Sādhanā (2020)45:11
https://doi.org/10.1007/s12046-019-1238-2

An information-theoretic graph-based approach for feature selection

AMIT KUMAR DAS[1], SAHIL KUMAR[1], SAMYAK JAIN[1], SAPTARSI GOSWAMI[2], AMLAN CHAKRABARTI[2] and BASABI CHAKRABORTY[3,*]

[1]Department of Computer Science and Engineering, Institute of Engineering and Management, Kolkata, India
[2]A K Choudhury School of Information Technology, University of Calcutta, Kolkata, India
[3]Iwate Prefectural University, Takizawa, Japan
e-mail: amitkrdas.kol@gmail.com; sahilriders@gmail.com; samj09091995@gmail.com; saptarsi007@gmail.com; achakra12@yahoo.com; basabi@iwate-pu.ac.jp

MS received 1 January 2019; revised 18 April 2019; accepted 17 October 2019

Abstract. Feature selection is a critical research problem in data science. The need for feature selection has become more critical with the advent of high-dimensional data sets especially related to text, image and micro-array data. In this paper, a graph-theoretic approach with step-by-step visualization is proposed in the context of supervised feature selection. Mutual information criterion is used to evaluate the relevance of the features with respect to the class. A graph-based representation of the input data set, named as feature information map (FIM) is created, highlighting the vertices representing the less informative features. Amongst the more informative features, the inter-feature similarity is measured to draw edges between features having high similarity. At the end, minimal vertex cover is applied on the connected vertices to identify a subset of features potentially having less similarity among each other. Results of the experiments conducted with standard data sets show that the proposed method gives better results than the competing algorithms for most of the data sets. The proposed algorithm also has a novel contribution of rendering a visualization of features in terms of relevance and redundancy.

Keywords. Graph-based feature selection; supervised learning; mutual information; vertex cover; feature graph.

1. Introduction

Twenty-first century has ushered in an era of data science and data-driven thinking [1]. There is a huge explosion of data available for analysis in every domain [2]. With this, there has been a rapid increase of dimensionality of the data sets used in data-science-related activities and feature selection has become extremely critical [3]. Feature selection is an important pre-processing step in machine learning, which helps in selecting a subset of features from the entire feature set. If the entire feature set is used, especially for the data sets with higher number of features, the computation cost of executing machine learning tasks will be very high.

Depending on whether the value of target variable is available or not, feature selection may be supervised or unsupervised. For supervised feature selection, relative importance of the features can be determined with respect to target variable. Mutual information (MI) between two random variables measures the amount of information contribution or the reduction in uncertainty that one variable can bring to the other variable. The uncertainty of a random variable is represented by its entropy. Conditional entropy of a random variable Y with respect to another random variable X indicates the quantum of information needed to describe the outcome Y when the value of X is known. MI of a variable X with respect to a variable Y is the difference between the entropy of variable Y and the conditional entropy of variable Y when the value of X is known. It is quite obvious that if X and Y are independent, MI is 0. As the dependence between X and Y increases, the value of MI also increases. Putting this in the context of data set features, the features as well as the class are random variables. MI between the features and the class variable is measured by the difference in entropy of the class variable and the conditional entropy of the class variable given the value of the feature variable. Hence, higher value of MI signifies higher information gain of the class variable from the feature variable [4]. Feature F_1 is preferred to feature F_2 if the information gain from F_1 is greater than that from F_2 [5].

For feature selection, both efficiency, i.e. time required to identify an optimal feature subset, and effectiveness, i.e. accuracy of the machine learning task using the feature subset identified, are equally important considerations.

*For correspondence
Published online: 30 December 2019

Filter-based feature selection methods are known for their superior efficiency. These methods use statistical measures to determine the optimality of the selected subset. Both relevance and redundancy of candidate feature subsets are essential parameters to optimize the performance of the machine learning task [6]. This is because irrelevant features do not help in deciding the target class value in any way [7]. On the other hand, a redundant feature contributes similar information as one or more other features and hence no additional information is contributed by it. Since both relevance and redundancy are essential parameters to decide the importance of a feature, respective statistical measures, like MI for measuring feature relevance and inter-feature correlation for measuring feature redundancy, are used.

Ideally, for feature selection, a complete search strategy should be used for selecting the most optimal feature subset. However, considering the computation cost of such a strategy, an approximate search technique is used in most of the cases. Graph-based approach can be adopted for modelling the relationship between the data set features [8–10]. Graph is an effective medium to model any combinatorial relationship. This graph, also known as feature graph, models features of a data set as the vertices and inter-feature relation as the edges. Later, some technique to identify a sub-graph of the feature graph is applied. The vertices of the sub-graph represent a potential feature subset. This approach has been adopted by a few researchers working on the area of feature selection. However, the extent of relevance or redundancy of the features is not represented in the feature graph. They are critical aspects, which can help the users to take decision on whether to select or reject a specific feature. Graphs give a visualization of the relationship between objects, and help the users in making useful inferences.

