INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Feature Selection in High Dimension Datasets using Incremental Feature Clustering

Damodar Patel[1], **Amit Kumar Saxena**[2]*

1 Research Scholar, Department of CSIT, Guru Ghasidas Vishwavidyalaya, Bilaspur, India
2 Professor, Department of CSIT, Guru Ghasidas Vishwavidyalaya, Bilaspur, India

OPEN ACCESS

Received: 24-06-2024

Accepted: 27-07-2024

Published: 14-08-2024

Citation: Patel D, Saxena AK (2024) Feature Selection in High Dimension Datasets using Incremental Feature Clustering. Indian Journal of Science and Technology 17(32): 3318-3326. https://doi.org/10.17485/IJST/v17i32.2077

* **Corresponding author**.

amitsaxena65@rediffmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To develop a machine learning-based model to select the most important features from a high-dimensional dataset to classify the patterns at high accuracy and reduce their dimensionality. **Methods**: The proposed feature selection method (FSIFC) forms and combines feature clusters incrementally and produces feature subsets each time. The method uses K-means clustering and Mutual Information (MI) to refine the feature selection process iteratively. Initially, two clusters of features are formed using K-means clustering (K=2) by taking features as the basis of clustering instead of taking the patterns (a traditional way). From these two clusters, the features with the highest MI value in each cluster are kept in a feature subset. Classification accuracies (CA) of the feature subset are calculated using three classifiers namely Support Vector Machines (SVM), Random Forest (RF), and k-nearest Neighbor (knn). The process is repeated by incrementing the value of K i.e. number of clusters; until a maximum user-defined value of K is reached. The best value of CA obtained from these trials is recorded and the corresponding feature set is finally accepted. **Findings**: The proposed method is demonstrated using ten datasets and the results are compared with the existing published results using three classifiers to determine the method's performance. The ten datasets are classified with average CAs of 92.72%, 93.13%, and 91.5%, using the SVM, RF, and K-NN classifiers respectively. The proposed method selects a maximum of thirty features from the datasets. In terms of selecting the most effective and the smallest feature sets, the proposed method outperforms eight other feature selection methods considering CAs. **Novelty**: The proposed model applies feature reduction using combined feature clustering and filter methods in an incremental way. This provides an improved selection of relevant features while removing those which are irrelevant at different trials.

Keywords: Feature selection; High-dimensional datasets; K-means algorithm; Mutual information; Machine learning

1 Introduction

The High-dimensional data affects learning task performance and causes challenges with data mining [1] and machine learning (ML) activities, including data analysis, information retrieval processing, and data/pattern classification. The Classifier models often cause overfitting, low effectiveness, and low classification accuracy. A typical dataset consists of patterns (or instances), features (or attributes), and classes (or labels). In these original datasets, some features are relevant while the remaining features can be irrelevant, and feature selection (FS) [2] is an essential preprocessing step to identify and remove the irrelevant features from the original dataset. In terms of data analysis and identification, ML and data mining applications use FS and classification techniques. It has the potential to significantly reduce the ”curse of dimensionality” and the high computing costs associated with data analysis. Therefore, financial, medical, intrusion detection and active power filters widely use FS as an important data preprocessing technique.

There are three types of FS methods commonly used based on filter [3,4], wrapper, and embedded approaches. The statistical measurements are used in filter methods to evaluate the ranks of features. The wrapper methods on the other hand, group features and evaluate the metric (like accuracy) produced by that group of features (subsets). The embedded methods combine the wrapper and filter approaches in a grouping of features. This paper applies ML for feature selection. The ML consists of efficient algorithms including Support Vector Machines (SVM), Random Forest (RF), k-nearest neighbor (knn), fuzzy logic [5], etc.

