Sādhanā (2020) 45:242
https://doi.org/10.1007/s12046-020-01475-2

A strong intuitionistic fuzzy feature association map-based feature selection technique for high-dimensional data

AMIT KUMAR DAS[1], SAPTARSI GOSWAMI[2], AMLAN CHAKRABARTI[1] and BASABI CHAKRABORTI[3,*]

[1]A. K. Choudhury School of Information Technology, University of Calcutta, Kolkata, India
[2]Bangabasi Morning College, University of Calcutta, Kolkata, India
[3]Iwate Prefectural University, Takizawa, Japan
e-mail: amitkrdas.kol@gmail.com; saptarsi007@gmail.com; acakcs@caluniv.ac.in; basabi@iwate-pu.ac.jp

MS received 27 February 2020; revised 19 July 2020; accepted 25 July 2020

Abstract. In this work, a graph-based approach has been adopted for feature selection in case of high-dimensional data. Feature selection intends to identify an optimal feature subset to solve the given learning problem. In an optimal feature subset, only relevant features are selected as "members" and features that have redundancy are considered as "non-members". This concept of "membership" and "non-membership" of a feature to an optimal feature subset has been represented by a strong intuitionistic fuzzy graph. The algorithm proposed in this work at first maps the feature set of the data as the vertex set of a strong intuitionistic fuzzy graph. Then the association between features represented as an edge-set is decided by the degree of hesitation between the features. Based on the feature association, the Strong Intuitionistic Fuzzy Feature Association Map (SIFFAM) is developed for the datasets. Then a sub-graph of SIFFAM is derived to identify features with maximal non-redundancy and relevance. Finally, the SIFFAM based feature selection algorithm is applied on very high dimensional datasets having features of the order of thousand. Empirically, the proposed approach SIFFAM based feature selection algorithm is found to be competitive with several benchmark feature selection algorithms in the context of high-dimensional data.

Keywords. Feature selection; strong intuitionistic fuzzy graph; mutual information; high-dimensional datasets.

1. Introduction

Digital world is moving towards a new paradigm. In place of the traditional business-driven decision-making, it is heading towards data-driven decision-making. Machine learning is extensively used to generate knowledge from the past data. In machine learning, based on the input dataset, the learning models are trained. Couple of decades back, the dimension of the input datasets used to be less than 100 for most of the application domains. In late 1990s, very few domains had more than 40 features [1]. However, over the last two decades, the data storage capacity is getting multiplied over the years. So the volume as well as dimension of input data fed into the machine learning models is also getting multiplied. From that time when machine learning models had to deal with less than 100 features, we have moved to a time when models are dealing with very high-dimensional datasets, especially the ones like microarray [2–4], image [5, 6] and text [7–9] have thousands of features. This makes the task of learning extremely time-consuming and resource-intensive. Also, the models generated are very complicated [10]. This is where a need is felt to step down the dataset dimensions by selecting a subset of features which are absolutely critical in context of the machine learning task.

The objective of feature subset selection is to select a subset of all features, which plays the most critical role with respect to the learning task [11]. An ideal way to identify the most optimal subset of features is to do an exhaustive search in the entire feature space - thus finding out all possible candidate subsets, evaluating each one of them and then selecting the best one. This approach, though being the ideal approach, may not be practically feasible when the number of dimensions of an input dataset is moderately high. For higher dimensional datasets, an approximate search strategy needs to be adopted.

There are different approximate search strategies which have been formulated by different researchers in the past. Graph is a great tool to model the combinatorial relationship between entities. Relationship between features in an

*For correspondence
Published online: 23 September 2020

input dataset is also combinatorial in nature - hence an ideal candidate for being modelled as a graph [12, 13]. This not only helps in the visualization of feature-to-feature association, but also triggers important thoughts regarding how to select a subset which is nearly optimal.

An optimal feature subset should consist of features which are relevant to the learning task and also not having redundancy i.e. similarity with one or more features in the selected subset. But a challenge to identify such an optimal subset is to prioritize amongst relevance and redundancy. Ideally the combined effect should be considered. But how can seamless integration between relevance and redundancy be achieved?

In this paper, an algorithm has been proposed which first models the feature set as vertex set of a strong intuitionistic fuzzy graph [14]. Membership and non-membership of vertices are calculated based on feature-to-class relevance and inter-feature similarity respectively. Then principles of strong intuitionistic fuzzy graph are applied to derive the membership and non-membership values of the edge set. A degree of hesitation is calculated to measure the strength of inter-feature association. Edges are drawn between vertices which have a low degree of hesitation. At the end, subset of connected vertices are picked using sub-graph derivation techniques - maximal independent set and minimal vertex cover. The sub-graph vertices along with the unconnected vertices form the final feature subset.

