(x_6) = (0.142, 0.750)$

Hence, dependency degree of feature a is $\gamma_{a_1}(D) = \frac{2.0679}{6} = 0.344$. Similarly, the value of dependency for remaining attribute is $\gamma_{a_2}(D) = 0$, $\gamma_{a_3}(D) = 0$, and $\gamma_{a_4}(D) = 0$. Since, a_1 has largest dependency degree it is selected. Consider the dataset corresponding to a_1 and eliminate the outliers. For example

1. Using process described in Section 6.1.2.1, for $\tau_o = 0.1$, instance $\{x_1, x_3, x_4\}$ is eliminated. Similarly, the dependency is recalculated by adding other features to this potentially reduced dataset and the re-performing instance selection. The entire process is iterated until some termination condition is met. So, the bireduct produced by this process consists of features $\{a_1, a_2\}$ or $\{a_1, a_3\}$ or $\{a_1, d\}$ and instances $\{x_2, x_5, x_6\}$.

2. Using process described in Section 6.1.2.2, the instance corresponding to minimum $\mid Pos_{a_1} \mid$ i.e. x_1 is eliminated. Further, the nearest similar instance to

x_1 is also eliminated. So, the distance of all the instances having same class as x_1 i.e. x_2, x_5, and x_6 is calculated as:

$\rho_{x_1,x_2} = 1.0$

$\rho_{x_1,x_5} = 0.675$

$\rho_{x_1,x_6} = 0.196$

Hence the nearest instance, the one corresponding to $min_i \rho_{x_1,x_i} = x_6$ is eliminated. Iterating the process yield bireducts consisting of $\{a_1, a_2\}$ features and $\{x_2, x_4, x_5\}$ instances or $\{a_1, a_4\}$ features and $\{x_2, x_3, x_4\}$ instances or $\{a_1, a_2\}$ features and $\{x_3, x_4\}$ instanes or $\{a_1, a_4\}$ features and $\{x_2, x_5\}$ instances.

The similarity measure given in equation (6.9) and (6.10) are respectively employed for instance and feature selection for experientation. As can be observed, many combinations of features and instances can be made. To constraint bireducts for containing atleast a proportion of the original instances thereby obtaining optimal bireduct, a concept of $\epsilon-$Bireduct has been introduced [137]. A parameter $\epsilon \in (0, 1]$ is used to stop instance elimination beyond certain limit, i.e. the number of instances $\mid Y \mid$ in bireduct (Y, A) must be more than $(1 - \epsilon) \mid U \mid$. Large values of ϵ leads to more number of elimination and vice versa.

6.1.4 Heuristic Search Strategy for IF Bireducts

In the previous section, the foundation for simultaneous feature and instance selection i.e. intuitionistic fuzzy bireduct was laid. However, there is a need for an effective and efficient search strategy that would reduce the data without performing exhaustive search [96].

Particle swarm optimization (PSO) is an evolutionary search strategy [82, 160] based on the unpredictable movement of flock of birds. PSO for generating bireducts is

initialised with m swarms. Each swarm $S^i = \{s_1^i, s_2^i, \ldots, s_n^i\}$ is a n dimensional random vector consisting of 0s and 1s, where n is the number of features in the dataset and the value of 0 or 1 represents the presence or absence of corresponding feature. Each swarm is characterized by $pBest^i = \{p_1^i, p_2^i, \ldots, p_n^i\}$, best previous position (vector configuration giving best fitness value). Globally best position $gBest^i = \{g_1, g_2, \ldots, g_n\}$ is given by the position which is best among all the swarms. Swarm's positions are updated by a velocity vector vel^i, which is computed using the following formula:

$$vel^i = w \times vel^i + c_1 \times r_1 \times \sum_{j=1}^{n} \left(p_j^i - s_j^i\right) + c_2 \times r_2 \times \sum_{j=1}^{n} \left(g_j - s_j^i\right) \tag{6.14}$$

where r_1 and r_2 are two random numbers lying between [0, 1], c_1 and c_2 are acceleration constants, w is an inertia weight for balancing between local and global search. The value of velocity governs the number of bits that should be changed in S^i in order to lead swarms head towards optimal solution. Let s_g be the number of different bits between swarm's current position and $gBest$. The position of swarm is updated via one of the two cases:

1. If $vel^i \leq s_g$, swarms's velocity is less than or equal to difference between s^i and $gBest$. vel^i number of bits of s^i is randomly flipped.

2. If $vel^i > s_g$, swarm's velocity overshoots the difference between s^i and $gBest$. Randomly change $(vel^i - s_g)$ bits in swarm's position that are different from $gBest$ apart from changing the different bits in s^i to be same as that in global best position $gBest$. This way heads the swarm towards optimal solution.

Maximum velocity of swarms is constrained to v^{max}, which is set to $n/3$ to prevent swarm from flying too away from optimal solution. So, if $vel^i < 1$, then $vel^i = 1$, if $vel^i > v^{max}$, then $vel^i = v^{max}$.

The fitness function is utilized to evaluate quality of the bireduct. Since, the quality of bireduct is influenced by the presence of subset of features that satisfy maximal number of instances. Hence taking into account the dual objective, the fitness function is defined as:

$$Fit_i = \alpha \times \Upsilon_A(D) + \beta \times \frac{(n+ \mid U \mid) - (r+o)}{n+ \mid U \mid} \tag{6.15}$$

where n is the total number of features in the dataset, r is the number of bits set in s^i and o is the number of instances covered by r after removing outliers. The two parameters α and β govern the importance to classification performance and subset length respectively, such that $\alpha = 1 - \beta, \alpha \in [0, 1]$. After the selection of globally best position of swarm, outliers are eliminated using the process described in Section 6.1.2. The entire methodology is iterated *generation* times. The whole methodology is described in Algorithm 6.1.4 and depicted in flowchart 6.3.

Algorithm 6.1.4

Input: *generation*: number of iterations; m: number of swarms; c_1, c_2: constants; w: inertia weight; v^{max}: maximum swarm's velocity;

$s^i = n$ bit vector generated randomly; $i = 1, 2, \ldots, m$

$pBestFit^i, gBestFit = 0; i = 1, 2, \ldots, m$

$itr = 0;$

while $itr <$ *generation* **do**

for $\forall$ swarm s^i **do**

$Fit_i =$ Fitness of swarm i;

if $Fit_i > pBestFit^i$ **then**

$pBestFit^i = Fit_i; pBest^i = s^i; j = 1, 2, \ldots, n$

end if

if $Fit_i > gBestFit$ **then**

$gBestFit = Fit_i; gBest = s^i; j = 1, 2, \ldots, n$

end if

end for

for $\forall$ swarm s^i **do**

$vel^i = w \times vel^i + c_1 \times r_1 \times \sum_{j=1}^{n} \left(p_j^i - s_j^i\right) + c_2 \times r_2 \times \sum_{j=1}^{n} \left(g_j - s_j^i\right)$;

if $vel^i > vel^{max}$ **then**

$vel^i = vel^{max}$;

end if

if $vel^i < 1$ **then**

$vel^i = 1$;

end if

$s_g = \sum_{j=1}^{n} \mid \left(gBest_j - s_j^i\right) \mid$;

if $s_g \leq vel^i$ **then**

Randomly change vel^i bits of s^i;

else

Randomly change $s_g - vel^i$ bits that are different from $gBest$ apart from changing the bits that are different to $gBest$;

end if

Remove outliers O from dataset containing $gBest$ features;

$U \leftarrow U - O$;

end for

$itr \leftarrow itr + 1$;

end while

return ϵ-bireduct

Let the bireduct obtained in first *generation* of i^{th} swarm be $\{a, b\}$ features and $\{2, 4, 5\}$ instances for the previous toy example, then the fitness value $Fit_i = 0.9 \times$

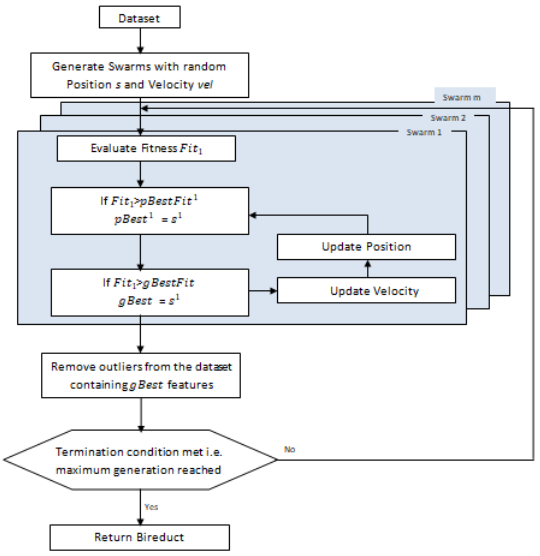


FIGURE 6.3: The flowchart of entire metoodology to calculate $\epsilon-$bireduct i.e for obtaining a reduced representation of the dataset

$0.833 + 0.1 \times \frac{(4+6)-(2+3)}{4+6} = 0.800$, $\alpha = 0.9$ and $\beta = 0.1$ is used.

The worst case time complexity of the algorithm is:

$O\left(generation* \mid swarms \mid * \mid U \mid * \mid U \mid * n\right)$.

In each iteration of the algorithm which is done *generation* times, fitness of all the swarms is evaluated. The fitness of the bireduct is based on dependency degree which has $O\left(\mid U \mid * \mid U \mid * n\right)$ time complexity.

6.2 Experimentation

This section details the experimental evaluation to show the effectiveness of the proposed approach for data reduction.