The development of the FS method has been widely published in the literature including a few research works presented as follows. Rani et al. [6] introduced GARFE, a hybrid FS method that utilized recursive and evolutionary feature reduction techniques, to deal with this issue. The classification system eliminates features that are not relevant. Jiajun et al. [7] proposed MFS-AGD, which incorporated adaptive graph diffusion for learning high-order structure, and HSIC to maximize class dependency. LMFS-AG extends this to enhance discriminative power. In order to extract relevant classification information, Zhang et al. [8] introduced a novel approach called CWJR-FS that made use of conditional-weight joint relevance. Gaussian vote feature selection (GVFS) is a revolutionary approach that Wang et al. [9] presented for diagnosing problems in permanent magnet DC motors (PMDCMs). By reducing unnecessary data using the Gaussian probability density function (GPDF), GVFS claims to improve diagnostic accuracy over other methods. Khan et al. [10] propose to hybridize a slime mold algorithm (SMA) with binary grey wolf optimization (BGWO) for FS. First, they used eight different optimization techniques then based on the outcomes combined the best two algorithms to create a hybrid method. Li et al. [11] proposed Deep Feature Screening (DeepFS), a two-step nonparametric method that extracted low-dimensional data representations and applied the feature screening using multivariate rank distance correlation. Based on clustering algorithms and a correlation coefficient, Atmakuru et al. [3] proposed two enhanced filter-based feature selection approaches. The first approach assigns a neighborhood to each feature based on feature correlation and then forms the subsets of features that are more similar than a specific threshold. The second technique finds features that are used to represent the entire cluster by applying clustering analysis to the correlation data. Wu et al. [12] propose an enhanced RIME (ERIME: Enhanced RIME) approach to combine the DBSCAN spatial clustering particle evolution method with feature information entropy pruning. Yang et al. [13] proposed an informative feature clustering and selection technique. The technique consists of two phases. In the first phase, a feature clustering (FC) technique is used, and in the second phases, a stratified feature selection (SFS) technique is used. Haq et al. [14] introduced a novel feature selection framework that combines an exemplar-based clustering algorithm of variables with multiple feature-ranking techniques for selecting the best feature subset. Hashemi et al. [15] proposed a fast feature selection algorithm based on clustering ranking in feature-label space and L2-norm, called the MLCR method. First, the k-means algorithm to cluster the features based on their correlation with labels. Then sorted the features in each cluster based on the L2-norm in descending order and finally set the rank to each feature. In the second step, the features with the same rank are sorted like in the first step and added to the feature ranking vector. Chavent et al. [16] proposed a methodology that combines the Hierarchical clustering of features and random forests FS method. Hierarchical clustering is used to build groups of correlated features and summarizes each group by a synthetic feature, and a random forest method is used to identify the most relevant synthetic features. Sun et al. [17], present a feature reduction technique that combines adaptive weighted k-nearest neighbors (AWKNN) with similarity-based feature clustering. Zhao et al. [18] propose that the FS method known as FSRRW enhances classification in data mining and ML by considering both relevance and redundancy. For the purpose of extracting important data, the method calculates a relevant-redundant weight (RRW) and evaluates feature relevance using this weight to extract important data. In this paper, for comparison of the FSRRW method, seven FS methods, namely CWJR, MIFS, CIFE, mRMR, JMI, MRI, and DCSF, are used. In the current research paper, the proposed method, FSIFC, is compared to eight FS methods (CWJR, MIFS, CIFE, mRMR, JMI, MRI, DCSF, and FSRRW). One of the objectives was also to obtain the average CA using the proposed method which is observed better than those noted using eight FS methods.

As seen, more often than not, in the feature clustering algorithms reviews, the number of clusters (K) is fixed at the beginning of any algorithm which prevents flexibility in deciding the formation of a cluster of appropriate size. This serves as a research gap in our perception. The present study attempts to address the problem of fixing the number of clusters tentatively by incrementing and provides a scope to consider the appropriate although approximate number of clusters in an incremental manner. Further, the arrangement of placing the features and pattern is reversed in the proposed method, and in the present work the features are taken in rows and the patterns are taken in columns (transpose of the traditional approach). The present study places distinct (non-common) features in each of the clusters to show the highest intra similarity among the feature in each of the clusters. This is the novel feature of the proposed work.

The remainder of the paper is structured as follows: Section 2 shows the Methodology. Section 3 includes the experiment results and discussions. Section 4 contains the conclusion and scope for future work.

2 Methodology

This section describes the proposed method, dataset description, model development, and experiment parameters.

2.1 Proposed Method

FS methods use different evaluation metrics to obtain the best feature subset but usually do not look for feature structure. A better method that eliminates irrelevant or noisy data while recognizing the structure of features is to enforce clustering. The dimensionality issue is resolved, and performance is improved when feature clustering and filter methods are combined rather than when filter evaluation measures are used separately. This research proposed a method in which the filter measure selects only relevant features and ranks the features according to their priority. The clustering algorithm eliminates irrelevant features. The K-means algorithm [19] is used for the clustering, and mutual information [20] is used for the filtering. The combination of filtering algorithm with clustering helps in, firstly clustering the features based on their similarities and secondly, deciding the ranks of individual features in a cluster on the basis of their respective MI values. Initially, two clusters are created out of all features, and then select best features of each cluster based on their individual mutual information values. After obtaining the best feature subset, calculate the classification accuracy (CA) using Support Vector Machines (SVM) [21], Random Forest (RF) [22], and k-nearest neighbor (knn) [23] classifiers. If the stopping criteria Equation (1) are not met, then increase the number of clusters by one and repeat the steps until the stopping criteria are reached. Once the stopping criterion is reached, compare CAs of all feature subsets and select the feature subset having the best CA. The proposed method selects only one representative feature from each cluster to address the redundancy issue. The complete flow chart of the proposed method is given in Figure 1.