The remaining part of the paper is organized as follows:

- Section 2 lays down the background concepts along with details of the proposed method.
- Section 3 sets the context for the experiments by providing brief details on datasets used, the benchmark algorithm with which the proposed SIFFAMFS is compared with and the criteria for comparison.
- Intermediate visualizations of the feature sets of each dataset are presented along with the detailed experimental results in section 4.
- At the end, the work and related outcome is summarized in section 5.

2. Proposed method

Before getting into details of the proposed approach and the feature selection algorithm, let's first try to understand some background concepts related to strong intuitionistic fuzzy graph.

2.1 *Background concepts*

The notion of intuitionistic fuzzy sets as coined by Atanassov [15–18] is a generalization of the concept of fuzzy sets [19]. In fuzzy sets, the concept of membership of an element to a given set exists. Obviously, degree of non-membership of the element to the set = 1 - degree of membership. However, Atanassov defined a new component in case of intuitionistic fuzzy sets - the degree of non-membership, which is independent of the degree of membership. The sum of the degrees of membership and non-membership is however less than or equal to 1.

The concept of fuzzy graph was introduced by Rosenfield [20]. A fuzzy graph is represented as $\widetilde{G} = (X, \widetilde{U})$, where X is the vertex set and $\widetilde{U} = \{\mu_U(x_i, x_j) \in X^2\}$ is a fuzzy set of edges with membership function $\mu_U : X^2 \rightarrow [0, 1]$ [21]. A membership function of a fuzzy graph maps each pair of vertices or edges to a membership grade or value between 0 and 1.

Atanassov [22–25] introduced the concept of intuitionistic fuzzy graphs. A strong intuitionistic fuzzy graph [14] is represented as G = (V, E), where V is the intuitionistic fuzzy vertex set such that $\mu_V : V \rightarrow [0, 1]$ denotes the degree of membership and $\gamma_V : V \rightarrow [0, 1]$ denotes the degree of non-membership of the element $v_i \in V$. Also,

$$0 \leq \mu_V(v_i) + \gamma_V(v_i) \leq 1 \tag{1}$$

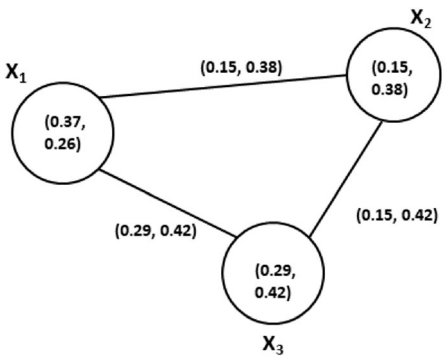
The amount $\pi_v(v_i) = 1 - (\mu_V(v_i) + \gamma_V(v_i))$ is known as degree of hesitation. It may cater to membership value or non-membership value or both [26]. Intuitionistic fuzzy edge set $E \subset V \times V$ [14], where $\mu_E : V \times V \rightarrow [0, 1]$ and $\gamma_E : V \times V \rightarrow [0, 1]$ are such that:

$$\mu_E(v_i, v_j) = min(\mu_V(v_i), \mu_V(v_j)) \tag{2}$$

$$\gamma_E(v_i, v_j) = max(\gamma_V(v_i), \gamma_V(v_j)) \tag{3}$$

$$0 \leq \mu_E(v_i, v_j) + \gamma_E(v_i, v_j) \leq 1 \tag{4}$$

The edge-set hesitation degree is defined by equation 5.



	X_1	X_2	X_3
μ_V	0.37	0.15	0.29
γ_V	0.26	0.38	0.42

	X_1X_2	X_2X_3	X_3X_1
μ_E	0.15	0.15	0.29
γ_E	0.38	0.42	0.42

Figure 1. Sample strong intuitionistic fuzzy graph.

$$\pi_E(v_i, v_j) = 1 - (\mu_E(v_i, v_j) + \gamma_E(v_i, v_j)) \quad (5)$$

In Figure 1, a strong intuitionistic fuzzy graph has been depicted.

For feature selection, we try to derive a near-optimal subset of features from the input dataset, which works best in solving a learning problem. In an optimal feature subset, only relevant features are selected as "members". Also, features which have redundancy are considered as "non-members" of the optimal feature subset. In other words, criteria for selection or "membership" of a feature in an optimal feature subset is decided by relevance. In the same way, criteria for rejection or "non-membership" of a feature in an optimal subset is decided by redundancy. This philosophy has been used to formulate the Strong Intuitionistic Fuzzy FAM based feature selection (SIFFAMFS). The steps of the SIFFAMFS algorithm have been detailed below.