The experimental setup for the proposed approach (IFBR) is as follows: $\alpha = 0.9$, $\beta = 0.1$ is chosen. The value of *generation* is set to 100 and 20 swarms are considered. Further, the constants c_1 and c_2 defined in equation (6.14) are set to 2 and value of inertia weight is modified using following equation:

TABLE 6.3: Benchmark Datasets

Dataset	Instance	Feature	Class	Classification accuracy	
				KNN	SVM
Diabetes	768	8	2	72.50±5.70	64.86±6.35
Glass	214	9	6	69.52±9.03	62.85±7.02
Appendicitis	106	7	2	83.00±9.48	85.00±7.07
Heart	267	13	2	82.69±8.92	74.23±10.42
Fertility-diagnosis	100	9	2	88.00±11.35	88.00±12.29
Wine	178	13	3	95.29±6.67	95.29±5.40
German	1000	24	2	70.80±6.35	76.60±3.53
Hepatitis	155	19	2	81.33±7.56	84.00±7.16
Flags-religion	194	28	8	48.42±10.46	44.21±9.98
Leaf	340	14	30	66.47±5.03	48.23±9.32
Lymphography	148	18	4	85.00±10.35	80.71±12.16
Seeds	210	7	3	91.90±5.04	92.38±4.60
Dbworld-bodies	64	4702	2	53.33±20.48	85.00±5.27
Dbworld-bodies-stemmed	64	3721	2	53.25±3.26	81.93±1.57
Micro-mass-mixed-spectra	360	1300	10	8t.11±7.94	79.16±8.71

$$w = (w_{max} - w_{min}) \times \frac{itr}{generation} + w_{min} \tag{6.16}$$

where itr is the current iteration number, $w_{min} = 0.4$ and $w_{max} = 1.4$. These are the default parameter settings as employed by Wang et. al. [160] for particle swarm optimization.

6.2.1 Results

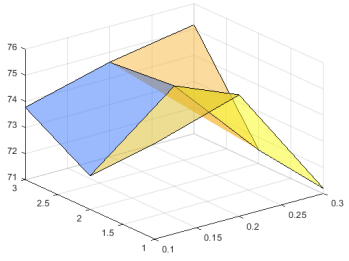
Fifteen benchmark datasets taken from UCI repository [108] are used to conduct experiments and are summarised in table 6.3. Five sets of experiments are performed illustrating the proposed approach and its variants. Further, a comparative study is done to demonstrate the effectiveness of the proposed model. All the respective accuracies are evaluated using 10×10-fold cross validation technique. Two classifiers namely kNN (k = 3) [131] and SVM [117] are employed to evaluate performances. Highest performances are bold-faced and rank of algorithm superscripted.

6.2.1.1 Using Parameter Variation

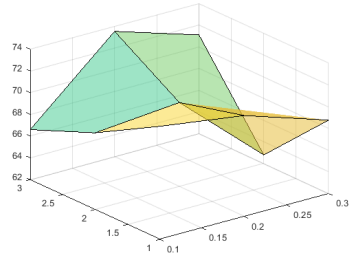
A series of experiments are performed by varying of k (used to compute lower approximations) as $1, 2$ and 3. Also, the results are evaluated for $\epsilon = 0.1, 0.2, 0.3$ to ensure data coverage of 90%, 80% and 70% respectively for generating intuitionistic fuzzy bireducts. The size of feature subset and instances is shown in table 6.4, while the classification accuracies (along with standard deviation) for different values of ϵ in tables 6.5, 6.6, and 6.7. There is not much difference in number of features and instances generated for $k = 1, 2, 3$ for fixed ϵ. Only the difference of one or two is observed in size of feature subsets for different value of ϵ. In terms of classification accuracy, $k = 2, 3$ has produced higher accuracy on some datasets. Again, not much difference between accuracies is observed for $\epsilon = 0.1$ except for few dataset like dbworld-bodies, etc. While there is a increase in number of datasets producing higher accuracy for $k = 2$ than $k = 3$ for 80% coverage of datasets i.e. $\epsilon = 0.2$ while decreasing the number of selected instances. The feature subset size is not much affected in the shift from $\epsilon = 0.2$ to $\epsilon = 0.3$. The classification accuracy for $\epsilon = 0.3$ follows nearly the same trend as for $\epsilon = 0.1$ i.e. a slight higher value for $k = 2$ and 3. For better understanding the difference, a visualization of accuracy for varying values of ϵ and k for kNN classifier is shown in Figure 6.4. There is a little increase in accuracy on increasing coverage for most of the datasets, which is to be expected as reduced data (smaller subtable) gives higher performance. However, on average $\epsilon = 0.2$, $k = 2$ chooses less number of features maintaining high classification accuracy and is henceforth used for subsequent experimental evaluation.

TABLE 6.4: IFBR results for various parameter combination

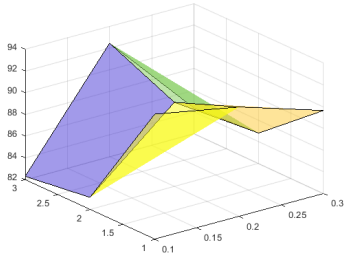
Dataset	k	$\epsilon = 0.1$			$\epsilon = 0.2$			$\epsilon = 0.3$		
		1	2	3	1	2	3	1	2	3
Diabetes	Instance	697.8	697.8	697.9	650.8	650.0	646.5	650.4	650.9	645.7
	Feature	6.7	6.7	6.6	6.9	6.2	6.5	7.2	6.5	6.3
Glass	Instance	193.9	193.9	193.6	174.5	174.7	174.5	155.5	155.8	155.7
	Feature	8.6	8.2	7.9	7.5	7.2	6.6	6.9	6.0	6.2
Appendicitis	Instance	95.6	95.6	95.6	85.5	85.7	85.6	76.6	76.6	76.4
	Feature	3.2	3.6	3.5	2.4	2.2	2.1	2.7	1.3	1.5
Heart	Instance	241.9	242.0	241.8	218.0	217.6	217.9	193.8	193.9	193.7
	Feature	6.5	6.2	5.9	6.2	5.9	5.6	6.0	4.9	4.6
Fertility-diagnosis	Instance	89.6	89.9	89.8	80.7	80.6	80.5	71.5	71.6	71.4
	Feature	5.2	4.3	5.1	4.6	4.1	5.4	3.7	4.6	5.4
Wine	Instance	160.8	160.5	160.7	144.8	144.6	144.8	128.8	128.1	128.8
	Feature	4.1	4.0	4.1	4.1	4.0	4.0	4.0	3.6	3.7
German	Instance	909.0	909.0	908.9	888.5	890.4	885.1	886.2	890.2	884.3
	Feature	13.7	12.7	12.1	15.1	14.1	14.1	14.9	14.2	13.8
Hepatitis	Instance	139.5	139.9	139.9	125.7	125.9	127.4	111.7	115.6	112.9
	Feature	6.4	6.2	6.0	5.7	5.0	4.7	4.7	5.1	5.3
Flags-religion	Instance	176.0	175.9	175.9	158.0	158.0	157.9	141.0	140.8	140.9
	Feature	12.1	12.5	12.0	11.2	11.0	11.0	10.6	10.6	10.5
Leaf	Instance	308.6	308.8	308.9	277.7	277.8	277.6	247.8	248.0	248.0
	Feature	8.3	8.4	8.6	7.7	7.3	7.6	7.1	6.6	7.1
Lymphography	Instance	133.7	133.9	133.8	120.8	120.9	120.6	106.8	106.6	106.7
	Feature	7.4	7.1	6.7	7.2	6.6	6.3	6.7	6.0	5.9
Seeds	Instance	190.6	190.5	190.5	171.6	171.6	171.8	152.4	152.7	152.4
	Feature	4.5	4.0	4.0	4.0	3.5	3.0	3.1	3.0	3.0
Dbworld-bodies	Instance	64.0	64.0	64.0	64.0	63.1	64.0	64.0	64.0	64.0
	Feature	2282.8	2273.9	2280.2	2270.3	2282.8	2270.4	2273.2	2276.6	2274.9
Dbworld-bodies-stemmed	Instance	64.0	64.0	64.0	64.0	63.0	64.0	64.0	64.0	64.0
	Feature	1777.0	1794.9	1799.0	1797.1	1789.1	1790.9	1785.3	1795.0	1788.5
Micro-mass-mixed-spectra	Instance	360.0	360.0	360.0	360.0	360.0	360.0	360.0	360.0	360.0
	Feature	600.2	605.8	600.2	605.8	600.2	608.2	600.2	605.8	608.2



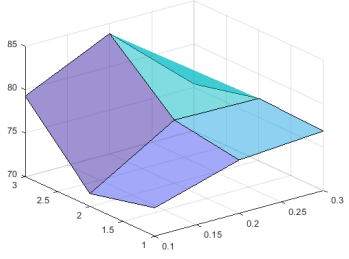
Diabetes



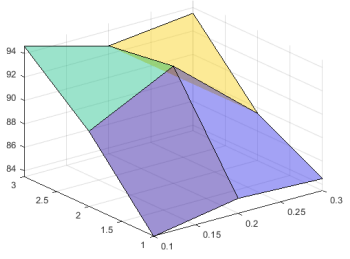
Glass



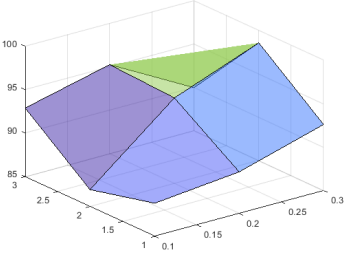
Appendicitis



Heart



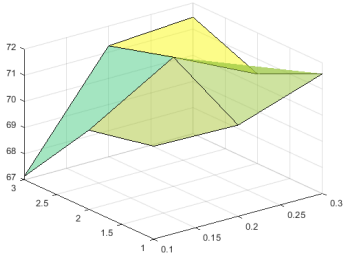
Fertility diagnosis



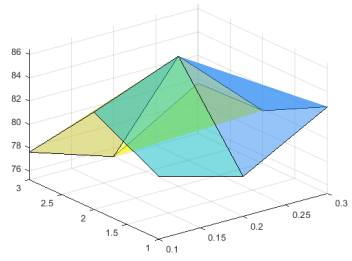
Wine

TABLE 6.5: IFBR classification accuracy for 90% coverage ($\epsilon = 0.1$)