Algorithm used in the FSIFC Method

- Given f_t: number of all features in the original dataset.
- Calculate the mutual information values of all features in the dataset.
- Initially, K = 2 (K is the number of clusters).
- Apply the K-means clustering algorithm in the dataset with all features. It is to be noted that all features will be placed in rows while patterns in columns.
- Select the feature with the highest mutual information value of each cluster. Keep all these features in a separate feature subset FS_K: K being the number of features in FS_K.
- Determine the CAs of selected feature set FS_K using classifiers SVM, RF, knn.
- Check the condition (stopping criteria), K_{max} the maximum number of clusters to be taken

$$K_{max} \leq \begin{Bmatrix} 15,\ 0 < f_t < 100 \\ 20,\ 100 \leq f_t < 300 \\ 25,\ 300 \leq f_t < 500 \\ 30,\ 500 \leq f_t \end{Bmatrix} \tag{1}$$

- $K = K + 1$.
- Repeat steps 4 to 8 till $K \leq K_{max}$.
- Find out the feature subset having the highest CA (Step-6) given by any classifier from all K trials ($K \leq K_{max}$).

The threshold values used in Equation (1) have been taken to match the maximum limit of a number of features used in [18].

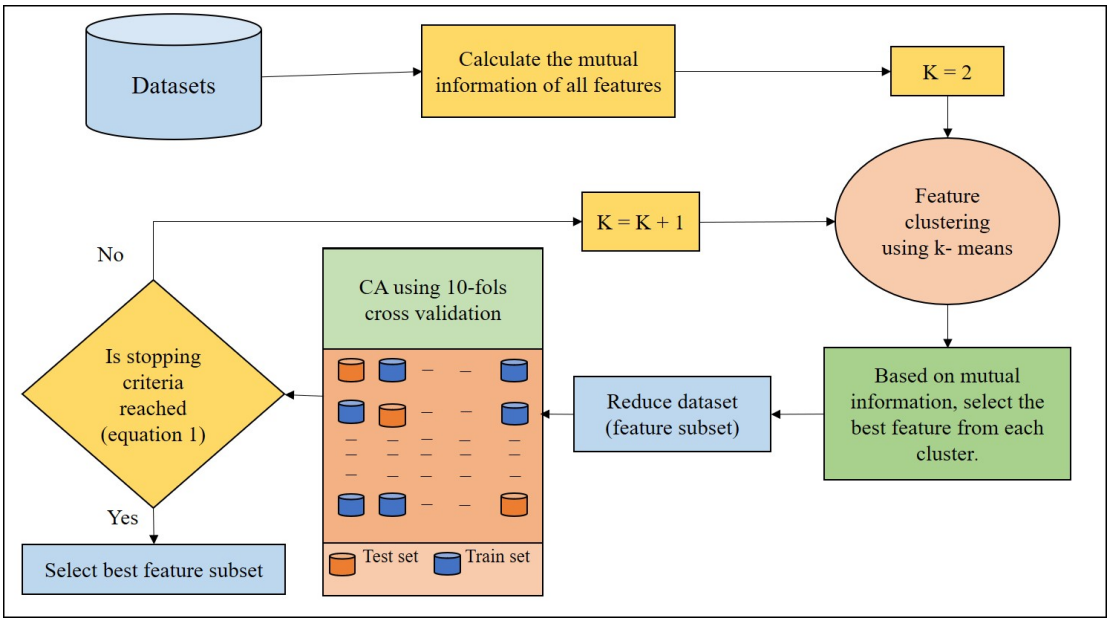

Fig 1. Proposed FSIFC feature selection method

2.1.1 K-means algorithms

K-means clustering is a popular unsupervised ML method for splitting a dataset into K-distinct clusters. First, the algorithm initializes K cluster centroids at random. The nearest centroid for each data point in the dataset is then determined using the Euclidean distance. The centroids are recalculated as the mean of all the points in each cluster once all the points have been allocated. Iteratively assigning and updating the centroids continues until either an established number of iterations is reached or the centroids no longer show significant changes. K-means aims to minimize the sum of squared distances within a cluster, hence maximizing the distance between data points within a cluster and minimizing their difference from points outside of the cluster.