2.2 *Steps for SIFFAMFS algorithm*

Step 1: Computation of vertex fuzzy membership values The feature-to-class normalized mutual information (NMI) is computed based on the mutual information between the features $I(F_i; C)$ normalized by the minimum of entropy of the features F_i i.e. $H(F_i)$ and class C i.e. H(C), as depicted in equation 6.

$$MI_{norm}(F_i,\ C) = \frac{I(F_i; C)}{min(H(F_i), H(C))} \quad (6)$$

The fuzzy membership value μ_V is calculated using Sigmoidal membership function in the interval $[b_1,\ b_3]$ on normalized mutual information MI_{norm} as shown in equation 7. The parameter $b_2 = \frac{b_1+b_3}{2}$ is termed as the crossover point, where $S(b_2; b_1, b_2, b_3) = 0.5$ [21].

$$\mu_V(F_i) = S(MI_{norm}(F_i,\ C), b_1, b_2, b_3) \quad (7)$$

The features having $\mu_V(F_i) = 0$ are to be marked as "red" as they have low relevance and other features are marked as "green" as depicted in figure 2.

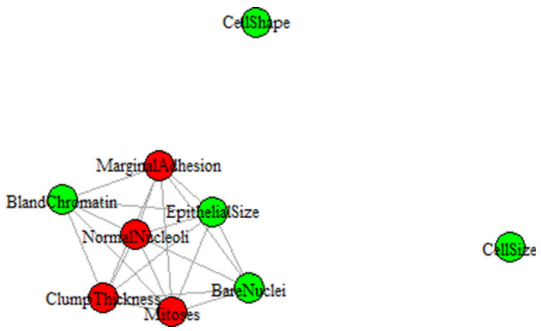


Figure 2. Stage 1 of SIFFAM generation.

Step 2: Computation of vertex fuzzy non-membership values In supervised learning, similarity between two features can be measured by the amount of similar information that the features contribute to decide class value. Amount of information contribution of a feature to class can be measured by the mutual information between the feature and the class. Hence, for calculating the feature-to-feature similarity, the difference between feature-to-class mutual information of the features can be considered, as depicted in Figure 3. In the figure, the feature-to-feature similarity between feature F_1 and F_2 for example, is calculated by the difference of the feature-to-class mutual information $MI(F_1; C)$ and $MI(F_2; C)$ respectively. At an overall level, to formulate the FAM, the average feature level (or vertex level) similarity for a vertex F_1 for example can be measured using the formula:

$$Sim(F_1) = \frac{1}{2}[\{MI(F_1; C) - MI(F_2; C)\} + \{MI(F_1; C) - MI(F_3; C)\}]$$

Generalizing the above formula for any feature F_i and assuming that there are 'n' features (excluding the target variable):

$$\begin{aligned} Sim(F_i) &= \frac{1}{n-1}[\{MI(F_i;C) - MI(F_1;C)\} \\ &\quad + \{MI(F_i;C) - MI(F_2;C)\} \\ &\quad + \ldots + \{MI(F_i;C) - MI(F_n;C)\}] \\ &= \frac{1}{n-1}\sum_{i=1}^{n-1} MI(F_i;C) - \sum_{\substack{j=1 \\ j\neq i}}^{n-1} MI(F_j;C) \\ &= MI(F_i;C) - \frac{1}{n-1}\sum_{\substack{j=1 \\ j\neq i}}^{n-1} MI(F_j;C) \\ &= MI(F_i;C) - \frac{1}{n-1}\{n * mean(MI_{dataset}) - MI(F_i;C)\} \\ &= MI(F_i;C) + \frac{1}{n-1}MI(F_i;C) - \frac{n}{n-1} * mean(MI_{dataset}) \\ &= \frac{n}{n-1}[MI(F_i;C) - mean(MI_{dataset})] \end{aligned}$$

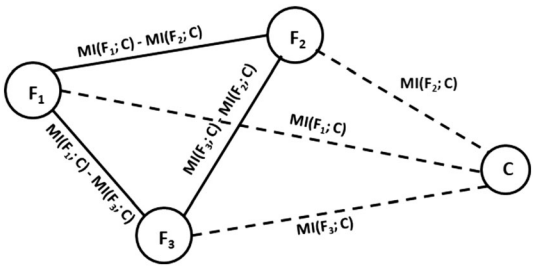


Figure 3. Vertex similarity.

Once the average similarity values for the features are calculated, the fuzzy non-membership value γ_V is computed applying Sigmoidal membership function in the interval $[b_1, b_3]$ on the similarity values using equation 8.

$$\gamma_V(F_i) = S(Sim(F_i), b_1, b_2, b_3) \tag{8}$$

The parameter $b_2 = \frac{b_1+b_3}{2}$ is the crossover point.