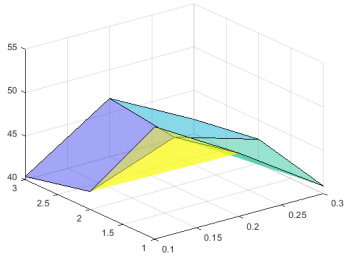
Dataset	**Classifier**	k = 1	k = 2	k = 3
Diabetes	KNN	74.67±1.01	**75.66±1.06**	71.19±1.32
	SVM	62.84±1.58	**67.26±1.97**	64.51±2.95
Glass	KNN	**72.40±2.10**	71.28±2.67	68.75±4.10
	SVM	59.23±4.47	63.30±3.67	**66.84±4.10**
Appendicitis	KNN	**93.52±2.56**	92.07±3.63	89.64±3.55
	SVM	88.96±2.37	**92.17±3.51**	91.43±2.72
Heart	KNN	73.28±3.10	76.14±1.16	**76.85±3.28**
	SVM	**58.18±2.13**	56.56±5.91	55.68±3.17
Fertility-diagnosis	KNN	83.50±1.92	**84.78±3.22**	84.51±5.68
	SVM	86.28±2.65	**91.48±5.01**	84.51±5.68
Wine	KNN	88.81±3.59	89.78±2.81	**92.59±1.30**
	SVM	**92.65±1.10**	89.97±1.13	89.29±3.13
German	KNN	70.57±0.70	70.50±1.09	**71.59±0.46**
	SVM	75.49±1.35	75.29±2.18	**75.59±1.34**
Hepatitis	KNN	80.65±5.62	78.69±1.38	**82.70±1.94**
	SVM	80.76±5.58	**85.09±4.87**	84.54±3.16
Flags-religion	KNN	**52.95±9.54**	47.27±2.58	40.85±2.85
	SVM	**52.68±4.74**	48.36±5.55	43.14±2.00
Leaf	KNN	65.19±2.23	65.19±1.66	**68.23±1.38**
	SVM	35.98±1.49	**39.33±1.44**	38.01±1.78
Lymphography	KNN	73.43±5.99	**78.91±5.67**	71.93±1.01
	SVM	78.34±4.60	**79.20±5.14**	71.48±6.02
Seeds	KNN	86.20±1.20	91.66±1.91	**93.52±1.35**
	SVM	90.83±1.38	**93.91±1.53**	93.16±0.94
Dbworld-bodies	KNN	**67.58±8.82**	50.22±6.48	57.71±11.25
	SVM	92.47±4.36	88.70±3.30	**96.10±2.41**
Dbworld-bodies-stemmed	KNN	47.47±2.49	**72.09±6.18**	61.32±10.95
	SVM	78.69±5.69	86.56±5.52	**93.20±4.43**
Micro-mass-mixed-spectra	KNN	**87.07±1.13**	80.63±1.02	**87.07±1.13**
	SVM	**72.49±3.62**	71.61±3.46	**72.49±3.62**



German



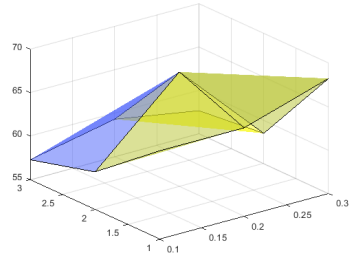
Hepatitis



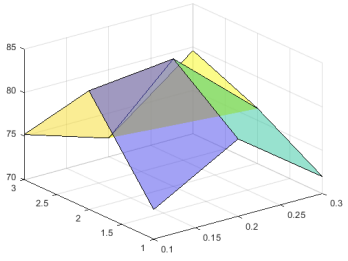
Flags religion

TABLE 6.6: IFBR classification accuracy for 80% coverage ($\epsilon = 0.2$)

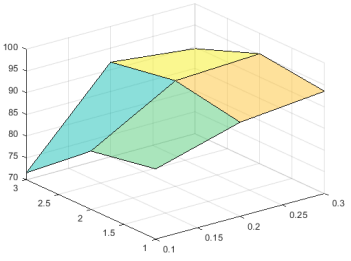
Dataset	**Classifier**	k = 1	k = 2	k = 3
Diabetes	KNN	72.29±1.14	**74.86±1.73**	71.52±2.81
	SVM	**69.57±1.46**	65.57±1.79	61.55±3.46
Glass	KNN	69.05±1.23	**69.73±2.28**	62.80±4.02
	SVM	61.15±3.91	**64.01±2.99**	61.65±2.68
Appendicitis	KNN	83.11±4.95	**89.75±4.19**	84.81±2.12
	SVM	83.11±4.95	**91.52±1.50**	81.62±7.23
Heart	KNN	71.48±3.90	**77.31±1.17**	77.15±3.87
	SVM	44.85±1.79	**72.07±1.74**	49.16±7.03
Fertility-diagnosis	KNN	89.91±3.53	**93.51±2.85**	87.51±4.51
	SVM	91.11±3.07	**93.51±2.85**	87.51±4.51
Wine	KNN	86.94±3.89	94.85±2.19	**98.55±1.47**
	SVM	86.83±3.94	**94.33±1.45**	92.06±1.48
German	KNN	70.04±0.65	**71.96±1.08**	70.43±1.01
	SVM	75.01±1.34	**75.12±1.97**	71.76±3.11
Hepatitis	KNN	83.61±2.26	**86.44±3.76**	79.80±2.50
	SVM	82.53±2.94	**88.53±5.35**	78.57±1.04
Flags-religion	KNN	42.06±1.96	**45.68±4.32**	42.80±2.44
	SVM	46.56±4.93	47.09±0.66	**52.46±4.07**
Leaf	KNN	59.32±2.63	**68.18±5.66**	58.48±1.63
	SVM	34.30±1.57	**35.15±3.36**	34.74±2.03
Lymphography	KNN	83.66±2.22	**84.69±3.68**	76.33±1.48
	SVM	76.33±2.39	**83.32±1.44**	73.60±2.41
Seeds	KNN	83.43±2.46	94.35±1.56	**95.28±2.25**
	SVM	89.61±1.77	**93.82±1.37**	93.69±0.97
Dbworld-bodies	KNN	52.95±2.48	**67.58±8.82**	55.71±5.40
	SVM	87.00±7.60	**92.43±4.36**	90.53±5.74
Dbworld-bodies-stemmed	KNN	48.51±2,78	**69.76±5.93**	61.32±10.95
	SVM	91.00±3.32	91.12±4.59	**93.20±4.43**
Micro-mass-mixed-spectra	KNN	80.65±1.02	**87.07±1.13**	79.71±2.54
	SVM	71.61±3.46	72.49±3.62	**78.37±2.28**



Leaf



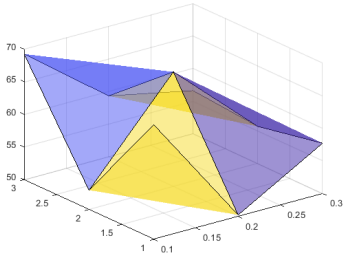
Lymphography



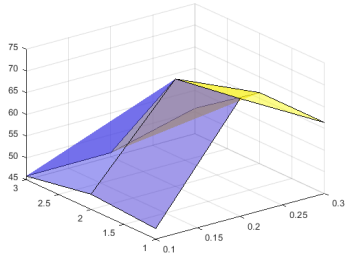
Seeds

TABLE 6.7: IFBR classification accuracy for 70% coverage ($\epsilon = 0.3$)