In this paper, we propose to represent the features in the form of a graph, highlighting the extent of irrelevance of the individual features. This graph, proposed to be named as feature information map (FIM), will help the user to visualize the key features of the data set from relevance perspective. We also propose to model the inter-feature similarity between the most informative features. This, as a whole, helps in identifying the feature subset having the maximum relevance and least redundancy in the case of supervised classification.

1.1 *Related works*

Supervised feature selection algorithms can be based on filter or wrapper approach. Algorithms based on wrapper approach use the predictive accuracy of the selected learning model for different candidate subsets and select the one that gives better result [5]. Filter algorithms use the general characteristics of the data to select the feature subset [11–13].

In filter approach, feature evaluation can be either univariate or multivariate. In both cases, information-theoretic measures can be used to select subset of features having highest discriminative ability. In case of univariate, criticality of each feature is evaluated and most critical features are chosen. In case of multivariate evaluation, features are considered as a group and not in isolation. MI is an efficient and easily interpretable information-theoretic measure popularly used in filter-based algorithms [14–17]. In the algorithms that have used univariate approach, feature to class MI has been calculated and feature subset has been identified either in a rank-based or threshold-based approach. Algorithms using multivariate approach have chosen features that have boosted the overall information contribution of the feature set to the class.

Minimal-redundancy-maximal-relevance (mRMR) [14] is a popular supervised feature selection algorithm, which uses MI to identify good features according to the maximal statistical dependence.

Often, data set features are modelled in the form of information-theoretic feature graphs. Later, some sub-graph selection mechanism is used to identify feature subsets. One such information-theoretic graph-based supervised feature selection algorithm [18] has been implemented using the concept of dominant-set clustering. In this algorithm, multidimensional interaction information is used as a feature selection criterion. In another similar work [19], hypergraph clustering technique is used to select the most informative feature subset, which has both low redundancy and strong discriminating power.

Structure learning, which can be thought as a more general problem than feature selection, also maps the relationships among features as graph. Generally it assumes that the data is generated using Bayesian Network. Structure learning also uses MI between features, often as an edge-weight of the MI graph [20–22]. Feature selection algorithms using structure learning among features have also been implemented [23]. However, it may be noted that effective and efficient learning of Bayesian structure, especially for high-dimensional data, is a challenging problem [24].

2. Preliminary concepts

2.1 *"More" and "less" informative features*

Features of a data set make different amounts of information contribution about the class variable. MI is an appropriate measure to indicate the level of information contribution of a feature. High value of feature to class MI indicates that the feature has significant contribution in deciding value of the class label. Hence, those features are the more informative features and need to be considered as a part of the final feature subset.

On the other hand, if, for a particular feature, the feature to class MI value is low, the feature is less informative. All less informative features are considered as candidates for rejection when the final feature subset is identified.

2.2 *MI*

MI of a feature F with respect to the class variable C, i.e. $I(F; C)$, is measured by the difference in entropy of the class variable, i.e. $H(C)$, and the conditional entropy of the class variable given the value of the feature variable, i.e. $H(C \mid F)$. Following equations are used to calculate feature-to-class MI:

$$I(F;C) = H(C) - H(C|F) \tag{1}$$

$$I(F;C) = H(C) + H(F) - H(C,F) \tag{2}$$

where $H(C)$ and $H(F)$ are the marginal entropies and $H(C, F)$ is the joint entropy of C and F. $H(C)$, $H(F)$ and $H(C, F)$ are defined by following equations:

$$H(C) = -\sum_{i=1}^{K} p(C_i)\log_2 p(C_i), \tag{3}$$

$$H(F) = -\sum_{i=1}^{n} p(F_i)\log_2 p(F_i), \tag{4}$$

$$H(C,F) = -\sum_{K}\sum_{n} p(C,F)\log_2 p(C,F). \tag{5}$$

In these equations, K is number of classes, C is class variable, F is feature set, which takes discrete values, n is number of particular values of F and $p(C, F)$ is joint probability of C and F.

2.3 *Similar features*

One or more features may have similar information. If two features have similar information, in spite of both being relevant, one of them can be rejected.

Pearson's product moment correlation coefficient, or simply inter-feature correlation, is a widely adopted measure of feature similarity [25]. Features having high correlation value with one or more features are candidates for elimination because of potential redundancy. At the same time, features having very less correlation with other features are assumed to have less redundancy and hence ideal candidates to be a part of the final feature subset.