2.1.2 Mutual Information

A metric from information theory called Mutual Information (MI) estimates the amount of information about one random variable that can be obtained about another random variable. It captures their dependence by measuring a reduction in uncertainty about one variable given knowledge of the other. The mathematical definition of MI between two discrete random variables, X and Y, is:

$$I(X;Y) = \sum_{x,y} P(x,y) log \frac{P(x,y)}{P(x)\,P(y)} \tag{2}$$

where I(X,Y) represent mutual information, P(x,y) represents X and Y joint probability distribution function, and P(X) and P(Y) represent X and Y marginal probability distribution functions. Additionally, present in terms of entropy:

$$I(X;Y) = H(X) - H(X\ )\ Y)$$

where H(X) is the marginal entropy, H(X | Y) is the conditional entropy of X given Y, which measures the remaining uncertainty of X when Y is known.

2.2 Datasets description

Evaluation is conducted on ten real benchmark datasets obtained from ASU [24] and the UCI machine learning repository [25] to determine the effectiveness of the proposed method. A detailed description of the datasets is provided in Table 1. The majority of the datasets are high-dimensional where four datasets are binary class and six are multiclass datasets.

Table 1. Description of the datasets

Datasets	Features	Records	Classes
ALLMAIL	7129	72	2
Ionosphere	34	351	2
Lung	3312	203	5
Lung discrete	325	73	7
lymphoma	4026	96	9
QUSAR_B (QSAR biodegradation)	41	1055	2
ORL	1024	400	40
orlraws10P	10304	100	10
Thoracic Surgery	17	470	2
warpPIE10P	2420	210	10

2.3 Experiments

Python 3.11.9 was used to execute the FSIFC method on a Windows 10 PC with an Intel (R) Core i5, 8 GB of RAM, a 2.30 GHz CPU, and a 500 GB SSD card. The proposed FSIFC approach evaluates the CA obtained by applying three classification models: SVM (kernel = rbf), RF, and knn (k = 3). Evaluate the models using the available labels at the data source. The experiment employed a standard scaler to normalize the data. A ten-stratified k-fold dataset was used in this experiment. Table 2 displays the performance measures of the ten-stratified k-fold cross-validation datasets.

2.4 Model Development

This study employed the three most widely used classifiers: SVM, RF, and knn. These classifiers were used to evaluate the performance of the proposed FS approach to each dataset.

2.4.1 Support Vector Machine (SVM)

SVM finds the maximum margin hyperplane, or linear separator, as far away from the positive and negative training information as possible. We can use kernel functions to project the data into a high-dimensional space, thereby making the linear separator highly non-linear in the input space.

2.4.2 Random Forest (RF)

RF is a popular ML algorithm in the supervised learning approach. ML tasks such as regression and classification utilize it. The concept of ensemble learning serves as its foundation. Combine the results of many decision trees to arrive at a single conclusion. To predict the final outcome, the random forest uses the majority vote on the predictions made by each tree.

2.4.3 K-Nearest Neighbors (knn)

In ML, knn is a simple and straightforward categorization technique. This non-parametric technique is used for regression as well as for classification. The knn classifies test data according to the known labels of the k closest neighbors using a distance metric similar to the Euclidean distance. The class of the test patterns is determined by most of the class labels predicted by the training patterns.

3 Results and Discussion

Table 2 illustrates the results of the proposed method comprehensively. In Table 2, the first column shows the name of each dataset used for the experiments. The second to fourth columns show the CA using SVM, RF, and knn classifiers with original (base) full feature sets. The fifth to seventh columns show the CA of the selected feature in the proposed FSIFC method using the same three classifiers. The last column shows the number of selected features (SF). The highest value in each row is presented in bold.

Table 2 and Figure 2 show a comparison of the CA of the original feature sets and selected features using three classifiers. As an example, the orlraws10P dataset is taken for explanation purpose. In this dataset, the highest CA (99%) is achieved with all features (10304). The highest base CAs using SVM, RF, and knn were 99%, 97.09%, and 92%, respectively. After applying

Table 2. Classification accuracies (in %) obtained with original datasets (All features) and with selected features using three classifiers