Step 3: Computation of edge-set fuzzy membership and non-membership values For the intuitionistic fuzzy edge set, the fuzzy membership matrix μ_E and fuzzy non-membership matrix γ_E should be derived using equations 9 and 10 [14].

$$\mu_E(F_i, F_j) = min(\mu_V(F_i), \mu_V(F_j)) \tag{9}$$

$$\gamma_E(F_i, F_j) = max(\gamma_V(F_i), \gamma_V(F_j)) \tag{10}$$

Step 4: Derive the fuzzy hesitation matrix for the edge set For the intuitionistic fuzzy edge set, fuzzy hesitation matrix π_E should be computed using equation 11.

$$\pi_E(F_i, F_j) = 1 - (\mu_E(F_i, F_j) + \gamma_E(F_i, F_j)) \tag{11}$$

Smaller π_E resembles higher value of $\mu_E(F_i, F_j)$ and $\gamma_E(F_i, F_j)$. This, according to equations 9 and 10, signify higher value of membership due to relevance and / or higher value of non-membership due to redundancy of the features. Hence, smaller value of π_E would mean less hesitation and better feature association. If the degree of hesitation is less than a threshold value, say θ, i.e. if $\pi_E(F_i, F_j) \leq \pi_\theta$, then make the value of the cell in adjacency matrix corresponding to SIFFAM as 1, else make it 0.

Out of the "green" features, the ones having adjacency with one or more features i.e. having sum of the corresponding row in the adjacency matrix of SIFFAM greater than 0, are marked as "blue" as depicted in figure 4.

Step 5: Generate final Strong Intuitionistic Fuzzy FAM (SIFFAM) The subgraphs of the vertex set marked in "blue" are derived using the concepts of minimal vertex cover and maximal independent set. Out of the vertices marked "blue", (as depicted in figure 4), the ones being part of the subgraph selected are marked "green" while the others are marked "red". This is the final stage of SIFFAM as shown in figure 5. At the end, all "green" features are selected as the final feature subset.

The proposed algorithm, named as SIFFAMFS, has been outlined as algorithm 1.

Algorithm 1 Strong Intuitionistic Fuzzy FAM based Feature Selection (SIFFAMFS)

Input: Dataset D_N with N-dimensions i.e. having feature set F (where F = $\{F_1, F_2, F_3, ., F_N\}$)
Output: Optimal feature subset, SS_{opt}
Notations used: b_1, b_2 and b_3 are thresholds of Sigmoidal membership function
M_{SIFFAM} - Strong intuitionistic fuzzy FAM adjacency matrix
F_i, F_j - i-th and j-th feature
f_r, f_b, f_g - features marked "red", "blue" and "green" respectively

Begin
for each $F_i \in F$, **do**
  $MI_{NORM}(F_i, C) = \frac{MI(F_i,C)}{min(H(F_i),H(C))}$
  $\mu_V(F_i) = S(MI_{NORM}, b_1, b_2, b_3)$
end for
if $(\mu_V(F_i) == 0)$ **then**
  Add F_i to f_r
else
  Add F_i to f_g
end if
for each $F_i \in F$, **do**
  $Sim(F_i) = \frac{(n+1)}{(n-1)} * MI(F_i, C) - \frac{n}{(n-1)} * mean(MI(D_n))$
  $\gamma_V(F_i) = S(Sim(F_i), b_1, b_2, b_3)$
end for
for each $F_i \in F$, **do**
  for each $F_j \in F$, **do**
    $\mu_E(F_i, F_j) = min(\mu_V(F_i), \mu_V(F_j))$
    $\gamma_E(F_i, F_j) = max(\gamma_V(F_i), \gamma_V(F_j))$
  end for
end for
$\pi_E(F_i, F_j) = 1 - (\mu_E(F_i, F_j) + \gamma_E(F_i, F_j))$
for each $F_i \in f_g$, **do**
  if $(rowsum(\pi_E(F_i)) == 0)$ **then**
    Add F_i to f_b
    Remove F_i from f_g
  end if
end for
$G_{\widetilde{SIFFFAM1}} = GenerateGraph(M_{FFAM}, f_r, f_g, f_b)$
Select a subset of f_b as a MVC or MIS of f_b. Add the selected subset to f_g and remaining to f_r.
$G_{\widetilde{SIFFFAM2}} = UpdateGraph(G_{\widetilde{SIFFFAM1}}, f_r, f_g, f_b)$
Output features f_g as an optimal feature subset SS_{OPT}.
End