2.4 *Inter-feature correlation*

Inter-feature correlation α is calculated using the following equation:

$$\alpha = \frac{\mathrm{cov}(F_1, F_2)}{\sqrt{\mathrm{var}(F_1)\mathrm{var}(F_2))}} \tag{6}$$

where the covariance between features F_1 and F_2 i.e.

$$\mathrm{cov}(F_1, F_2) = \sum (F_{1i} - \overline{F_1})(F_{2i} - \overline{F_2}) \tag{7}$$

and the variances of features F_1 and F_2 are calculated using the following equations:

$$\mathrm{var}(F_1) = \sum (F_{1i} - \overline{F_1})^2, \tag{8}$$

$$\mathrm{var}(F_2) = \sum (F_{2i} - \overline{F_2})^2. \tag{9}$$

2.5 *Vertex cover*

In a graph, a subset of the vertices is known as vertex cover if all the edges in the graph are incident on one of the vertices of that subset. There can be more than one vertex cover for a particular graph. The smallest amongst all vertex covers, that is the one having least number of vertices, is known as the minimum vertex cover. If the features in a data set are represented in the form of a feature graph, a vertex cover, or more specifically a minimum vertex cover, represents the most optimal sub-graph of the feature graph. Hence, the features forming the vertices of the minimum vertex cover can be said to represent the optimal feature subset. However, finding minimum vertex cover is considered to be a classical NP-complete problem. Hence, a minimal vertex cover that can be identified in polynomial time can be used as the closest possible approximation of the minimum vertex cover. There are multiple algorithms for deriving minimal vertex cover [26, 27]. One simple approach is to identify vertex covers by node traversal and choosing the one with lowest cardinality.

3. Proposed method

3.1 *Stages of FIM formation*

3.1.1 *Stage 1 – highlight irrelevant (or least informative) features* In the proposed method, MI between the feature variables and the class variable is calculated. Then the FIM is generated with the features represented as the vertices of FIM. Default colour "green" is assigned to all vertices. Features that have MI value lower than that of the average MI value are identified as least informative features. They are coloured "red", indicating that these features are candidates for elimination due to low relevance. This is the stage 1 of FIM.

3.1.2 *Stage 2 – highlight potentially redundant (or similar) features* In the next step, similarity between

features coloured in "green", i.e. features expected to be relevant, are measured by their inter-feature correlation. A threshold value of correlation, α_0, is taken as user input. A correlation value above α_0 indicates high correlation and hence high degree of similarity.

A similarity matrix is then created in a way such that each cell in the matrix holds a value equal to the correlation between the features represented by the respective row and column for that cell. From the similarity matrix, an adjacency matrix is created with value of all cells greater than α_0 replaced by 1 and those that are less than α_0 replaced by 0. The values of the cells that are part of the leading diagonal of the similarity matrix are made 0, as it reflects the autocorrelation of each feature and that is always 1. Using this matrix, the FIM is redrawn. Out of the "green" nodes, the ones that get connected are marked "blue". These nodes have good amount of information contribution as well as potential redundancy. This is the stage 2 of FIM.

3.1.3 *Stage 3 – highlight redundant features* From the connected features marked in "blue", a subset is selected as a representative of the whole set. For subset selection, the concept of minimal vertex cover is applied. Identification of vertex covers has been done as a part of the proposed algorithm by traversing the nodes of the feature graph. From the multiple vertex covers identified, the one with lowest cardinality has been selected. The features corresponding to the minimal vertex cover are marked as "green" as they are going to be a part of the feature subset. The remaining "blue" vertices are marked "red". This is the stage 3 of FIM.

From the stage 3 of FIM, all features corresponding to the vertices marked "green" are selected to be a part of the final feature subset.

3.2 *Algorithm*

Feature information map-based feature selection (FIMFS)

Input
1. Data set D_N with N dimensions, i.e. having $N-1$ feature set F (where $F = \{F_1, F_2, F_3, ., F_{N-1}\}$) and C, the class field.
2. Similarity threshold α_0.