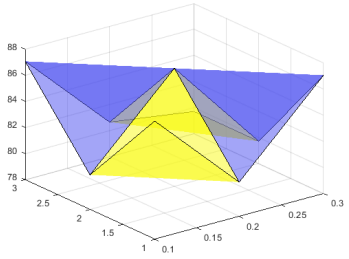
Dataset	**Classifier**	k = 1	k = 2	k = 3
Diabetes	KNN	73.77±1.92	74.62±0.08	**75.18±1.13**
	SVM	**65.64±1.13**	63.85±1.05	65.62±1.03
Glass	KNN	66.60±4.02	**73.53±4.90**	71.09±3.82
	SVM	63.10±1.49	**65.48±3.80**	63.97±5.19
Appendicitis	KNN	82.30±2.84	**92.44±2.14**	84.37±2.40
	SVM	82.30±2.84	**84.76±2.20**	83.56±2.26
Heart	KNN	79.26±5.6	**83.81±4.02**	75.38±1.89
	SVM	45.59±1.02	**55.83±5.49**	53.92±4.06
Fertility-diagnosis	KNN	**94.62±2.71**	92.69±1.76	93.52±4.70
	SVM	93.19±2.42	92.69±1.76	**93.52±4.70**
Wine	KNN	92.89±2.08	**95.24±3.63**	89.96±1.94
	SVM	89.62±1.28	91.29±1.85	**93.10±3.08**
German	KNN	67.12±1.52	71.25±2.67	**71.47±0.91**
	SVM	74.45±0.63	72.71±1.70	**75.30±0.86**
Hepatitis	KNN	77.62±2.05	75.27±6.44	**79.60±2.78**
	SVM	78.02±2.57	76.00±6.73	**80.88±2.44**
Flags-religion	KNN	40.35±2.26	**46.70±2.36**	41.66±1.58
	SVM	45.57±2.75	44.76±6.75	**55.23±3.23**
Leaf	KNN	57.28±4.77	**59.31±2.21**	57.65±1.63
	SVM	24.44±3.25	30.04±1.89	**30.17±2.89**
Lymphography	KNN	75.19±1.98	72.13±1.97	**79.57±4.97**
	SVM	79.58±0.97	69.63±4.54	**84.36±5.10**
Seeds	KNN	71.53±5.20	**91.74±1.99**	89.58±1.66
	SVM	74.94±5.81	**91.20±1.71**	86.96±1.53
Dbworld-bodies	KNN	**69.26±12.61**	59.32±4.64	56.61±3.70
	SVM	**90.67±4.08**	87.77±3.52	88.27±3.70
Dbworld-bodies-stemmed	KNN	45.81±3.36	45.90±7.10	**50.87±4.05**
	SVM	**89.44±4.30**	79.06±5.08	89.11±4.38
Micro-mass-mixed-spectra	KNN	**87.07±1.13**	80.65±1.02	79.71±2.54
	SVM	72.49±3.62	71.61±3.46	**78.73±2.28**



Dbworld-bodies



Dbworld-bodies-stemmed

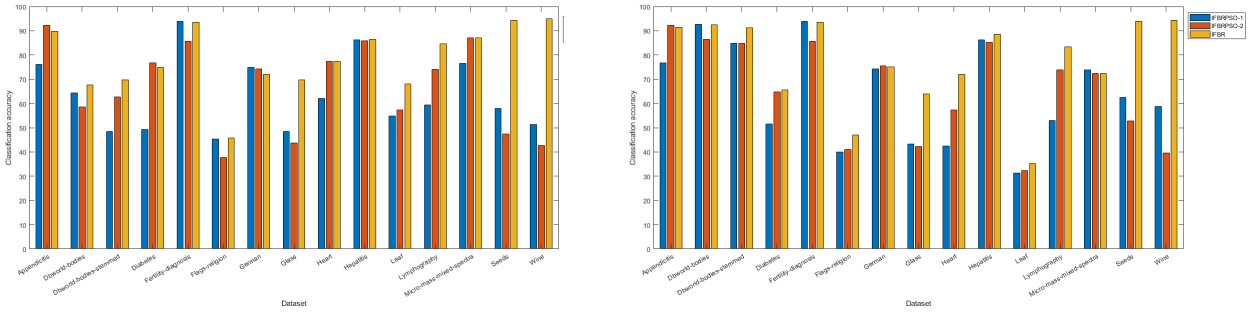


Micro-mass-mixed-spectra

FIGURE 6.4: Graphical visualization showing the variation of classification accuracy with dataset coverage (ϵ) and noise parameter k

6.2.1.2 Using Variants of Instance Selection

This section compares the performance and the size of bireducts as generated by two instance selection approaches described in Section 6.1.2, abbreviated as IFBRPSO-1 and IFBRPSO-2. Further comparison is made with ϵ-bireduct generated by IFBRPSO-2, abbreviated as IFBR. The entire comparison is summarised in table 6.8, 6.9 and 6.10. It can be clearly seen that IFBRPSO-1 removes considerable number of instances for few datasets, which might lead to information loss. IFBRPSO-2 also removes too many instances, but a fair reduction is observed for few datasets. To ensure minimum coverage, however, ϵ-bireduct is employed. In terms of feature subset size, all the three approaches gave identical results for most of the dataset. IFBR produced comparable or higher classification accuracy than IFBRPSO-1 and IFBRPSO-2 for most of the dataset-classifier combination as can be seen from corresponding box plot in figure 6.5. The overall reduction can be seen from table 6.10.



KNN SVM

FIGURE 6.5: Graphical visualization showing the classification accuracy with variants of IS

TABLE 6.8: IFBR results for variants of IS

Dataset	IFBRPSO-1		IFBRPSO-2		IFBR (ϵ-bireduct)	
	Instance	Feature	Instance	Feature	Instance	Feature
Diabetes	85.7	1.8	652.7	6.5	650.0	6.2
Glass	77.7	4.3	63.9	3.9	174.7	7.2
Appendicitis	39.8	2.6	57.5	2.0	85.7	2.2
Heart	76.7	2.7	149.9	4.0	217.6	5.9
Fertility-diagnosis	81.9	4.9	52.9	4.9	80.6	4.1
Wine	30.0	1.7	28.1	2.5	144.6	4.0
German	648.0	10.1	888.9	13.4	890.4	14.1
Hepatitis	92.6	4.5	122.7	7.5	125.9	5.0
Flags-religion	115.5	10.7	82.8	10.6	158.0	11.0
Leaf	205.2	6.4	207.7	7.2	277.8	7.3
Lymphography	71.5	4.5	112.3	8.8	120.9	6.6
Seeds	47.5	1.3	38.8	1.4	171.6	3.5
Dbworld-bodies	64.0	2276.9	64.0	2271.0	63.1	2282.8
Dbworld-bodies-stemmed	64.0	1796.3	64.0	1791.9	63.0	1789.1
Micro-mass-mixed-spectra	360.0	606.6	360.0	600.2	360.0	600.2

TABLE 6.9: IFBR classification accuracy employing variants of IS

Dataset	Classifier	IFBRPSO-1	IFBRPSO-2	IFBR (ϵ-bireduct)
Diabetes	KNN	49.31±3.73	**76.67±1.51**	74.86±1.73
	SVM	51.53±3.70	64.75±2.67	**65.57±1.79**
Glass	KNN	48.49±3.29	43.66±3.95	**69.73±2.28**
	SVM	43.29±1.89	42.23±3.87	**64.01±2.99**
Appendicitis	KNN	76.15±2.71	**92.24±4.20**	89.75±4.19
	SVM	76.83±3.16	**92.24±4.20**	91.52±1.50
Heart	KNN	62.04±9.50	**77.38±1.39**	77.31±1.17
	SVM	42.55±4.07	57.26±2.74	**72.07±1.74**
Fertility-diagnosis	KNN	**93.97±4.12**	85.57±2.72	93.51±2.85
	SVM	**93.97±4.12**	85.57±2.72	93.51±2.85
Wine	KNN	51.40±5.03	42.76±3.23	**94.85±2.19**
	SVM	58.84±5.55	39.56±6.71	**94.33±1.45**
German	KNN	**74.98±1.66**	74.22±2.86	71.96±1.08
	SVM	74.34±1.57	**75.50±2.31**	75.12±1.97
Hepatitis	KNN	86.27±3.12	85.91±1.38	**86.44±3.76**
	SVM	86.19±2.64	85.19±1.99	**88.53±5.35**
Flags-religion	KNN	45.43±5.11	37.79±5.49	**45.68±4.32**
	SVM	39.91±1.83	41.08±4.15	**47.09±0.66**
Leaf	KNN	54.94±4.82	57.40±5.03	**68.18±5.66**
	SVM	31.37±3.85	32.35±4.63	**35.15±3.36**
Lymphography	KNN	59.40±6.29	74.03±6.34	**84.69±3.68**
	SVM	52.92±7.81	73.77±2.50	**83.32±1.44**
Seeds	KNN	57.98±4.39	47.48±3.80	**94.35±1.56**
	SVM	62.54±2.58	52.82±5.81	**93.82±1.37**
Dbworld-bodies	KNN	64.41±8.00	58.61±4.45	**67.58±8.82**
	SVM	**92.75±5.51**	86.42±7.67	92.43±4.36
Dbworld-bodies-stemmed	KNN	48.44±7.06	62.65±7.79	**69.76±5.93**
	SVM	84.85±6.73	84.90±1.43	**91.12±4.59**
Micro-mass-mixed-spectra	KNN	76.65±2.81	**87.07±1.13**	**87.07±1.13**
	SVM	73.89±4.32	**72.49±3.62**	**72.49±3.62**

TABLE 6.10: IFBR overall reduction rate employing variants of IS

Dataset	IFBRPSO-1	IFBRPSO-2	IFBR (ϵ-bireduct)
Diabetes	97.48926	30.94808	34.40755
Glass	82.65265	87.06075	34.69159
Appendicitis	86.05391	84.50135	74.5903
Heart	94.03371	82.72544	63.01238
Fertility-diagnosis	55.41	71.19889	63.28222
Wine	97.79602	96.96413	75.00432
German	72.73	50.36975	47.689
Hepatitis	85.85059	68.75212	78.62479
Flags-religion	77.24871	83.84242	68.00442
Leaf	72.41008	68.58319	57.39622
Lymphography	87.9223	62.9039	70.0473
Seeds	95.79932	96.30476	59.14286
Dbworld-bodies	51.57593	51.7014	52.1331
Dbworld-bodies-stemmed	51.72534	51.84359	52.6701
Micro-mass-mixed-spectra	53.33846	53.83077	53.83077