3. Methods and materials

3.1 *Datasets used*

The SIFFAMFS algorithm developed is applied on gene microarray datasets taken from open-source feature

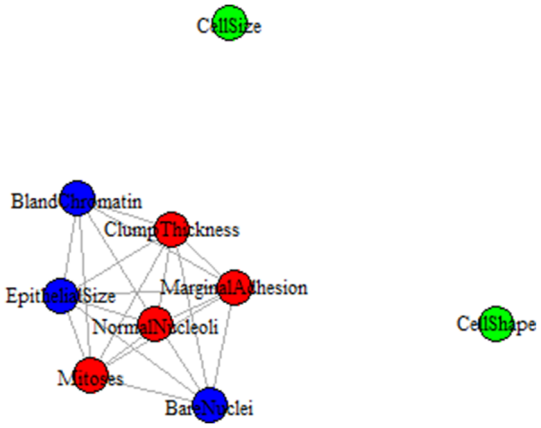


Figure 4. Stage 2 of SIFFAM generation.

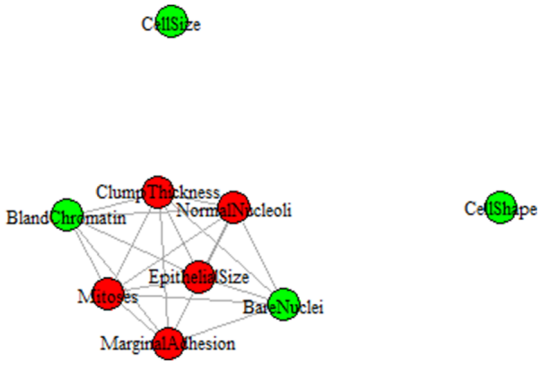


Figure 5. Final stage of SIFFAM generation.

selection repository developed at Arizona State University [27]. All the datasets have more than 1000 features - the dataset with lowest dimensions having 2,000 features and the highest one having close to 50,000 features. More details about the datasets used in the experiments have been presented in the table 1.

Table 1. High-dimensional datasets [27] used in experiments.

Dataset	Number of Features	Number of Records	Number of Classes
colon	2,000	62	2
lymphoma	4,027	96	9
TOX-171	5,749	171	4
ALLAML	7,130	72	2
CLL-SUB-111	11,341	111	3
SMK-CAN-187	16,383	187	2
GLA-BRA-180	49,152	180	4

Since the datasets are very high-dimensional, an initial relevance-based filtering have been applied, selecting top ζ relevant features. The value of ζ has been heuristically set to a value within $\sqrt{n}$ and $\sqrt{n} * \log(n)$, where n is the number of features of the dataset. This has helped in reducing the time complexity from $O(n^2)$ to $O(\zeta^2)$, where $\sqrt{n} \leq \zeta \leq \sqrt{n} * \log(n)$.

3.2 *Competing algorithms*

For performance comparison, the following algorithms have been selected:

- The Relief-F algorithm [28], a well-known feature selection algorithm based on univariate technique or evaluating each feature separately based on its discriminating ability.
- A well-known multivariate, sequential forward selection based algorithm named FOCUS-SF [29].
- A graph-based algorithm named Fast clustering based feature selection algorithm (FAST) [30] which applies graph-theoretic clustering methods to features and has demonstrated superior performance on high-dimensional datasets.

The results of the SIFFAMFS algorithm have been compared with the results of the three benchmark algorithms i.e. Relief-F, FOCUS-SF and FAST, taken from [30]. Comparison is done based on two aspects - classification accuracy and execution time. Accuracy value based on 100 iterations, along with the corresponding execution time, have been reported.

3.3 *Evaluation criteria*

The experiments have been done to evaluate the performance of the proposed algorithm compared to the competing algorithms with respect to two main criteria - classification accuracy and time of execution as these are the measures for the effectiveness and efficiency of any algorithm.

For accuracy, the measurement has been done as follows: Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$, where TP, TN, FP and FN represent the number of true positives, true negatives, false positives and false negatives respectively.

For measuring the classification accuracy with the derived feature subset from each algorithm, Naive Bayes classifier - has been used. Naive Bayes is a basic classification model which can be used to test the efficacy of any feature selection algorithm.