Output
Optimal feature subset SS_{opt}

// Stage 1: Calculate MI of every feature with respect to the class field, find average MI, draw FIM with features having MI less than average MI as "red" vertices.
Step 1: For $i = 1$ to $N-1$
Step 2: $MI_i = MI(F_i, C)$
Step 3: *If* $MI_i <$ average (MI_i) *then*
Step 4: colour $(F_i) \leftarrow$ "red"
Step 5: *End If*
Step 6: Next
Step 7: $g_{v1} \leftarrow$ generate $FIM(F)$
// Stage 2: For features corresponding to green vertices in FIM, calculate pair-wise correlation. Connect the vertices having high pair-wise correlation and mark them "blue".
Step 8: $F' = \{x : x \subseteq F$ and colour(x) = "green" $\}$
Step 9: $M_{corr} \leftarrow$ correlation $(D_N[F'])$
Step 10: For $i = 1$ to $|F'|$
Step 11: *For* $j = 1$ to $|F'|$
Step 12: *If* $M_{corr}[i, j] > \alpha_0$ and $i \neq j$, *then*
Step 13: add-edge (F_i, F_j, g_{v1})
Step 14: colour $(F_i) \leftarrow$ "blue"
Step 15: colour $(F_j) \leftarrow$ "blue"
Step 16: *End If*
Step 17: *Next*
Step 18: Next
// Stage 3: On the blue vertices, apply minimal vertex cover to identify a subset. The subset of vertices which are part of vertex cover are marked green and the remaining marked red.
Step 19: $F'' = \{x : x \subseteq F$ and colour(x) = "blue" $\}$
Step 20: $V_{VERTCOV} \leftarrow$ MVC (F'')
Step 21: colour $(V_{VERTCOV}) \leftarrow$ "green"
Step 22: colour $(F'' - V_{VERTCOV}) \leftarrow$ "red"
Step 23: $SS_{opt} = \{x : x \subseteq F$ and colour(x) = "green" $\}$

Table 1. Details of data sets used.

Data set	Number of features	Number of records	Number of classes
CTG	34	2126	10
Texture	40	5500	11
Sonar	60	208	2
Optdgt	63	5620	10
Digits	257	1593	10
Madelon	500	2000	2
Isolet	618	1559	26
Mfeat	649	2000	10

4. Methods and materials

4.1 *Data sets used*

For conducting the experiments, 8 standard data sets, details provided in table 1, from the machine learning repository of UCI (University of California, Irvine) [28] have been used. Out of the 8 data sets, 4 data sets have number of features less than 100 while the remaining 4 data sets have number of features greater than 100. This has been done to get an understanding of the relative performance of the algorithms in low- as well as high-dimensional data sets. The data sets have been split in 70:30 ratio between training and test data, respectively.

4.2 *Competing algorithms*

The proposed FIMFS algorithm has been compared to three benchmark algorithms for supervised feature selection. These algorithms are

- correlation-based feature selection (CFS) [25],
- minimum redundancy maximum relevance (mRMR) [14],
- dominant set clustering algorithm (DSCA) [18].

Both CFS and mRMR are well-known algorithms used as benchmarks in many works related to supervised feature selection. CFS is a correlation-based filter method, which identifies relevant features as well as redundancy among relevant features; mRMR uses MI values between individual feature and class as the primary parameter to identify feature subset having mRMR. DSCA is a recent work based on graph-theoretic approach of dominant set clustering, which also uses MI between features in the form of multidimensional interaction information as the feature selection criterion.

Out of the three competing algorithms, mRMR and DSCA use MI as the basis of feature selection and both of them have a filter approach, i.e. they generate feature subsets that are independent of any classifier. In this respect, these algorithms are similar in approach to the proposed algorithm. DSCA algorithm even models the data set features as a feature graph. This way, it is even more similar to FIMFS in terms of solution approach. However, the mRMR algorithm follows a two-stage process. As a first stage, it selects a smaller set of candidate features using filter approach. Then on this smaller set, it applies a wrapper approach (both forward and backward selection) in the second stage to obtain a further smaller and more effective subset. Due to a combination of two stages and a wrapper approach being followed in the second stage, the efficiency of the algorithm gets impacted. DSCA algorithm on the other hand follows a filter approach. It uses multidimensional interaction information as the basis for selecting feature subset. However, the cost of execution of DSCA is higher than that of the proposed approach as FIMFS uses pair-wise interaction information. This is reflected by the higher values of the execution time of DSCA algorithm compared with FIMFS, as presented in the next section. Another novelty of FIMFS is providing the stage-wise visualization of critical and non-critical features, which can be used for the optimal subset selection.

4.3 *Evaluation criteria*

Experiments have been done to evaluate the performance of the proposed algorithm compared to the competing algorithms with respect to two main criteria – accuracy and time taken for execution, as these are the measures for the effectiveness and efficiency, respectively, of any algorithm.

For accuracy, the measurement has been done as follows: accuracy = $\frac{CC}{CC+MC}$, where CC represents the number of correct classifications and MC represents the number of misclassifications. For measuring the classification accuracy with the derived feature subset from each algorithm, three standard classifiers – Decision Tree, Naive Bayes and Support Vector Machine (SVM), have been used. Decision Tree and Naive Bayes are basic classification models, which can be used to test the efficacy of any feature selection algorithm. SVM is one of the most robust and accurate classification algorithms, especially in a high-dimensional feature space. Hence, it has been chosen as the third classifier to compare the performance of the algorithms.