Datasets	SVM	RF	knn	FSIFC for SVM	FSIFC for RF	FSIFC for knn	SF(%)
ALLMAIL	84.64	95.23	81.79	95.71	**99.86**	97.14	16 (0.22)
Ionosphere	**96**	95.6	90.75	93.25	93.55	87.5	15 (44.12)
Lung	93.1	93.4	96.02	96.55	94.82	**97.02**	26 (0.79)
Lung discrete	79.46	82.95	86.07	**91.61**	88.11	84.82	17 (5.23)
lymphoma	81.11	83.76	90.22	91.44	91.54	**92.56**	21(0.52)
QUSAR_B	**86.81**	85.24	82.26	83.69	84.72	84.26	12 (29.27)
ORL	96	**96.05**	90.75	91.75	93	86	30(2.93)
orlraws10P	**99**	97.9	92	98	**99**	97	28(0.27)
Thoracic Surgery	**85.11**	84.74	82.13	**85.11**	84.85	82.55	9 (52.94)
warpPIE10P	95.71	96.48	95.24	95.71	97.24	**99.05**	28 (1.16)
Average	90.24	91.62	89.32	92.72	**93.13**	91.5	20.2(13.74)

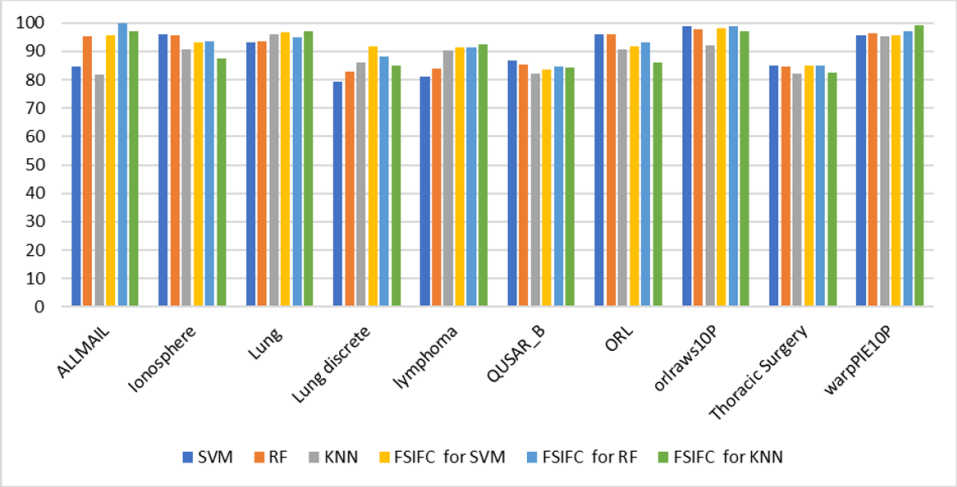

Fig 2. Classification Accuracies obtained from all ten datasets with all features versus selected features using three classifiers

FSIFC, the highest CAs using SVM, RF, and knn were 98%, 99%, and 97%, respectively. The important fact is that these CAs are achieved with only 28 features which is 0.27%. Table 2 refers to all results obtained after experimenting with FSIFC. The average CA of ten datasets with base datasets CA using SVM, RF, and Knn were 90.24%, 91.62%, and 89.32%, respectively, while after applying FSIFC, the average CA for all same ten datasets using SVM, RF, and knn were 92.72%, 93.13%, and 91.5%, respectively which is higher than those achieved with base datasets. Moreover, the RF classifier shows the maximum average CA (93.13%) with the selected feature set. The proposed FS method performs effectively, providing better CA with just a few features. From Table 2, it is also noted that the average SF is 20.2%. These findings indicate that the proposed FSIFC approach could reduce the dimensionality of features while improving model performance.

Tables 3, 4 and 5 compare the FSIFC method with the other eight FS methods for SVM, RF, and knn classifiers respectively. Each of the Tables 3, 4 and 5 lists the names of each dataset in the first column. Columns two to nine show the CA using eight FS methods namely: CWJR, MIFS, CIFE, mRMR, JMI, MRI, DCSF, and FSRRW. The CA of the FSIFC method is presented in the tenth column.

Table 3. Comparison of Classification accuracies (in %) obtained with proposed FSIFC method to other eight methods with SVM classifier

Datasets	CWJR	MIFS	CIFE	mRMR	JMI	MRI	DCSF	FSRRW	FSIFC
ALLAML	95.31	94.25	82.04	94.37	95.39	92.47	91.47	**96.58**	95.71
Ionosphere	92	91.29	91.03	91.54	92.51	91.56	92.16	92.61	**93.25**
Lung	93.03	88.11	77.79	89.84	91.91	90.83	89.37	92.5	**96.55**