In addition to accuracy and time taken for execution, the extent of feature reduction is also considered to evaluate the efficacy of the proposed algorithm. Extent of feature reduction, measured as feature reduction percentage, is calculated as follows:

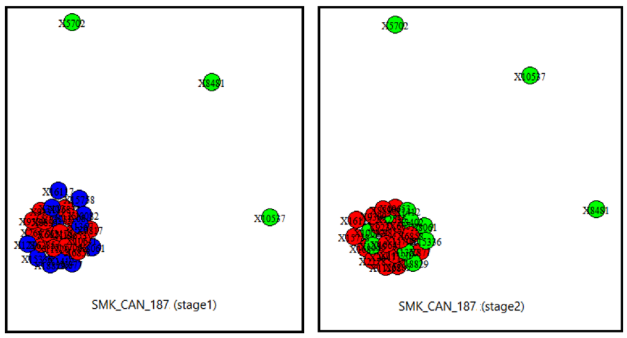


Figure 6. SIFFAM formation for SMK-CAN-187 dataset.

$$\text{Feature reduction} = (\text{Number of features rejected} / \text{Total number of features}) \times 100$$

4. Results and analysis

4.1 *Dataset SIFFAM generated*

As a part of the SIFFAMFS algorithm, a stage-wise visual representation of the features of the dataset in the form of Strong Intuitionistic Feature Information Map (SIFFAM) is first generated. Figure 6 gives the stages of SIFFAM formation for SMK-CAN-187 dataset. This view can give significant information related to potentially irrelevant and redundant features in datasets.

4.2 *Experimental results*

In table 2, 3 and 4, the experimental results for SIFFAMFS algorithm compared to the benchmark algorithms have been presented. The relative comparison of classification accuracy values yielded by the different algorithms can be observed in table 2. The accuracy values given SIFFAMFS and FAST algorithms are best for most of the datasets. None of the other algorithms have accuracy close to these algorithms. The F1-score for SIFFAMFS algorithm have also been presented in table 2 for more rigorous evaluation of the classifier performance with the features extracted by the proposed SIFFAMFS algorithm. The F1-score values are however same or almost same to the classification accuracy for all the datasets. This firmly shows the goodness of classification using the extracted features.

The relative comparison of time of execution needed by the different algorithms can be observed in the table 3:

- The efficiency of SIFFAMFS algorithm is by far the best for all datasets.

Table 2. Classification accuracy - SIFFAMFS vs. other algorithms.

		Classification Accuracy				F1-score
Dataset [27]	Dimension	ReliefF	FOCUS-SF	FAST	SIFFAMFS	SIFFAMFS
colon	2000	0.68	0.85	**0.95**	0.94	0.94
lymphoma	4,027	0.79	0.82	0.98	**1.0**	1.0
TOX-171	5,749	0.8	0.74	**0.86**	0.84	0.84
ALLAML	7,130	0.85	0.77	**1.0**	**1.0**	1.0
CLL-SUB-111	11,341	0.69	0.74	0.85	**0.91**	0.91
SMK-CAN-187	16,383	0.68	0.74	**0.86**	0.8	0.8
GLA-BRA-180	49,152	0.69	0.63	0.76	**0.83**	0.82

Table 3. Execution time - SIFFAMFS vs. other algorithms.

		Execution time (secs.)			
Dataset [27]	Dimension	ReliefF	FOCUS-SF	FAST	SIFFAMFS
colon	2000	7.44	9.6	**1.66**	2.5
lymphoma	4,027	70	23.48	6.35	**3.97**
TOX-171	5,749	97.57	84.46	17.5	**5.52**
ALLAML	7,130	18.84	47.34	16.3	**7.04**
CLL-SUB-111	11,341	127.2	123.07	16.87	**10.54**
SMK-CAN-187	16,383	348.81	386.06	43.07	**16.76**
GLA-BRA-180	49,152	976.41	1525.36	298.54	**41.93**

Table 4. Feature reduction - SIFFAMFS vs. other algorithms.

		Proportion of selected features (in %)			
Dataset [27]	Dimension	ReliefF	FOCUS-SF	FAST	SIFFAMFS
colon	2000	39.13	**0.30**	**0.30**	3.55
lymphoma	4,027	98.24	**0.12**	2.06	1.71
TOX-171	5,749	64.60	**0.19**	0.28	0.94
ALLAML	7,130	50.50	0.06	**0.04**	0.72
CLL-SUB-111	11,341	54.35	0.08	**0.04**	0.71
SMK-CAN-187	16,383	14.23	**0.06**	0.13	0.45
GLA-BRA-180	49,152	53.06	**0.02**	0.03	0.15

Table 5. Summary of experiments.