Time taken for execution of the proposed and competing algorithms is measured in a standard computer with 16.0 GB RAM, Intel i5 processor and 64-bit Operating System.

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:

feature reduction = (number of features rejected / total number of features) $\times$ 100.

4.4 *Other details*

For identifying potentially redundant features, pair-wise correlation between features has been calculated. Features having correlation value more than a threshold have been marked as potentially redundant. The value of threshold for correlation coefficient has been assumed as 0.67 [29].

5. Results and analysis

5.1 *Data set FIM generated*

As a part of the FIMFS algorithm, a stage-wise visual representation of the features of the data set in the form of FIM is first generated. Figure 1 gives a summarized view of the stages of FIM formation for all the 8 data sets. This view can give significant information related to potentially irrelevant and redundant features in data sets. Also, the algorithm can be tuned up to generate views for different levels of redundancy (by changing the threshold for correlation coefficient) and relevance (by adopting some other threshold than average MI, e.g. top 'p' features having the highest value of MI).

5.2 *Comparison of efficiency of the algorithms*

Two important parameters indicate the performance of a feature selection algorithm – efficiency and effectiveness.

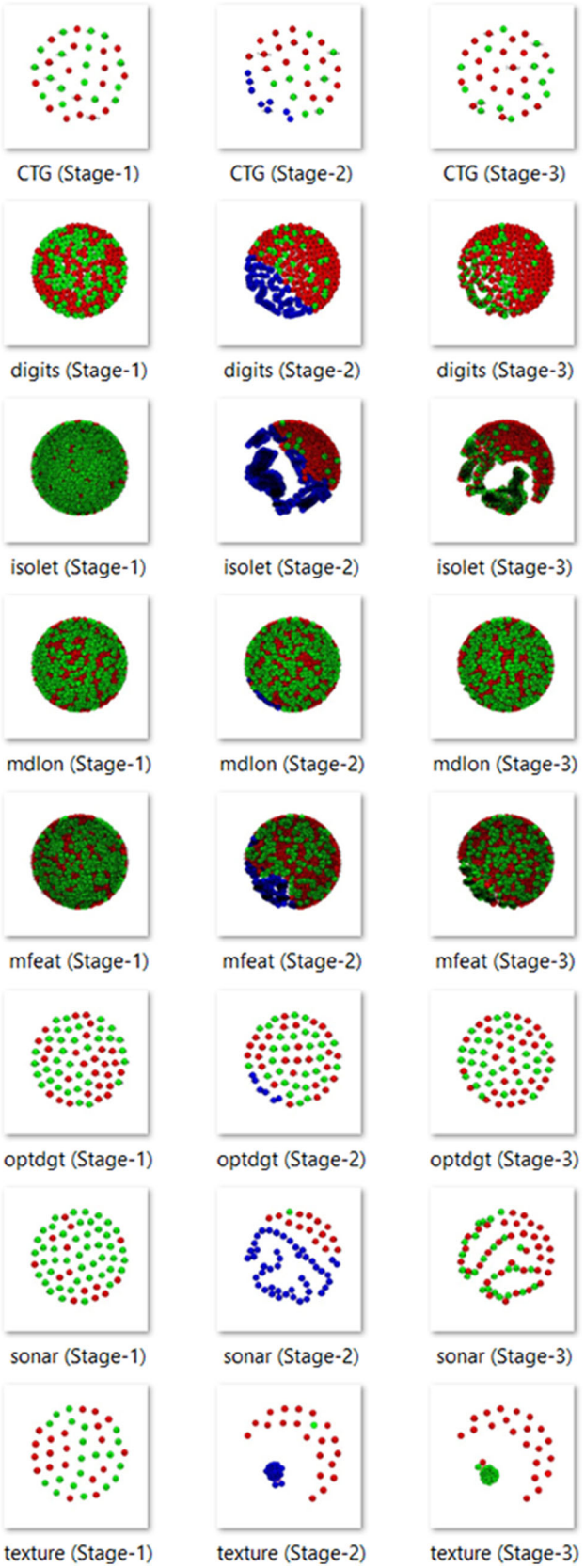

Figure 1. FIM formation – all data sets.

First, the results related to the efficiency of the proposed and competing algorithms are presented. Then the effectiveness measured in the form accuracy, using three different classifiers, has been provided. The comparative results show how good the proposed algorithm is from the perspective of effectiveness. At the end, an inference is drawn about the overall performance to show the suitability of the proposed algorithm in order to address the feature subset selection problem.