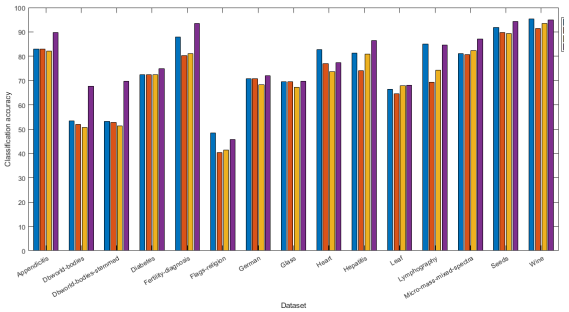
6.2.1.3 Comparisons with Other FS Algorithms

A comparison of proposed approach with other state of the art methods is dealt in this section. An IF rough feature selection (IFFS) approach [140] and a fuzzy rough feature selection (FRPSO) [104] both employing particle swarm optimization are used for comparisons. The experimental results are shown in table 6.11, 6.12, 6.13. The proposed intuitionistic fuzzy ϵ-bireduct (IFBR) while reducing number of instances decreases the size of feature subset for all the datasets except for Glass, German, Dbworld-bodies and Dbworld-bodies-stemmed, in which case the difference in feature subset size between various approaches is insignificant. IFBR produces a significant increase in classification accuracy for the all datasets than IFFS, and FRPSO except for Leaf and Micro-mass-mixed-spectra. IFFS and FRPSO gave poor performance for thirteen datasets. Since, intuitionistic fuzzy rough sets and particle swarm search heuristic is employed in this approach, superiority of IFBR over IFFS and FRPSO clearly emphasis the effectiveness of the proposed work. The bar plot (figure 6.6) clearly shows the superiority of the approach. The performance

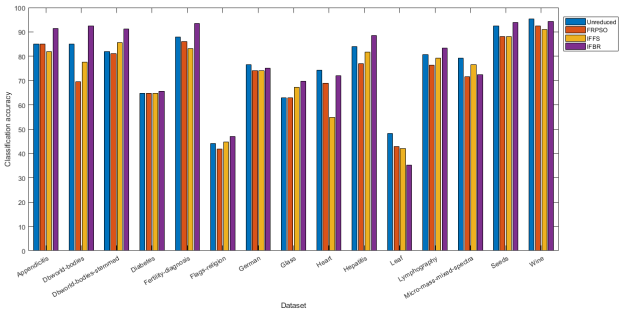
TABLE 6.11: Comparison with other state of art feature selection algorithms

Dataset	**FRPSO**	**IFFS**	**IFBR**	
	Feature	Feature	Instance	Feature
Diabetes	8.0	8.0	650.0	**6.2**
Glass	9.0	**6.0**	174.7	7.2
Appendicitis	7.0	4.0	85.7	**2.2**
Heart	12.9	7.4	217.6	**5.9**
Fertility-diagnosis	9.0	6.0	80.6	**4.1**
Wine	9.7	6.0	144.6	4.0
German	16.8	**12.1**	890.4	14.1
Hepatitis	13.4	8.2	125.9	**5.0**
Flags-religion	18.9	14.5	158.0	**11.0**
Leaf	14.0	9.0	277.8	**7.3**
Lymphography	11.7	7.4	120.9	**6.6**
Seeds	7.0	4.0	171.6	**3.5**
Dbworld-bodies	**2274.9**	2281.6	63.1	2282.8
Dbworld-bodies-stemmed	**1783.7**	1786.4	63.0	1789.1
Micro-mass-mixed-spectra	605.8	609.7	360.0	**600.2**

is even higher than unreduced case except for flags-religion, heart and wine for 3NN and heart, leaf, mixed-mass-micro-spectra and wine for svm. The value of $F(2, 28) = 3.34$ at $\alpha = 5\%$ level for significance, therefore the null hypothesis is rejected using Freidman test, i.e. four algorithms are statistically different. For Bonferroni Dunn test, $q_{0.05} = 2.241$ so $Cd_{0.05} = 0.818$. Hence, IFBR is statistically better than FRPSO and IFFS for both the classifiers at 5% level of significance.



KNN



SVM

TABLE 6.12: Classification accuracy comparison with other state of art feature selection algorithms

Dataset	Classifier	FRPSO	IFFS	IFBR
Diabetes	KNN	72.50±5.70$^{2.5}$	72.50±5.70$^{2.5}$	**74.86**±1.73^{1}
	SVM	64.86±6.35$^{2.5}$	64.86±6.35$^{2.5}$	**65.57**±1.79^{1}
Glass	KNN	69.52±9.03^{2}	67.24±5.17^{3}	**69.73**±2.28^{1}
	SVM	62.85±7.02^{2}	56.64±1.40^{3}	**64.01**±2.99^{1}
Appendicitis	KNN	83.00±9.48^{2}	82.21±4.68^{3}	**89.75**±4.19^{1}
	SVM	85.00±7.07^{2}	81.91±4.59^{3}	**91.52**±1.50^{1}
Heart	KNN	77.01±2.04^{2}	73.62±3.40^{3}	**77.31**±1.17^{1}
	SVM	68.95±3.25^{2}	54.78±4.33^{3}	**72.07**±1.74^{1}
Fertility-diagnosis	KNN	80.27±2.96^{3}	81.12±2.33^{2}	**93.51**±2.85^{1}
	SVM	86.06±5.24^{2}	83.19±2.19^{3}	**93.51**±2.85^{1}
Wine	KNN	91.36±3.51^{3}	93.54±1.27^{2}	**94.85**±2.19^{1}
	SVM	92.50±1.69^{2}	91.02±1.82^{3}	**94.33**±1.45^{1}
German	KNN	70.81±1.25^{2}	68.21±1.34^{3}	**71.96**±1.08^{1}
	SVM	74.10±1.01^{3}	74.12±1.68^{2}	**75.12**±1.97^{1}
Hepatitis	KNN	74.02±7.46^{3}	80.83±2.44^{2}	**86.44**±3.76^{1}
	SVM	76.87±4.49^{3}	81.73±2.92^{2}	**88.53**±5.35^{1}
Flags-religion	KNN	40.41±2.40^{3}	41.45±2.39^{2}	**45.68**±4.32^{1}
	SVM	41.77±2.96^{3}	44.65±1.69^{2}	**47.09**±0.66^{1}
Leaf	KNN	64.47±1.67^{3}	67.87±2.80^{2}	**68.18**±5.66^{1}
	SVM	**42.94**±3.03^{1}	41.98±1.97^{2}	35.15±3.36^{3}
Lymphography	KNN	69.41±8.75^{3}	74.28±3.92^{2}	**84.69**±3.68^{1}
	SVM	76.36±4.21^{3}	79.22±3.30^{2}	**83.32**±1.44^{1}
Seeds	KNN	89.74±3.90^{2}	89.34±3.18^{3}	**94.35**±1.56^{1}
	SVM	88.12±2.54^{2}	88.04±4.55^{3}	**93.82**±1.37^{1}
Dbworld-bodies	KNN	52.06±2.40^{2}	50.63±6.92^{3}	**67.58**±8.82^{1}
	SVM	69.55±4.98^{3}	77.64±6.07^{2}	**92.43**±4.36^{1}
Dbworld-bodies-stemmed	KNN	52.89±4.01^{2}	51.43±7.30^{3}	**69.76**±5.93^{1}
	SVM	81.10±6.48^{3}	85.72±1.73^{2}	**91.12**±4.59^{1}
Micro-mass-mixed-spectra	KNN	80.65±1.02^{3}	82.38±1.79^{2}	**87.07**±1.13^{1}
	SVM	71.61±3.46^{3}	**76.50**±4.54^{1}	72.49±3.62^{2}
Average Rank	KNN	2.5	2.5	1
	SVM	2.43	2.36	1.20
F statistics	KNN	42.0		
	SVM	11.7		

TABLE 6.13: Overall redcution rate comparison with other state of art feature selection algorithms

Dataset	FRPSO	IFFS	IFBR
Diabetes	0	0	34.40755
Glass	0	33.33333	34.69159
Appendicitis	0	42.85714	74.5903
Heart	0.769231	43.07692	63.01238
Fertility-diagnosis	0	33.33333	63.28222
Wine	25.38462	53.84615	75.00432
German	30	49.58333	47.689
Hepatitis	29.47368	56.84211	78.62479
Flags-religion	32.5	48.21429	68.00442
Leaf	0	35.71429	57.39622
Lymphography	35	58.88889	70.0473
Seeds	0	42.85714	59.14286
Dbworld-bodies	51.61846	51.47597	52.1331
Dbworld-bodies-stemmed	52.06396	51.9914	52.6701
Micro-mass-mixed-spectra	53.4	53.1	53.83077

FIGURE 6.6: Graphical visualization showing comparison of the classification accuracy with state of the art feature selection approach

6.2.1.4 Comparison with Instance Selection and Feature Selection + Instance Selection Approaches

The above presented comparative approaches are based on feature selection alone and it lacks proper comparative analysis of IFBR. Since there does not exist any previous intuitionistic fuzzy bireduct approach, so a combination of feature selection and instance selection is employed for comparison purpose. Two sets of algorithms namely FSIS and ISFS are used. In FSIS, irrelevant features are removed using IFFS [140] using paticle swarm search heuristic and then outliers are eliminated from this reduced dataset using instance selection [70]. While in ISFS, problematic instances