	ReliefF	FOCUS-SF	FAST	SIFFAMFS
Average accuracy	0.74	0.76	0.89	**0.91**
Average execution time (secs.)	235.18	314.2	57.18	**12.61**
Average reduction (%)	46.56	**99.88**	99.59	98.83

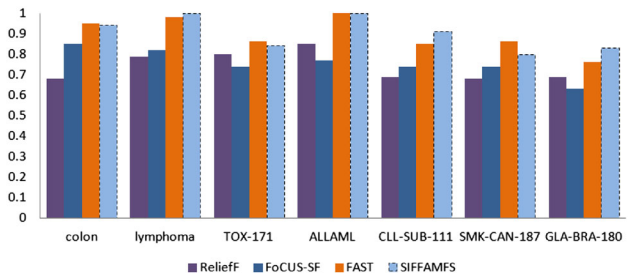


Figure 7. Comparison of accuracy for all high-dimensional datasets.

- FAST algorithm, which gives better accuracy than SIFFAMFS for a couple of datasets, has a higher average execution time.

The comparative values of feature reduction have been given in table 4. This is the only area where the traditional feature selection algorithm FOCUS-SF and the competitive graph-based feature selection algorithm FAST have performed better than our proposed SIFFAMFS algorithm. For all datasets, the FOCUS-SF and FAST algorithms have given a higher feature reduction and hence selected a lower proportion of features. The average proportion of features selected by SIFFAMFS algorithm is 1.17% compared to 0.12% and 0.41% selected by FOCUS-SF and FAST respectively. However, even 1.17% features selected by SIFFAMFS i.e. 98.83% reduction given by SIFFAMFS is extremely high amount of reduction. It can be tuned further by adjusting the value of ζ used at the beginning of the algorithm.

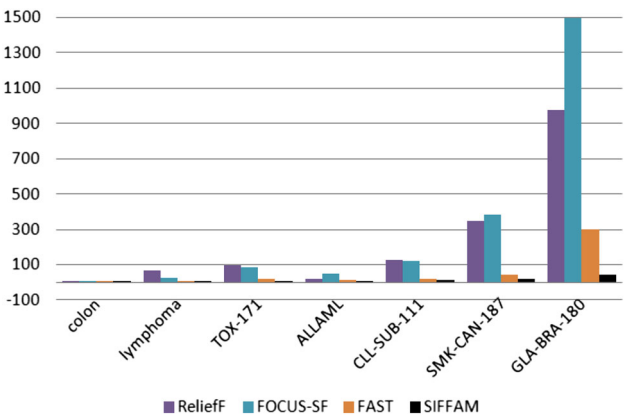


Figure 8. Comparison of execution time (in secs.) for all high-dimensional datasets.

Figures 5 and 6 give a visual comparison of the FAM-based algorithms with the benchmark algorithms.

- SIFFAMFS accuracy curve gives higher peaks than most of the other algorithms.
- Barring the GLA-BRA-180 dataset, FAST gives a decent accuracy value for the other datasets.
- For datasets of very high number of dimensions (e.g. SMK-CAN-187, GLA-BRA-180), traditional feature selection algorithms like ReliefF and FOCUS-SF end up with a very high execution time. Graph-based algorithms like FAST or SIFFAMFS algorithms have given a much better performance in terms of execution time.

- With the adopted threshold ζ, SIFFAMFS has given consistently fast performance, much better than the other algorithms.

All the experimental results have been summarized in table 5, based on which following inferences can be drawn:

- For gene microarray data, SIFFAMFS algorithm is undoubtedly a very good choice for feature selection. It has outperformed all benchmark algorithms, both in terms of efficiency (demonstrated by fast execution) as well as effectiveness (high average accuracy value and F1-score yielded).
- Traditional feature selection algorithms like ReliefF and FOCUS-SF do not yield favourable results for high-dimensional data. Other than low accuracy, the time of execution is also very high in case of traditional feature selection algorithms.
- Amongst other graph-based algorithms, FAST is a good choice for feature selection in high-dimensional datasets. However, execution time of FAST is much higher than the SIFFAMFS algorithm.

5. Conclusion

In this paper, the strong intuitionistic fuzzy FAM framework has been used to integrate feature relevance and redundancy seamlessly rather than considering them separately as two different parameters for feature selection. Using the concept of Strong Intuitionistic fuzzy FAM (or SIFFAM), feature relevance (as membership value) and feature redundancy (as non-membership value) have been considered together to formulate association between features. This has helped to consider feature association in a holistic way. This has given a significant gain in efficiency, demonstrated in the form of much reduced time of execution.

Results of the experiments conducted using publicly available gene microarray datasets are quite promising. The proposed algorithm demonstrates better results, with respect to both efficiency and effectiveness, compared to the competing algorithms. The results keep on improving as the dataset dimension increases justifying its suitability to be adopted as feature selection algorithm for high-dimensional datasets such as gene microarray.