Continued on next page

Table 3 continued

lung-discrete	77.85	69.5	58.45	66.89	76.43	75.73	76.12	78.5	**91.61**
lymphoma	86.44	77.69	55.06	78.1	84.26	84.47	84.09	86.73	**91.44**
ORL	69.78	61.95	39.18	67.19	67.46	62.64	63.41	69.79	**83.69**
orlraws10P	**95.27**	82.2	53.9	87.1	92.83	84.03	78.57	93.67	91.75
QSAR_B	80.77	80.02	77.9	81.24	78.58	79.08	80.86	81.22	**98**
Thoracic Surgery	83.47	83.54	81.49	82.63	81.71	81.57	82.07	82.09	**85.11**
warpPIE10P	87.32	86.56	74.22	86.78	86.8	88.31	88.54	88.66	**95.71**
Average	85.42	80.63	66.35	82.07	84.05	81.79	81.06	85.58	**92.72**

Table 3 shows that the FSIFC method, outperformed all other FS methods on eight out of the ten datasets except in the ALLAML dataset where the FSRRW method performed better with CA 96.58% against the CA 95.71% by the FSIFC method. In orlraws10p dataset, the CWJR method performed better with 95.27% CA against the CA 91.75%% by the FSIFC method. On an average, out of ten datasets, the FSIFC achieved a better CA of 92.72% than other FS methods.

Table 4. Comparison of Classification accuracies (in %) obtained with proposed FSIFC method to other eight methods with RF classifier

Datasets	CWJR	MIFS	CIFE	mRMR	JMI	MRI	DCSF	FSRRW	FSIFC
ALLAML	93.64	94.69	82.73	94.92	97.44	93.89	93.81	99.02	**99.86**
Ionosphere	88.24	88.26	83.19	83.19	84.37	83	82.75	87.54	**93.55**
Lung	**95.75**	90.79	77.61	92.04	94.36	93.46	93.34	95.14	94.82
lung-discrete	90.17	88.29	69.5	81.98	88.77	88.22	88.29	**91.96**	88.11
lymphoma	95.73	90.58	56.98	90.09	94.45	94.37	93.56	**96.28**	91.54
ORL	84.86	82.03	58.55	82.88	83.38	83.64	85.22	**86.02**	84.72
orlraws10P	**97.07**	92.5	72.97	95.43	95.73	92.9	91.23	95.93	93
QSAR_B	79.72	80.1	77.76	77.76	78.48	79.73	80.83	81.48	**99**
Thoracic Surgery	85.11	85.11	85.11	85.11	85.11	85.11	85.11	**85.11**	84.85
warpPIE10P	91.89	91.88	90.14	94.71	94.71	93.86	95.36	96.52	**97.24**
Average	89.49	88.26	73.94	87.68	88.67	88.17	88.24	91.46	**93.13**

Table 4 shows that the FSIFC method outperformed all other FS methods on four out of ten datasets except in the Lung dataset where the CWJR method performed better with CA 95.75% against the CA 94.82% by the FSIFC method. in the lung-discrete dataset where the FSRRW method performed better with CA 91.96% against the CA 88.11% by the FSIFC method. In the lymphoma dataset where the FSRRW method performed better with CA 96.28% against the CA 91.54% by the FSIFC method. in the ORL dataset where the FSRRW method performed better with CA 86.02% against the CA 84.72% by the FSIFC method. in the orlraws10p dataset where the CWJR method performed better with CA 97.07% against the CA 93% by the FSIFC method. In the Thoracic Surgery dataset, the FSRRW method performed better with 85.11% CA against the CA 84.85% by the FSIFC method. On average, out of ten datasets, the FSIFC achieved a better CA of 93.13% than other FS methods.

Table 5. Comparison of Classification accuracies (in %) obtained with proposed FSIFC method to other eight methods with knn classifier

Datasets	CWJR	MIFS	CIFE	mRMR	JMI	MRI	DCSF	FSRRW	FSIFC
ALLAML	96.78	95.03	75.53	95.95	97.42	91.58	89.02	**98.39**	97.14
Ionosphere	89.45	89.83	87.82	89.91	88.92	89.29	87.34	**89.92**	87.5
Lung	95.36	89.17	73.13	93.74	95.56	94.93	94.37	96.04	**97.02**
lung-discrete	90.08	84.42	66.28	80.46	88.37	87.87	87.3	**90.84**	84.82
lymphoma	92.3	89.21	58.56	89.68	89.9	92.01	92.62	**94.33**	92.56
ORL	83.54	81.03	52.69	82.92	81.12	80.63	82.43	83.78	**84.26**
orlraws10P	**94.7**	89.5	63.17	93.93	92.6	87.6	85.37	94.6	86
QSAR_B	76.21	75.67	75.96	77.94	75.14	76.86	77.79	78.46	**97**
Thoracic Surgery	81.16	81.01	80.94	80.57	80.75	81.49	81.35	81.84	**82.55**
warpPIE10P	91.12	82.63	80.41	94.38	93.85	93.14	94.59	95.58	**99.05**
Average	88.22	85.02	68.31	87.05	87.44	85.46	85.31	89.12	**91.50**