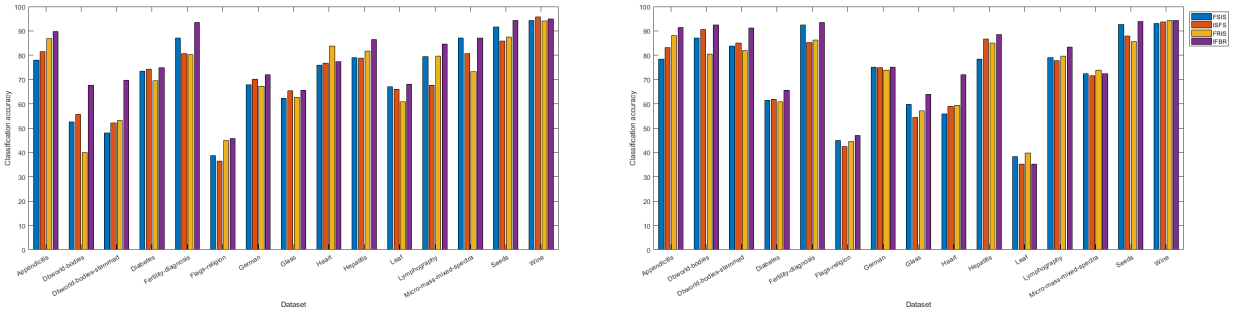
TABLE 6.14: Comparison with Instance Selection-Feature Selection combination

Dataset	FSIS		ISFS		FRIS		IFBR	
	Instance	Feature	Instance	Feature	Instance	Feature	Instance	Feature
Diabetes	760.2	8	759.8	8	768	8	**650**	**6.2**
Glass	207	**6**	208.2	**6**	212.4	9	**174.7**	7.2
Appendicitis	101.6	3.4	106	3.7	104.1	7	**85.7**	**2.2**
Heart	255.1	7.4	265.4	7.4	267	13	**217.6**	**5.9**
Fertility-diagnosis	98.4	6.3	98.6	5	99	9	**80.6**	**4.1**
Wine	177.8	5.9	176.4	6	165.7	13	**144.6**	**4**
German	999	12.5	996.2	**12.3**	996.3	24	**890.4**	14.1
Hepatitis	151.1	8	155	8	148.7	19	**125.9**	**5**
Flags-religion	183	14.8	191	14.2	185.8	28	**158**	**11**
Leaf	292.5	9	315.5	9	340	14	**277.8**	**7.3**
Lymphography	146.8	7.4	148	7.8	147.4	18	**120.9**	**6.6**
Seeds	198.9	4	193	4	191.8	7	**171.6**	**3.5**
Dbworld-bodies	64.0	**2270.3**	64.0	2270.4	64.0	4702	**63.1**	**2282.8**
Dbworld-bodies-stemmed	64.0	1790.9	64.0	1791.1	64.0	3721	**63.0**	**1789.1**
Micro-mass-mixed-spectra	**360**	**600.2**	**360**	605.8	**360**	1300	**360**	**600.2**

are first eliminated using [70] followed by feature selection with paticle swarm search heuristic in IFFS [140]. Finally, these two reduced datasets are evaluated for performance. A comparison with instance selection (FRIS) [70] alone is also undertaken. Table 6.14 shows the number of features and instances selected and overall reduction thus achieved in table 6.16 while model performance is illustrated in table 6.15. IFBR produces the best reduction in data size except for Micro-mass-mixed-spectra for which the reduction size are comparable. IFBR outperforms FSIS, ISFS, and FRIS for all the datasets except Heart, Wine, Leaf, and Micro-mass-mixed-spectra in which case, the difference is insignificant. FSIS, ISFS performs very poorly for Dbworld-bodies-stemmed and FRIS for Dbworld-bodies dataset respectively. Figure 6.7 gives a better visualization for comparing performance. In summary, IFBR generates more consistent bireducts that performs better than other instance selection and feature selection combinations. For statistical testing, $M = 15, N = 4$, so the value of $F(3, 42) = 2.847$ is used. The null hypothesis is rejected implying the significant difference between algorithms. Here, $q_{0.05} = 2.394$ such that $Cd_{0.05} = 1.12$. The null hypothesis is again rejected by Bonferroni Dunn test for both classifiers demonstrating the superiority of the proposed approach.

TABLE 6.15: Classification accuracy comparison with Instance Selection-Feature Selection combination

Dataset	Classifier	FSIS	ISFS	FRIS	IFBR
Diabetes	KNN	73.50±2.10^{3}	74.28±3.08^{2}	69.55±2.12^{4}	**74.86**±1.73^{1}
	SVM	61.44±1.39^{3}	61.95±2.03^{2}	60.95±4.13^{4}	**65.57**±1.79^{1}
Glass	KNN	62.30±3.53^{4}	65.39±4.47^{2}	62.69±2.85^{3}	**69.73**±2.28^{1}
	SVM	59.80±4.34^{2}	54.40±6.60^{4}	57.17±5.36^{3}	**64.01**±2.99^{1}
Appendicitis	KNN	77.90±6.77^{4}	81.61±4.41^{3}	86.85±2.08^{2}	**89.75**±4.19^{1}
	SVM	78.51±3.50^{4}	83.06±4.88^{3}	88.16±2.91^{2}	**91.52**±1.50^{1}
Heart	KNN	76.01±1.18^{4}	76.83±2.10^{3}	**83.72**±2.74^{1}	77.31±1.17^{2}
	SVM	55.80±2.12^{4}	58.92±0.80^{3}	59.32±8.75^{2}	**72.07**±1.74^{1}
Fertility-diagnosis	KNN	87.18±2.07^{2}	80.75±4.16^{3}	80.20±8.14^{4}	**93.51**±2.85^{1}
	SVM	92.41±3.12^{2}	85.23±4.07^{4}	86.35±6.56^{3}	**93.51**±2.85^{1}
Wine	KNN	94.31±2.34^{3}	**95.73**±0.96^{1}	94.03±0.75^{4}	94.85±2.19^{2}
	SVM	93.06±2.53^{4}	93.74±1.98^{3}	94.24±1.27^{2}	**94.33**±1.45^{1}
German	KNN	67.85±1.28^{3}	70.14±1.25^{2}	67.30±1.83^{4}	**71.96**±1.08^{1}
	SVM	75.10±0.59^{2}	74.92±1.09^{3}	73.86±1.32^{4}	**75.12**±1.97^{1}
Hepatitis	KNN	79.01±3.15^{3}	78.79±2.47^{4}	81.67±3.34^{2}	**86.44**±3.76^{1}
	SVM	78.51±3.18^{4}	86.65±3.60^{2}	85.03±2.31^{3}	**88.53**±5.35^{1}
Flags-religion	KNN	38.65±1.42^{3}	36.51±3.78^{4}	44.92±5.97^{2}	**45.68**±4.32^{1}
	SVM	44.99±3.34^{2}	42.55±2.81^{4}	44.55±3.96^{3}	**47.09**±0.66^{1}
Leaf	KNN	67.09±1.29^{2}	65.91±1.77^{3}	60.76"±1.22^{4}	**68.18**±5.66^{1}
	SVM	38.24±4.73^{2}	35.33±3.62^{3}	**39.76**±5.00^{1}	35.15±3.36^{4}
Lymphography	KNN	79.52±4.25^{2}	67.62±4.40^{4}	78.77±3.01^{3}	**84.69**±3.68^{1}
	SVM	78.93±0.93^{3}	77.72±3.67^{4}	79.73±3.21^{2}	**83.32**±1.44^{1}
Seeds	KNN	91.64±3.22^{2}	85.82±1.41^{4}	87.40±1.81^{3}	**94.35**±1.56^{1}
	SVM	92.67±2.74^{2}	87.86±2.70^{3}	85.63±1.12^{4}	**93.82**±1.37^{1}
Dbworld-bodies	KNN	52.59±2.48^{3}	55.77±5.40^{2}	39.90±9.64^{4}	**67.58**±8.82^{1}
	SVM	87.00±7.60^{3}	90.53±5.74^{2}	80.40±4.61^{4}	**92.43**±4.36^{1}
Dbworld-bodies-stemmed	KNN	48.13±6.13^{4}	52.18±0.15^{3}	53.25±3.26^{2}	**69.76**±5.93^{1}
	SVM	83.87±7.48^{3}	85.05±0.12^{2}	81.93±1.57^{4}	**91.12**±4.59^{1}
Micro-mass-mixed-spectra	KNN	**87.07**±1.13$^{1.5}$	80.65±1.02^{3}	73.25±1.10^{4}	**87.07**±1.13$^{1.5}$
	SVM	72.49±3.62$^{2.5}$	71.61±3.46^{4}	**73.88**±1.91^{1}	72.49±3.62$^{2.5}$
Average Rank	KNN	2.90	2.86	3.06	1.16
	SVM	2.83	3.06	2.79	1.30
F statistics	KNN	11.91			
	SVM	8.19			



KNN SVM

FIGURE 6.7: Graphical visualization showing comparison of the classification accuracy with other IS-FS combination

TABLE 6.16: Overall reduction rate comparison with Instance Selection-Feature Selection combination

