Advances in Artificial Intelligence and Machine Learning; Research 2 (4) 500-515 Received 8-10-22; Accepted 31-10-22; Published 18-11-22

A Novel Unsupervised Feature Selection Approach Using Genetic Algorithm on Partitioned Data

Amit Saxena
Department of Computer Science and IT,
Guru Ghashidash University, Bilashpur,
India.

Deepesh Chugh
Department of Computer Engineering,
Netaji Subhas Institute of Technology, Delhi,
India.

Himanshu Mittal himanshu.mittal224@gmail.com
Department of AIDS,
Indira Gandhi Delhi Technical University for Women, Delhi, India
India

Mohammad Sajid
Department of Computer Science,
Aligarh Muslim University, Aligarh,
India.

Ritu Chauhan
Centre for computational Biology and Bioinformatics,
Amity University, Noida,
India.

Eiad Yafi
Centre for computational Biology and Bioinformatics,
Amity University, Noida,
India.

Jian Cao
Department of Computer Science and Engineering,
Shanghai Jiaotong University, Shang-hai,
China.

Mukesh Prasad Mukesh.Prasad@uts.edu.au
School of Computer Science, FEIT,
University of Technology Sydney, Sydney,
Australia.

Corresponding Author: Mukesh Prasad

Citation: Amit Saxena, et al. A Novel Unsupervised Feature Selection Approach Using Genetic Algorithm on Partitioned Data. Advances in Artificial Intelligence and Machine Learning. 2022;2(4):34.

Abstract

A novel feature selection approach is presented in this paper. Sammon's Stress Function transforms the high dimension data to a lower dimension data set. A data set is divided into small partitions. The features are assigned randomly to these partitions. Using GA with Sammon Error as fitness value, a small, desired number of features are selected from every partition. The combination of the reduced subsets of the features from these partitions is again divided into small partitions. After a certain number of iterating the process, a desired small number of features is obtained. For experimental validation, the proposed method has been tested on 11 standard datasets with three classifiers namely, Decision Tree, MLP and KNN. The classification accuracies obtained by the proposed method is highest on most of the considered datasets against the results reported in literature. Moreover, the proposed method selects comparatively less number of features in comparison to considered methods. The optimistic results obtained from the proposed method justify its strength.

Keywords: Feature Extraction, Unsupervised Feature Selection, Genetic Algorithm, Classification Techniques.

1. INTRODUCTION

With the introduction of sophisticated electronic gadgets, there has been a tremendous growth in capturing vivid features of a data. This has resulted in exponential increase in data size. However, it is seen that some features are redundant and do not contribute to the inference of a learnable model. In fact, such features degrade the performance of a learning model. For example, microarray-datasets like, leukemia dataset, are large datasets which consist of thousands of features. However, a significant portion of these datasets is redundant only. Thus, the abundance of redundant and un-significant features leads to the problem of 'curse of dimensionality' which is one of the major problems of data mining [1] and pattern classification [2]. To mitigate this, it is necessary to identify features which are not only important but dominant over the entire dataset in representing the dataset and helps in achieving better learnable model. These features are termed as representative features of the dataset. In literature, there are number of feature selection approaches to retain representative features of a data. These methods differ from each other in terms of nature of the mapping function, how mapping function is learned and what optimization criterion is used [3-7]. Generally, the mapping can be linear or nonlinear and can be learned through either supervised or unsupervised methods. Feature selection advantages in saving the measurement cost as un-significant features get discarded. Additionally, the extracted features have better discriminating capability that leads to better performance sometimes [8]. Meanwhile, the selected features retain the original interpretation of data which is important to understand the underlying process that generates the data.

To perform feature selection, neural network, fuzzy logic, evolutionary algorithms, and statistical methods are some of the common approaches [6-16]. In general, each instance in a dataset is either labelled or unlabelled. Based on the use of label information, different approaches for feature selection are classified in two categories, namely supervised feature selection and unsupervised feature selection. In supervised feature selection, methods take

the advantage of such knowledge to perform selection of features. However, the association of a label with a data pattern requires extra cost and efforts. To minimize this, unsupervised feature selection methods work without label information. Some of the unsupervised feature selection (UFS) methods can be found in [17-19]. UFS methods benefit in saving the cost of finding labels for various patterns.