Table 5 shows that theFSIFC method outperformed all other FS methods on five out of ten datasets except in the ALLAML dataset where the FSRRW method performed better with CA 98.39% against the CA 97.14% by the FSIFC method. in the Ionosphere dataset where the FSRRW method performed better with CA 89.92% against the CA 87.5% by the FSIFC method. In the lung-discrete dataset where the FSRRW method performed better with CA 90.84% against the CA 84.82% by the FSIFC method. in the lymphoma dataset where the FSRRW method performed better with CA 94.33% against the CA 92.56% by the FSIFC method. in the orlraws10p dataset where the CWJR method performed better with CA 94.7% against the CA 86% by the FSIFC method. On average, out of ten datasets, the FSIFC achieved a better CA of 91.50% than other FS methods.

In general, these findings show that the proposed FSIFC approach in high-dimensional datasets using incremental feature clustering effectively determines the optimal feature subset and improves the classifiers' performance on the majority of the datasets. This justifies the FS approach's robustness and effectiveness in improving CA on the one hand while reducing feature dimensionality on the other hand to an average of 20.2 features.

4 Conclusion

This paper presents a new method FSIFC for feature selection. The method iteratively forms different subsets of features and evaluates the classification accuracy (CA) every time. The formation of subsets is through a novel approach. The original datasets with all features are divided into K number of clusters using the famous K means clustering algorithm. After clustering into K groups, the features of each cluster are ranked on the basis of their individual Mutual Information (MI) values in that particular cluster. The feature with the highest MI value of a cluster is added to a feature subset. Thus, K features are kept in the resulting feature subset. The CA of the feature subset is calculated using three benchmark classifiers SVM, RF, and knn. The process is repeated with incremental values of K = 2,3,4,..to a predefined maximum value of K. At the stopping state, the feature set producing the highest CA by any of the three classifiers is recorded and the classifier with that feature subset is used to serve as the model.

The FSIFC method is compared with the existing other eight methods on ten datasets. It was observed that the FSIFC method has a competitive; mostly superior performance compared to the other eight existing methods with a trade-off on a few of the ten databases. In addition, the FSIFC also reduces the dimensionality of the datasets to a good extent. On an average 20.2 number of features i.e. 13.74% of the total number of features, were required to produce the CA competitive and even better than the other existing methods which were reported in the literature. The FSISC although attempts to succeed in its objectives of feature selection with iterative and incremental clustering, yet has a few limitations as well. The decision of estimating K as the stopping criterion, the case of empty clusters or clusters with a very small number of features, considering other ranking metrics of features while selecting representative feature(s) from each cluster or even imposing some penalty to excess of clusters or features are some of the areas which need attention to researchers on their upcoming researches.

References

1) Yang XS. Introduction to algorithms for data mining and machine learning. . Academic press. 2019;p. 109–128. Available from: https://doi.org/10.1016/B978-0-12-817216-2.00013-2.
2) Silaich S, Gupta S. Feature Selection in High Dimensional Data: A Review. *Lecture Notes in Networks and Systems* . 2023. Available from: https://doi.org/10.1007/978-981-19-9225-4_51.
3) Atmakuru A, Fatta GD, Nicosia G, and AB. Improved Filter-Based Feature Selection Using Correlation and Clustering Techniques. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2024. Available from: https://doi.org/10.1007/978-3-031-53969-5_28.
4) Patel D, Saxena AK, Laha S, Ansari GM. A Novel Scheme for Feature Selection Using Filter Approach. In: and others, editor. Proceedings of the 2022 7th International Conference on Computing, Communication and Security, ICCCS 2022 and 2022 4th International Conference on Big Data and Computational Intelligence;vol. 2022. Institute of Electrical and Electronics Engineers Inc. 2022. . Available from: https://doi.org/10.1109/ICCCS55188.2022.10079604.
5) Bharill N, Tiwari A, Malviya A, Patel OP, Gupta A, Puthal D, et al. Fuzzy knowledge based performance analysis on big data. *Neurocomputing*. 2020;389:218–228. Available from: https://doi.org/10.1016/j.neucom.2018.10.088.
6) Rani P, Kumar R, Jain A, Chawla SK. A hybrid approach for feature selection based on genetic algorithm and recursive feature elimination. *International Journal of Information System Modeling and Design*. 2021;12(2):17–38. Available from: https://doi.org/10.4018/IJISMD.2021040102.
7) Ma J, Xu F, Rong X. Discriminative multi-label feature selection with adaptive graph diffusion. *Pattern Recognition*. 2024;148:110154–110154. Available from: https://doi.org/10.1016/j.patcog.2023.110154.
8) Zhang P, Gao W, Hu J, Li Y. A conditional-weight joint relevance metric for feature relevancy term. *Eng Appl ArtifIntell*. 2021;106:104481–104481. Available from: https://doi.org/10.1016/J.ENGAPPAI.2021.104481.
9) Wang W, Lu L, Wei W. A Novel Supervised Filter Feature Selection Method Based on Gaussian Probability Density for Fault Diagnosis of Permanent Magnet DC Motors. *Sensors*. 2019;22(19):7121–7121. Available from: https://doi.org/10.3390/s22197121.
10) Khan A, Nisha SS. Efficient hybrid optimization based feature selection and classification on high dimensional dataset. *Multimed Tools Appl*. 2023;83(20):58689–58727. Available from: https://doi.org/10.1007/s11042-023-17724-5.