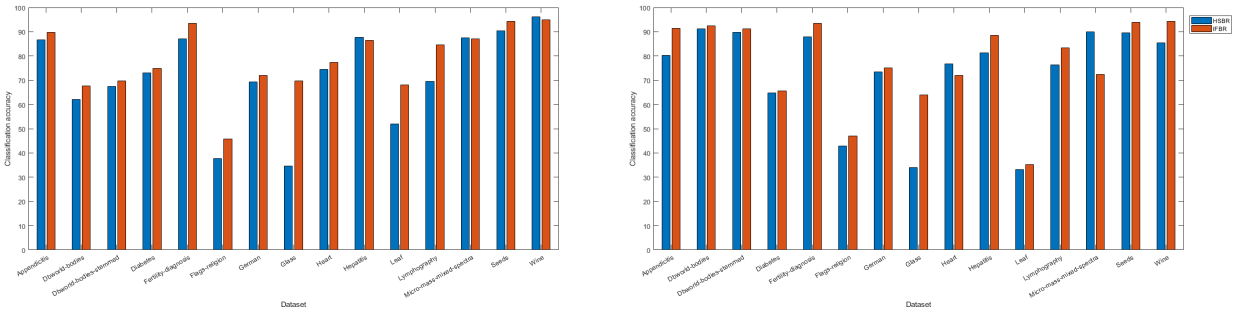
Dataset	FSIS	ISFS	FRIS	IFBR
Diabetes	1.015625	1.067708	0	34.40755
Glass	35.51402	35.14019	0.747664	34.69159
Appendicitis	53.44474	47.14286	1.792453	74.5903
Heart	45.61394	43.41804	0	63.01238
Fertility-diagnosis	31.12	45.22222	1	63.28222
Wine	54.66638	54.26102	6.910112	75.00432
German	47.96875	48.94475	0.37	47.689
Hepatitis	58.95416	57.89474	4.064516	78.62479
Flags-religion	50.13991	50.06996	4.226804	68.00442
Leaf	44.69538	40.34664	0	57.39622
Lymphography	59.22222	56.66667	0.405405	70.0473
Seeds	45.87755	47.48299	8.666667	59.14286
Dbworld-bodies	51.71629	51.71416	0	52.1331
Dbworld-bodies-stemmed	51.87046	51.86509	0	52.6701
Micro-mass-mixed-spectra	53.83077	53.4	0	53.83077

6.2.1.5 Comparison with existing Bireduct approach

The above presented comparative approaches are based on feature selection, instance selection or combination of both. However, a proper comparative analysis can be done on comparison with existing bireduct approach. A fuzzy rough bireduct approach [103] (HSBR) based on harmony search is employed for comparison purpose. Table 6.17 shows the number of features and instances selected and the overall reduction rate in table 6.19 while model performance is illustrated in table 6.18. IFBR produces the best reduction in data size except for Dbworld-bodies and Micro-mass-mixed-spectra for which the reduction size are comparable while a considerable decrease in feature subset and hence increase in overall reduction rate except for Glass. IFBR outperforms HSBR for nearly all the datasets as can be illustrated from Figure 6.8. To show statistical significance, the $q_{0.05} = 1.96$ is used thereby $Cd_{0.05} = 0.50$. The null hypothesis is rejected demonstrating the superiority of proposed approach.

TABLE 6.17: Comparison with Bireduct approach

Dataset	HSBR		IFBR	
	Instance	Feature	Instance	Feature
Diabetes	698.0	7.0	**650.0**	**6.2**
Glass	194.0	**2.0**	**174.7**	7.2
Appendicitis	96.0	6.0	**85.7**	**2.2**
Heart	242.7	8.5	**217.6**	**5.9**
Fertility-diagnosis	91.2	7.0	**80.6**	**4.1**
Wine	175.0	7.0	**144.6**	**4.0**
German	910.0	16.0	**890.4**	**14.1**
Hepatitis	145.2	12.7	**125.9**	**5.0**
Flags-religion	193.0	18.0	**158.0**	**11.0**
Leaf	309.0	10.1	**277.8**	**7.3**
Lymphography	138.0	11	**120.9**	**6.6**
Seeds	191.0	6.0	**171.6**	**3.5**
Dbworld-bodies	**63.0**	2365.0	63.1	**2282.8**
Dbworld-bodies-stemmed	**63.0**	1837.0	**63.0**	**1789.1**
Micro-mass-mixed-spectra	**359.0**	631.0	360.0	**600.2**



KNN SVM

FIGURE 6.8: Graphical visualization showing comparison of the classification accuracy with Bireduct approach

6.3 Application to Cancer Treatment

Oxygen and nutrients are the most essential elements of mammalian cells for their survival. Consequently, these cells are located within 100 to 200 mm of blood vessels,

TABLE 6.18: Classification accuracy comparison with Bireduct approach

Dataset	Classifier	HSBR	IFBR
Diabetes	KNN	73.03±3.63[2]	**74.86**±1.73[1]
	SVM	64.71±2.11[2]	**65.57**±1.79[1]
Glass	KNN	34.70±6.38[2]	**69.73**±2.28[1]
	SVM	34.09±3.44[2]	**64.01**±2.99[1]
Appendicitis	KNN	86.73±8.37[2]	**89.75**±4.19[1]
	SVM	80.18±2.77[2]	**91.52**±1.50[1]
Heart	KNN	74.50±10.86[2]	**77.31**±1.17[1]
	SVM	**76.77**±9.60[1]	72.07±1.74[2]
Fertility-diagnosis	KNN	87.00±4.83[2]	**93.51**±2.85[1]
	SVM	88.00±4.22[2]	**93.51**±2.85[1]
Wine	KNN	**96.08**±3.77[1]	94.85±2.19[2]
	SVM	85.36±9.97[2]	**94.33**±1.45[1]
German	KNN	69.40±4.88[2]	**71.96**±1.08[1]
	SVM	73.40±3.34[2]	**75.12**±1.97[1]
Hepatitis	KNN	**87.71**±5.65[1]	86.44±3.76[2]
	SVM	81.21±7.81[2]	**88.53**±5.35[1]
Flags-religion	KNN	37.71±11.02[2]	**45.68**±4.32[1]
	SVM	42.79±7.39[2]	**47.09**±0.66[1]
Leaf	KNN	52.06±5.38[2]	**68.18**±5.66[1]
	SVM	33.24±8.09[2]	**35.15**±3.36[1]
Lymphography	KNN	69.62±13.34[2]	**84.69**±3.68[1]
	SVM	76.38±10.46[2]	**83.32**±1.44[1]
Seeds	KNN	90.48±2.24[2]	**94.35**±1.56[1]
	SVM	89.52±7.71[2]	**93.82**±1.37[1]
Dbworld-bodies	KNN	62.14±11.53[2]	**67.58**±8.82[1]
	SVM	91.19±12.19[2]	**92.43**±4.36[1]
Dbworld-bodies-stemmed	KNN	67.38±14.98[2]	**69.76**±5.93[1]
	SVM =	89.76±11.88[2]	**91.12**±4.59[1]
Micro-mass-mixed-spectra	KNN	**87.50**±2.70[1]	87.07±1.13[2]
	SVM	64.17±6.47[2]	**72.49**±3.62[1]
Average Rank	KNN	1.80	1.13
	SVM	2.06	1.06

TABLE 6.19: Overall reduction rate comparison with Bireduct approach

Dataset	HSBR	IFBR
Diabetes	20.47526	34.40755
Glass	79.85462	34.69159
Appendicitis	22.37197	74.5903
Heart	40.56612	63.01238
Fertility-diagnosis	29.06667	63.28222
Wine	47.06137	75.00432
German	39.33333	47.689
Hepatitis	37.38404	78.62479
Flags-religion	36.04566	68.00442
Leaf	34.43487	57.39622
Lymphography	43.01802	70.0473
Seeds	22.04082	59.14286
Dbworld-bodies	50.48816	52.1331
Dbworld-bodies-stemmed	51.40293	52.6701
Micro-mass-mixed-spectra	51.59637	53.83077

which is the diffusion boundary for oxygen [20]. Therefore, multi-cellular organisms need new blood vessels to grow beyond this size for maintaining homeostasis and support growth [169]. Angiogenesis is known as a process of new blood vessel formations from pre-existing vessels which incorporates numerous biological behaviours, such as migration, apoptosis, endothelial cell proliferation, cell-cell and cell-matrix adhesion [13]. Angiogenesis is an extremely organized physiological process in growth as well as development. This process plays a key role in the formation of malignant tumours. Tumours employ angiogenesis to produce the vascular network, which is used to supply the cancer cells with oxygen and nutrients. Thus, anti-angiogenic

peptides are always capable candidates in the treatment of cancer [147]. Angiogenesis is a crucial physical process and is responsible for many diseases, such as cancer, myocardial ischemia, arthritis, myocardial infarction, and psoriasis. In the recent years, recognition of the anti-angiogenic peptides among other therapeutic peptides has drawn great attentions of the researchers in the cancer treatment area [46, 54, 169]. Cancer is a leading public health problem as it is one of the most fatal diseases world-wide. WHO has reported that cancer is the major cause of mortality in economically developed countries while the second major cause of mortality in developing countries. Cancer is still the third major cause of death after stroke and heart disease despite the fact that there are several advanced treatment schemes such as radiation, surgery, chemotherapy, and various diagnostic tests available in the literature [47, 149]. Nowadays, inhibiting angiogenesis is a pioneering area of research in cancer therapy [27, 170]. However, computational detection of anti-angiogenic peptides is rarely discussed in the literature.

To facilitate comparisons with the previous study for anti-angiogenic peptide prediction, the benchmark dataset introduced by Ramaprasad et. al. [123] is utilized. This dataset consisted of 135 positive (anti-angiogenic peptides) and 135 negative samples (non anti-angiogenic peptides). Ramaprasad et. al. selected 257 anti-angiogenic peptides from different research articles and patents to construct positive instances. CD-HIT technique is applied to eliminate highly similar sequences to ensure no two sequences contain more than 70% sequence similarity. Moreover, 135 random peptide regions from proteins available in Swiss-Prot database is extracted to construct negative instances.