In literature, a substantial amount of work on unsupervised feature selection method can be found [20-22]. In [23], a method is proposed that partitions the considered feature set into distinct clusters in such a way that features in a cluster are highly similar, while features in different clusters are quite dissimilar. From each cluster, a feature is selected to form the feature subset. Dash and Liu [24] discussed the clustering approach for feature selection. Dy and Brodley [25] presented a wrapper framework for feature selection, clustering, and order identification concurrently. They also compared two different feature selection criteria, namely scatteredness that prefers those features which are far apart and the maximum-likelihood that prefers those features which lead to clusters that fit Gaussian models. In [26], several methods are discussed for feature selection based on maximum entropy and maximum likelihood criteria but the proposed strategy for feature selection depends on the method used to estimate uni-variate data. Pal et al. [27] proposed an unsupervised neuro-fuzzy feature ranking method in which a criterion is used to measure the similarity between two patterns in the original feature space and in the transformed feature space. The transformed feature space is obtained by multiplying each feature with a coefficient '*w*'. This coefficient is learned through a feed-forward neural network and features are ranked according to the learned weights, where higher value indicates higher importance and hence higher rank. Using this rank, the required number of features is selected. In [28], a new correlation-based approach for feature selection (CFS) is presented which uses features predictive performances and inter-correlations to guide its search for a good subset of features. Experiments on discrete and continuous class datasets reveal that CFS can drastically reduce the dimensionality of datasets while maintaining or improving the performance of learning algorithms.

Heydorn [29] gives a definition of redundancy between two random variables X and Y which is used to define a test of redundancy also. This test is used to eliminate redundant features without degrading performance of classifiers. Features that are linearly dependent on other features do not contribute toward pattern classification. In order to detect the linearly dependent features, a measure of linear dependence is proposed in [30-33]. This measure is used as an aid to feature selection which is demonstrated by employing the Speaker verification experiment. Devaney and Ram [32] proposed an unsupervised feature-ranking scheme by formulating a measure, termed as category utility, to evaluate the goodness of a feature subset. In each step, one feature is added to the existing feature set and the COBWEB [34] is run using this feature subset. The value of category utility for the created partition at the first level is computed and the feature subset yielding the highest score is retained. The iteration is continued until there is no significant improvement in the category utility value. Talavera proposed another unsupervised method of feature ranking for categorical data [35] using Fisher's feature dependency measure [34]. The algorithm exploits the assumption that in the absence of the class labels, features exhibiting higher dependencies with other features are going to be more relevant from clustering point of view. Let's assume that there is a dataset which consists of *p* features (A1, A2, …., Ap) and each feature (*Ai*) can take on different values *xij*. For a certain value of *xij* of feature *Ai* and a given partition (C1, C2,, …,

Cp), clusters having high value of P(Ck |Ai = xij) are termed as “distinct” clusters, whereas clusters having high value of P(Ai = xij|Ck) are termed as “cohesive” clusters. Based on this definition of distinctness and cohesiveness, Talavera [35] formulated an expression for the relevance of a feature for capturing the dependency among features and ranked the features. Similarly, Pena et al. [36] proposed a feature selection method in connection with clustering using Conditional Gaussian Networks (CGN). Before learning the network, the training data is pre-processed by selecting only relevant features for learning CGNs. By assuming that features exhibiting low correlation with other features are considered irrelevant, features yielding higher relevance measure than the considered relevance threshold are selected.

An entropy-based unsupervised feature ranking scheme is proposed which is applicable on both categorical and numerical data [37]. In this, the proposed method uses a sequential backward selection scheme, which iteratively rejects the least important feature based on the entropy measure. One very popular clustering algorithm is k-means [38] which performs clustering of data instances based on distances from the considered k centroids. However, there are some issues with k-means like, predefining the values of k and local minima if wrong cluster centroids are selected. Some improvements in k-means are suggested in [39]. Density-based clustering considers density of point as a measure to cluster data instances and predict the classes based on number of data instances lying in a cluster. A comprehensive review of clustering methods can be found in [40-43]. In UFS, the knowledge about the class label for data instances is not available. In literature, some measures are used to evaluate the strength of a feature in the dataset to perform UFS. Correlation-based measure finds the relation between two features by computing the Pearson’s coefficient of correlation between them [44]. The bigger value of the coefficient indicates higher dependency between two features [45-49]. Another feature evaluation measure is computing the Laplace Score [50] which is measured for each feature and ranked accordingly. Variance is also used for measuring the strength of a feature.