11) Li K, Wang F, Yang L, Liu R. Deep feature screening: Feature selection for ultra high-dimensional data via deep neural networks. *Neurocomputing*. 2023;538:26186–26186. Available from: https://doi.org/10.1016/j.neucom.2023.03.047.
12) Wu H, Chen Y, Zhu W, Cai Z, Heidari AA, Chen H. Feature selection in high-dimensional data: an enhanced RIME optimization with information entropy pruning and DBSCAN clustering. *International Journal of Machine Learning and Cybernetics*. 2024;p. 1–44. Available from: https://doi.org/10.1007/s13042-024-02143-1.
13) Yang Y, Yin P, Luo Z, Gu W, Chen R, Wu Q. Informative feature clustering and selection for gene expression data. *IEEE Access*. 2019;7:169174–169184. Available from: https://doi.org/10.1109/ACCESS.2019.2952548.
14) Haq AU, Zhang D, Peng H, Rahman SU. Combining Multiple Feature-Ranking Techniques and Clustering of Variables for Feature Selection. *IEEE Access*. 2019;7:151482–151492. Available from: https://doi.org/10.1109/ACCESS.2019.2947701.
15) Hashemi A, Dowlatshahi MB. MLCR: A Fast Multi-label Feature Selection Method Based on K-means and L2-norm. In: and others, editor. 2020 25th International Computer Conference Computer Society of Iran, CSICC 2020. 2020. Available from: https://doi.org/10.1109/CSICC49403.2020.9050104.
16) Chavent M, Genuer R, Saracco J. Combining clustering of variables and feature selection using random forests. *Communications in Statistics-Simulation and Computation*. 2021;50(2):426–445. Available from: https://doi.org/10.1080/03610918.2018.1563145.
17) Sun L, Zhang J, Ding W, Xu J. Feature reduction for imbalanced data classification using similarity-based feature clustering with adaptive weighted K-nearest neighbors. *Information Sciences*. 2022;593:591–613. Available from: https://doi.org/10.1016/j.ins.2022.02.004.
18) Zhao S, Wang M, Ma S, Cui Q. A feature selection method via relevant-redundant weight. *Expert Systems with Applications*. 2022;207:117923–117923. Available from: https://doi.org/10.1016/J.ESWA.2022.117923.
19) Saxena A, Wang J, Sintunavarat W. An Empirical Study on Initializing Centroid in K-Means Clustering for Feature Selection. *International Journal of Software Science and Computational Intelligence (IJSSCI)*. 2021;13(1):1–16. Available from: https://doi.org/10.4018/IJSSCI.2021010101.
20) Zhou H, Wang X, Zhu R. Feature selection based on mutual information with correlation coefficient. *Applied Intelligence*. 2022;52(5):5457–5474. Available from: https://doi.org/10.1007/s10489-021-02524-x.
21) Wang Q. Support Vector Machine Algorithm in Machine Learning. In: and others, editor. 2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). IEEE. 2022. Available from: https://doi.org/10.1109/ICAICA54878.2022.9844516.
22) Schonlau M, Zou RY. The random forest algorithm for statistical learning. *The Stata Journal: Promoting communications on statistics and Stata*. 2020;20(1):3–29. Available from: https://dx.doi.org/10.1177/1536867x20909688.
23) Vinayakumar G. A Comparison of KNN Algorithm and MNL Model for Mode Choice Modelling. *European Transport/Trasporti Europei*. 2023;92(92):1–14. Available from: https://dx.doi.org/10.48295/et.2023.92.3.
24) Arizona State University. 2023. Available from: http://www.asu.edu.
25) UCI machine learning repository. 2023. Available from: http://archive.ics.uci.edu.