Efficiency of the algorithms is reflected by the execution time of the algorithms. Table 2 presents the execution time (in seconds) of the different algorithms. Following observations are made from table 2:

- mRMR takes the highest amount of time amongst all the algorithms, both for lower- as well as higher-dimensional data sets. The time taken is exceptionally high when the feature space is very large (in the current experiments for Digits, Madelon, Isolet and Multi-features data sets).
- Though the performance of CFS closely matches, and is sometimes better than, that of the proposed FIMFS algorithm, for certain data sets (e.g. Digits and Madelon), the execution time is quite high.
- DSCA gives the best performance in terms of execution time after FIMFS algorithm.
- The average execution time of the proposed FIMFS algorithm for the data sets used in the experiments is much better than the competing algorithms. The average execution time of FIMFS is 51 s, whereas those of CFS, mRMR and DSCA are 1674, 2700 and 78 s, respectively.

Overall, the proposed algorithm FIMFS exhibits a much better efficiency than that of the benchmark algorithms, especially in case of high-dimensional data sets, followed by the other graph-based algorithm DSCA.

5.3 *Comparison of effectiveness of the algorithms*

To understand the effectiveness of the proposed FIMFS algorithm compared to two benchmark algorithms namely CFS and mRMR, and also the competing graph-based

Table 2. Time taken for execution (s).

Data set	FIMFS	CFS	mRMR	DSCA
CTG	1.2	*0.3*	16.0	0.6
Texture	0.8	*0.1*	45.6	2.1
Sonar	*0.5*	2.9	129.4	0.6
Optdgt	1.8	*0.6*	108.1	3.1
Digits	*10.1*	316.9	1019.4	14.6
Madelon	*97.3*	12828.0	8244.8	155.3
Isolet	*121.2*	145.9	6947.3	200.9
Mfeat	177.6	*101.0*	5094.1	249.2

Italic values indicate the best performance

Table 3. Classification accuracy (%) of FIMFS vs. other algorithms.

Data set	FIMFS	CFS	mRMR	DSCA	FIMFS	CFS	mRMR	mRMR
	Decision Tree				Naive Bayes			
CTG	*98.3*	43.2	81.6	96.4	93.8	41.7	81.6	*97.5*
Texture	*83*	56.1	79.8	80.7	78.3	58.5	*80.9*	76.8
Sonar	*72.6*	62	57.2	61.3	*72.6*	65.1	57.2	67.8
Optdgt	75.7	50.1	74.3	*76.2*	*91.9*	42.4	90	82.2
Digits	60.6	53.6	56.5	*62.7*	75.9	61.6	63.9	*79.5*
Madelon	*75.4*	66.4	67.7	66	*60.3*	55.4	54.4	56.7
Isolet	*72.2*	6.7	68.9	68.8	76.5	9.2	77.8	*80.1*
Mfeat	91	77.9	*93*	91	*98.5*	91.4	98	95.7
	FIMFS	CFS	mRMR	DSCA	FIMFS	CFS	mRMR	mRMR
	SVM				Combined			
CTG	*99.1*	42.8	81.6	99.1	97.1	42.6	81.6	*97.7*
Texture	98.4	65.1	*99*	98.4	*86.6*	59.9	*86.6*	85.3
Sonar	*87.1*	69.9	73.1	83.9	*77.5*	65.7	62.5	71
Optdgt	*99.1*	72.7	99	98.6	*88.9*	55.1	87.7	85.7
Digits	91.9	85.8	76.4	*94.6*	76.1	67	65.6	*78.9*
Madelon	*56.5*	56.2	55.9	*56.5*	*64.1*	59.3	59.3	59.8
Isolet	89.6	6.2	92.1	*94.5*	79.4	7.4	79.6	*81.1*
Mfeat	98.3	96.7	99.5	98	96	88.7	*96.9*	94.9

Italic values indicate the best performance

algorithm DSCA, the classification accuracies of the algorithms for each of the data sets have been measured. Table 3 presents a detailed view of this comparison. A visual comparison for each of the three classifiers Decision Tree, Naive Bayes and SVM has been presented in figure 2.

Table 4 gives a summary of the performance based on number of "Win"s (or "Tie"s in case more than one algorithm have the highest accuracy) for each data set. The proposed FIMFS algorithm has the highest number of "Win"s (or "Tie"s) followed by the DSCA algorithm. Following are a few observations made from table 3 and figure 2:

- Proposed FIMFS gives a better accuracy than benchmark algorithms for most of the data sets with all three classifiers.
- The only competing algorithm whose results come close to those of FIMFS is DSCA, which is also based on graph-theoretic principles.