References

[1] Guyon I and Elisseeff A 2003 An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3:1157–1182
[2] Swarnkar T and Mitra P 2015 Graph-based unsupervised feature selection and multiview clustering for microarray data. *J. Biosci.* 40:755–767
[3] Shukla A K, Singh P, and Vardhan M 2018 A hybrid gene selection method for microarray recognition
[4] Yu L and Liu H 2004 Redundancy based feature selection for microarray data. In *KDD '04*
[5] Gal Y, Islam R, and Ghahramani Z 2017 Deep bayesian active learning with image data. ArXiv abs/1703.02910
[6] Ganesan J and Inbarani H H 2016 Hybrid tolerance rough set-firefly based supervised feature selection for mri brain tumor image classification. *Appl. Soft Comput.* 46: 639–651
[7] Leopold E and Kindermann J 2002 Text categorization with support vector machines. How to represent texts in input space? *Mach. Learn.* 46: 423–444
[8] Nigam K, McCallum A, Thrun S, and Mitchell T M 2000 Text classification from labeled and unlabeled documents using em. *Mach. Learn.* 39: 103–134
[9] Feng G, Guo J, Jing B Y, and Sun T 2015 Feature subset selection using naive bayes for text classification. *Pattern Recogn. Lett.* 65: 109–115
[10] Bolón-Canedo V, Sánchez-Maroño N, and Alonso-Betanzos A 2015 Feature selection for high-dimensional data. In: *Artificial Intelligence: Foundations, Theory, and Algorithms*
[11] Chandrashekar G and Sahin F 2014 A survey on feature selection methods. *Comput. Electric. Eng.* 40: 16–28
[12] Bandyopadhyay S, Bhadra T, Mitra P, and Maulik U 2014 Integration of dense subgraph finding with feature clustering for unsupervised feature selection. *Pattern Recogn. Lett.* 40: 104–112
[13] Das A K, Kumar S, Jain S, Goswami S, Chakrabarti A, and Chakraborty B 2019 An information-theoretic graph-based approach for feature selection. *Sādhanā* 45(1): 11
[14] Akram M and Davvaz B 2012 Strong intuitionistic fuzzy graphs
[15] Atanassov K T 1986 Intuitionistic fuzzy sets. *Fuzzy Sets Syst.* 20: 87–96
[16] Atanassov K T 1986 More on intuitionistic fuzzy sets. *Fuzzy Sets Syst.* 33: 37–46
[17] Atanassov K T 1999 Intuitionistic fuzzy sets. Physica-Verlag, Heidelberg
[18] Atanassov K T 2012 On intuitionistic fuzzy sets theory. *Stud. Fuzziness Soft Comput.*
[19] Zadeh L A 1965 Fuzzy sets. *Inf. Control* 8: 338–353
[20] Rosenfeld A 1975 Fuzzy graphs, fuzzy sets and their applications. Academic Press, New York, pp. 77–95
[21] Pal S K and Chakraborty B 1986 Fuzzy set theoretic measure for automatic feature evaluation. *IEEE Trans. Syst. Man Cybern.* 16: 754–760
[22] Atanassov K T and Shannon A 1994 A first step to a theory of the intuitionistic fuzzy graphs. In: *Proceeding of FUBEST (D. Lakov, Ed.) Sofia*, pp. 59–61
[23] Atanassov K T, Pasi G, Yager R, and Atanassova V 2003 Intuitionistic fuzzy graph interpretations of multi-person multi-criteria decision making. In: *EUSFLAT Conference*, pp. 177–182
[24] Karunambigai M G and Parvathi R 2006 Intuitionistic fuzzy graphs. *J. Comput. Intell. Theory Appl.* 139–150
[25] Atanassov K T, Parvathi R and Karunambigai M G 2009 Operations on intuitionistic fuzzy graphs, fuzzy systems. In: *IEEE International Conference, FUZZ-IEEE 2009*, pp. 1396–1401
[26] De S K, Biswas R and Roy A R 2001 An application of intuitionistic fuzzy sets in medical diagnosis. *Fuzzy Sets Syst.* 117: 209–213

[27] Li J, Cheng K, Wang S, Morstatter F, Trevino R P, Tang J and Liu H 2017 Feature selection: a data perspective. *ACM Comput. Surv.* 50: 94:1–94:45

[28] Kononenko I 1994 Estimating attributes: analysis and extension of relief. In: *Proceedings of the Sixth European Conference on Machine Learning*, pp. 171–182

[29] Dietterich T G, Almuallim H 1994 Learning boolean concepts in the presence of many irrelevant features. *Artif. Intell.* 69(1–2):279–305

[30] Song Q, Ni J and Wang G 2013 A fast clustering-based feature subset selection algorithm for high-dimensional data. *IEEE Trans. Knowl. Data Eng.* 25: 1–14