6.3.0.1 Results

The selection of a comprehensive and appropriate feature vector from peptide samples that can actually reflect their intrinsic correlation with the properties to be predicted is an essential task to establish a powerful predictor. A suitable feature representation is the key to success of classifier learning as it facilitates the classifiers to easily identify underlying regularities. Six features namely Amino acid composition (AAC), Dipeptide composition (DPC), Pseudo amino acid composition (PAAC), Amphiphilic pseudo amino acid composition (AmPAAC), C/T/D composition (CTD) and Amino acid index (AAI) are extracted using iFeature [24] web server. Applying the proposed IFBR model to the dataset reduces the size of dataset thereby enhancing classification accuracy. Ensembles of classifiers [85] namely RealAdaBoost [15] with Random Forest classifier, Random Forest [16], and Rotation Forest [125] is formed via Vote [2] based classification technique in Weka [56], which is deployed to measure prediction performance of reduced anti-angiogenic peptides dataset using 10×10-fold cross validation. The flowchart of the entire methodology is shown in figure 6.9. Accuracy, which is the number of correctly classified peptides (including both anti-angiogenic and angiogenic peptides), sensitivity is the number of correctly predicted anti-angiogenic peptides while specificity, the number of correctly predicted angiogenic peptides are measured for the reduced dataset produced by IFBR. The value of the above defined performance parameter is noted for $\epsilon = 0.1, 0.2, 0.3$, as recorded in table 6.20 and 6.21. Highest accuray of 78.2% is obtained by covering 70% of the dataset in the generated bireduct.

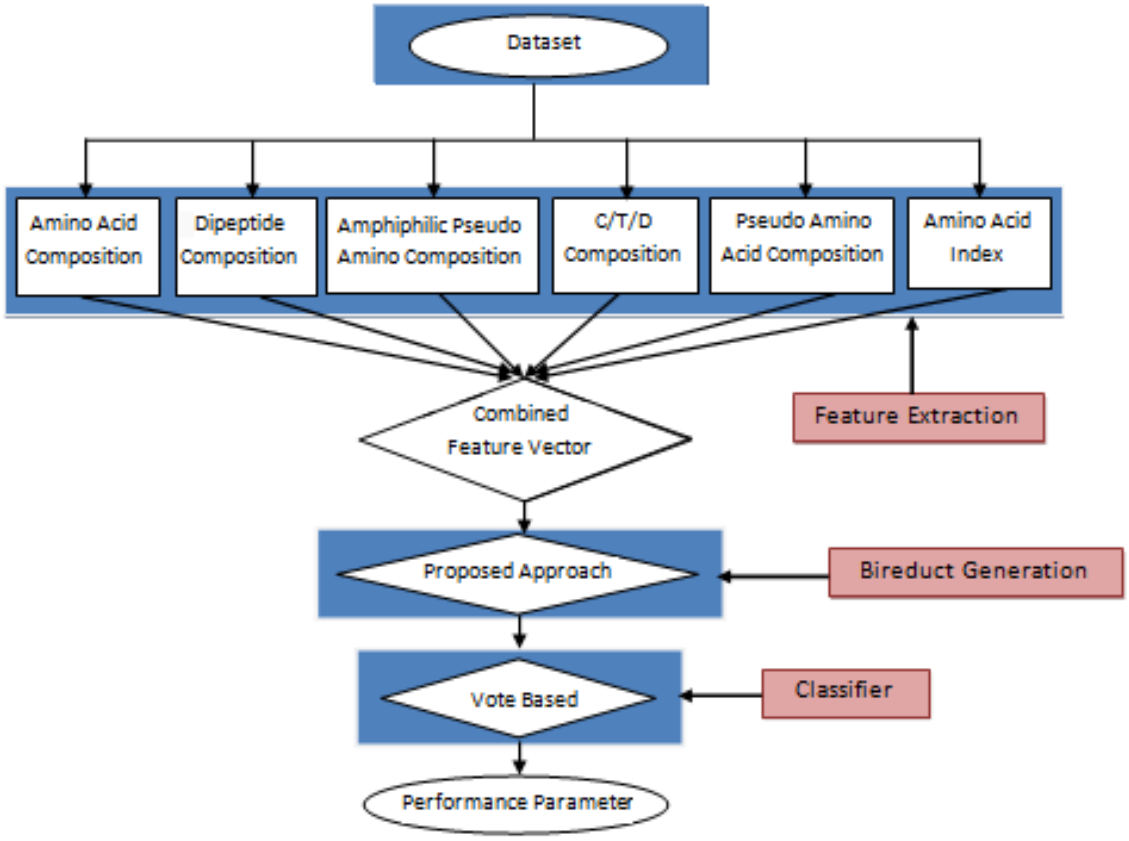


FIGURE 6.9: The flowchart of proposed cancer treatment model

TABLE 6.20: IFBR results for various ϵ values on Anti-angiogenic dataset

Dataset	**Original**		**Coverage** (ϵ)	**IFBR** (ϵ-**bireduct**)	
	Instance	Feature		Instance	Feature
Anti-angiogenic Dataset	270	460	0.1	242.0	233.0
			0.2	215.0	254.0
			0.3	188.0	240.0

TABLE 6.21: Performance evaluation metrics values on Anti-angiogenic dataset with IFBR

Dataset	**Coverage** (ϵ)	**Sensitivity**	**Specificity**	**Accuracy**
Anti-angiogenic Dataset	0.1	81.0	67.2	74.4
	0.2	81.0	72.7	76.7
	0.3	84.8	70.8	78.2

TABLE 6.22: Comparison of IFBR on Anti-angiogenic dataset with unreduced dataset

Method	Sensitivity	Specificity	Accuracy
Original	77.0	70.4	73.7
IFBR	84.8	70.8	78.2

6.3.0.2 Comparison with Unreduced Dataset

From table 6.20, it can be effectively seen that the size of dataset is considerably reduced in terms of both number of instances and feature vector size. The comparison with unreduced dataset is reported in table 6.22, which clearly demonstrates the superiority of IFBR. IFBR not only reduces the size but also increases the performance at the same time.

6.3.0.3 Comparison with Existing Approaches

A comparative analysis of IFBR with the HSBR, AntAngioCOOL proposed by Zahiri et. al. [169], AntiAngioPred by Ramaprasad et. al. [123] and TargetAntiAngio by Laengsri et. al. [87] is performed (table 6.23 and visualized in figure 6.10). Zahiri et. al. used the anti-angiogenic dataset for predicting the classification performance employing their proposed methodology. The whole dataset was divided in the ratio of 80:20, the model was trained on training dataset and evaluated on testing (or independent) dataset. While Ramaprasad et. al. and Laengsri et. al. employed the whole dataset to apply their model and thereby evaluate the performance parameters. An accuracy of 68.9%, 75%, 74.8%, 77.5%, sensitivity of 74.8%, 82%, 75.7%, 84.7% and specificity of 63.0%, 71%, 73.8%, 69.4% was obtained by HSBR, Zahiri et. al., Ramaprasad et. al. and Leangsri et. al. respectively. Though a slight decrease in specificity is reported by IFBR on comparing with methodology of Zahiri et. al. and Ramaprasad et. al., an increase in sensitivity and overall accuracy belittle its

TABLE 6.23: Comparison of IFBR on Anti-angiogenic dataset with HSBR, AntAngioCOOL [169], AntiAngioPred [123] and TargetAntiAngio [87]

Method	Sensitivity	Specificity	Accuracy
IFBR	84.8	70.8	78.2
HSBR	74.8	63.0	68.9
AntAngioCOOL	82.0	71.0	75.0
AntiAngioPred	75.7	73.8	74.8
TargetAntiAngio	84.7	69.4	77.5

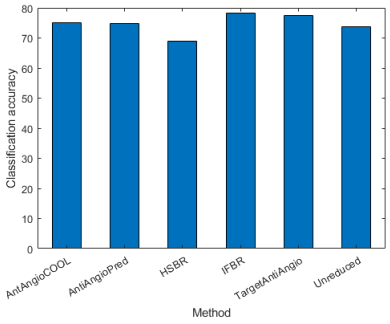


FIGURE 6.10: Graphical visualization showing comparison of the classification accuracy with Bireduct approach

effect. IFBR is thus clearly enhancing prediction performance and is outperforming the existing works.

6.4 Summary

The proposed work has employed intuitionistic fuzzy rough set for generating intuitionistic fuzzy bireducts. Bireducts generation is an effective data reduction process that reduces the data, both in terms of number of features and number of instances. It clearly demonstrates the increment in prediction performances whilst reducing the data and hence the complexity. To quantify the intuitionistic fuzzy bireducts to cover a specific percentage of dataset, ϵ-bireducts is introduced. It has helped in further enhancing the performance by allowing a balance between feature reduction and instance elimination. Particle swarm heuristic search technique is employed for

intuitionistic fuzzy bireduct generation to achieve optimal results. The proposed model of intuitionistic fuzzy bireducts generation has been applied for enhancing the prediction of anti-angiogenic peptides, which is leading therapeutic peptide for cancer treatment. IFBR has increased the prediction accuracy of peptides to 78.2%, thereby outperforming previous works. The effectiveness of IFBR is demonstrated on various benchmark datasets and by comparative analysis with existing approaches. All the works discussed so far are based on supervised datasets. Feature selection based on unsupervised domain will be discussed in the next chapter.

****