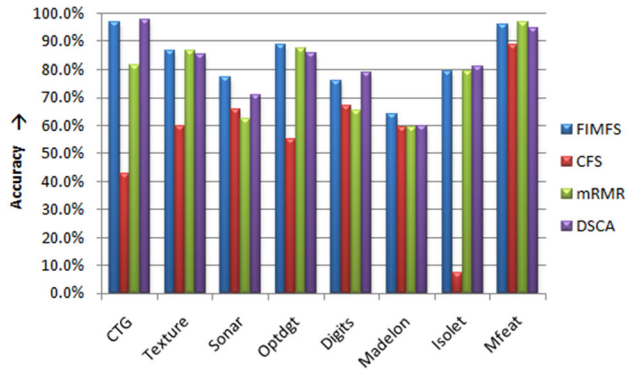

Figure 2. Summary comparison of accuracy.

Table 4. Number of best results (accuracy) of different algorithms.

Data set	FIMFS	CFS	mRMR	DSCA
CTG	2	0	0	2
Texture	1	0	2	0
Sonar	3	0	0	0
Optdgt	2	0	0	1
Digits	0	0	0	3
Madelon	3	0	0	1
Isolet	1	0	0	2
Mfeat	1	0	2	0
Overall	13	0	4	9

Overall, as can be observed in figure 2, the proposed algorithm FIMFS exhibits better effectiveness than that of the competing algorithms.

Table 5. Number of best results (accuracy) of different algorithms.

Data set	Number of features	Subset count	Reduction (%)
Wbdc	30	12	60.00
CTG	34	13	61.80
Texture	40	20	50.00
Sonar	60	14	76.70
Optdgt	63	32	49.20
Digits	257	73	71.60
Madelon	500	184	63.20
Isolet	618	183	70.40
Mfeat	649	224	65.50

Table 6. Summary of results of different algorithms.

	FIMFS	CFS	mRMR	DSCA
Average accuracy (lower-dimension data sets, %)	*87.5*	55.8	79.6	84.9
Average accuracy (higher-dimension data sets, %)	*78.9*	55.6	75.4	78.7
Average accuracy (overall, %)	*83.2*	55.7	77.5	81.8
Mean rank (accuracy)	*1.75*	3.63	2.50	2.00
Average execution time (s)	*51*	1674	2700	78

Italic values indicate the best performance

5.4 *Feature reduction of FIMFS*

Table 5 presents the feature subset count and the feature reduction percentage achieved using FIMFS. As is evident, a significant amount of reduction (average 63.5%) is obtained using FIMFS algorithm. This percentage can be further increased by tuning user input values to the algorithm, e.g. similarity threshold α_0.

5.5 *Analysis of results*

Table 6 captures a summary of performance of all four algorithms. Following inferences can be made based on the analysis of results:

- Proposed FIMFS gives a better average accuracy with all three classifiers – Decision Tree, Naive Bayes and SVM. Also, the mean rank of FIMFS calculated using accuracy-based ranking of algorithms is better than those of the other algorithms. Hence, it demonstrates better effectiveness based on the experiments conducted.
- Average execution time is also the least for the proposed algorithm. Hence, FIMFS demonstrates superior efficiency compared with the competing algorithms.
- Results are equally promising as the data set dimension increases, thus making it a good candidate for supervised feature selection of high-dimensional data sets.

6. Conclusion

In this paper an information-theory-based graph theoretic approach of feature selection has been proposed. As a part of the approach, feature-to-class MI is measured based on which the input data set is modelled as a graph. The graph is termed as feature information map or FIM. FIM helps in representing feature relevance and redundancy in a novel way. The vertices of FIM representing features that have high relevance and low redundancy are ultimately marked as "green", indicating that they are most suitable candidates to be chosen as the final feature subset.

Experiments have been conducted using eight publicly available benchmark data sets, with some having high number of features. The results of the experiments are quite promising. The proposed algorithm demonstrates better results, with respect to both efficiency and effectiveness, compared with the competing algorithms. The average accuracy obtained using subsets derived by proposed FIMFS algorithm is 83.5% which is much higher compared with the competing CFS, mRMR and DSCA algorithms, giving average accuracy of 55.7%, 77.5% and 81.8%, respectively. The average execution time for FIMFS algorithm is 51 s compared with the closest execution time of 78 s of DSCA algorithm. Both CFS and mRMR have average execution time of a few 1000 s. The results are equally good in case of high-dimensional data sets, proving the efficacy of the proposed algorithm for feature selection of high-dimensional data sets.

References

[1] Cao L 2016 Data science and analytics: a new era. *Int. J. Data Sci. Anal.* 1: 1–2

[2] Morgulev E, Azar O H and Lidor R 2017 Sports analytics and the big-data era. *Int. J. Data Sci. Anal.* 5(4): 213-222

[3] Moujahid A and Dornaika F 2017 Feature selection for spatially enhanced LBP: application to face recognition. *Int. J. Data Sci. Anal.* 5: 11–18

[4] Bandyopadhyay S, Bhadra T, Mitra P and Maulik U 2014 Integration of dense subgraph finding with feature clustering for unsupervised feature selection. *Pattern Recognit. Lett.* 40: 104–112

[5] Liu H and Yu L 2005 Toward integrating feature selection algorithms for classification and clustering. *IEEE Trans. Knowl. Data Eng.* 17: 491–502

[6] Dash M and Liu H 1997 Feature selection for classification. *Intell. Data Anal.* 1: 131–156

[7] John G H, Kohavi R and Pfleger K 1994 Irrelevant features and the subset selection problem. In: *ICML Proceedings*, pp. 121–129

[8] Das A K, Goswami S, Chakraborty B and Chakrabarti A 2016 A graph-theoretic approach for visualization of data set feature association. In: *Advanced Computing and Systems for Security*, vol. 4, pp. 109–124

[9] Goswami S, Das A K, Chakrabarti A and Chakraborty B 2017 A feature cluster taxonomy based feature selection technique. *Expert Syst. Appl.* 79: 76–89

[10] Goswami S, Guha P, Tarafdar A, Das A K, Chakraborty S, Chakrabarti A and Chakraborty B 2017 An approach of

feature selection using graph-theoretic heuristic and hill climbing. *Pattern Anal. Appl.* 22(2): 615–631

[11] Liu H and Motoda H 2009 Computational methods of feature selection. *Inf. Process. Manag.* 45: 490–493

[12] Tang J, Alelyani S and Liu H 2014 Feature selection for classification: a review. In: *Data Classification: Algorithms and Applications*, pp. 37–64

[13] Das A K, Goswami S, Chakrabarti A and Chakraborty B 2017 A new hybrid feature selection approach using Feature Association Map for supervised and unsupervised classiffication. *Expert Syst. Appl.* 88: 81–94

[14] Peng H, Long F and Ding C H 2005 Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27: 1226–1238

[15] Estvez P A, Tesmer M, Perez C A and Zurada J M 2009 Normalized mutual information feature selection. *IEEE Trans. Neural Netw.* 20: 189–201

[16] Battiti R 1994 Using mutual information for selecting features in supervised neural net learning. *IEEE Trans. Neural Netw.* 5(4): 537–550

[17] Hoque N, Bhattacharyya D K and Kalita J K 2014 MIFS-ND: a mutual information-based feature selection method. *Expert Syst. Appl.* 41: 6371–6385

[18] Zhang Z and Hancock E R 2011 A graph-based approach to feature selection. In: *GBRPR Proceedings*, pp. 205–214

[19] Zhang Z and Hancock E R 2012 Hypergraph based information-theoretic feature selection. *Pattern Recognit. Lett.* 33: 1991–1999

[20] Tsamardinos I, Brown L E and Aliferis C F 2006 The max–min hill-climbing Bayesian network structure learning algorithm. *Mach. Learn.* 65(1): 31–78

[21] Gasse M, Aussem A and Elghazel H 2014 A hybrid algorithm for Bayesian network structure learning with application to multi-label learning. *Expert Syst. Appl.* 41(15): 6755–6772

[22] Chow C K and Liu C N 1968 Approximating discrete probability distributions with dependence trees. *IEEE Trans. Inf. Theory IT* 14(3): 462–467

[23] Zare H and Niazi M 2016 Relevant based structure learning for feature selection. *Eng. Appl. Artif. Intell.* 55: 93–102

[24] Huang S *et al* 2013 Alzheimer's disease neuroimaging initiative—a sparse structure learning algorithm for Gaussian Bayesian network identification from high-dimensional data. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(6): 1328–1342

[25] Hall M A 2000 Correlation-based feature selection for discrete and numeric class machine learning. In: *ICML Proceedings*, pp. 359–366

[26] Khan I and Khan S 2014 Experimental comparison of five approximation algorithms for minimum vertex cover. *Int. J. Sci. Technol.* 7: 69–84

[27] Li S *et al* 2011 An algorithm for minimum vertex cover based on Max-I share degree. *J. Comput.* 6: 1781–1788

[28] Lichman M and Bache K 2013 UCI machine learning repository [online]. Available: http://archive.ics.uci.edu/ml. Accessed 10 Oct 2018

[29] Taylor R 1990 Interpretation of the correlation coefficient: a basic review. *J. Diagn. Med. Sonogr.* 6: 35–39