In this paper, an efficient and novel unsupervised algorithm is introduced to predict better classification accuracy of unknown patterns. The major contribution of the present work is three fold: (i) A new unsupervised feature selection method is presented, (ii) GA and Sammon error function are used to obtain optimal number of feature subset of the datasets, and (iii) For experimental analysis, the proposed method is compared against the eight state-of-the-art feature selection method in terms of classification accuracy and number of reduced features.

The rest of the paper is organized as follows. Section 2 discusses about the feature selection methods. Section 3 describes the proposed feature selection method. The experimental setup and obtained results are discussed in Section 4. Lastly, the paper is concluded in Section 5 along with future scope.

2. PRELIMINARIES

2.1 Feature Selection

The problem of feature selection can be formulated as follows: Given a dataset($X \subset R^p$) i.e., each xi$\in$ X has p features, a subset of q features needs to be selected that leads to the smallest (or highest) value with respect to some criterion. Let FC be the given set of features and FS be the set of selected features of cardinality m, i.e. FS$\subseteq$FC. Let the feature selection criterion for dataset (X) be represented as J(F, X) where lower value of J(.) indicates a better selection. When the training instances are labelled, the label information can be used, but in case of unlabeled data, this cannot be done. According to Kohavi and John [31], feature selection schemes can be broadly classified into two categories, wrapper models and filter models. Wrapper models use feedback principle to evaluate a subset of selected features by using the classifier itself to measure the goodness of the selected feature-subset on the either test data or training data [31]. While, filter models use some intrinsic property of the data set, which is assumed to affect the performance of the learning model. Therefore, filter models are generally characterized as unsupervised feature selection methods.

2.2 Genetic Algorithm

Genetic algorithm (GA), proposed by John Holland in 1975 [51] and later extended by Goldberg [52], works on evolutionary theory of natural genetics and is based on the principle of 'survival of the fittest' [53-54]. The algorithm starts by initializing a population of potential solutions encoded as chromosomes. Each solution has a fitness value which is used to determine the fittest parents for reproduction of next population. A typical GA executes three sequential phases: (i) selection of parents, (ii) crossover operation, and (iii) mutation operation. To perform selection of parents, few of the popular methods are roulette-wheel, rank-based, random selection, and tournament selection. In this paper, tournament selection is used in the proposed method. After crossover and mutation operations, the updated population and initial population are combined to form the next population by selecting the fittest chromosome. This completes one generation of GA which are repeated for a large number until some specific criterion is met or the solution converges to some optimal value.

2.3 K-Nearest Neighbor

Hodges and Fix [55] introduced k-nearest neighbor (K-NN) to perform classification task and is considered as one of the simplest methods of machine learning as the samples are classified by a majority of the vote of its nearest neighbor. [17]. Basically, it follows the approach of instance-based learning or lazy learning, where all computation is deferred until classification and the function is only approximated locally [56]. Generally, KNN is mostly used as a classifier for testing or evaluation purpose and is quite popular for data classification in the real world.

2.4 Multi-layer Perceptron

Multi-layer perceptron (MLP) is a popular method of machine learning which is based on artificial neural networks (ANN) [57- 58] and back propagation algorithm. MLP have a set of inputs, a set of hidden layers and an output layer. The inputs and targets are given in advance for training. After the training of MLP, an unknown set of inputs is supplied to trained MLP and the predicted outputs are used to test the performance of the trained MLP[59].

2.5 Decision Tree

A decision tree (DT) [60] uses a training set of patterns. Each pattern consists of a set of features and a class label. A decision tree may have no or few internal nodes and one or more leaf nodes. All internal nodes have at least two children. All internal nodes test the value of an expression for the features. Each leaf node corresponds to a class label.

3. PROPOSED UNSUPERVISED FEATURE SELECTION METHOD

This paper presents a new unsupervised feature selection method, unsupervised feature selection genetic algorithm (USFSGA), by incorporating genetic algorithm to obtain the optimal subset of features. In the proposed method, the features of the considered dataset are randomly arranged. For illustration, suppose the sequence of features in a dataset is represented as 1, 2, ..., F. This sequence is re-arranged randomly to obtain a new feature sequence, e.g., 4, 1, 80, 34, …, F. This set of features is further divided into (F/D) number of partitions in such a way that each partition contains a maximum of D features. This makes a fair and unbiased distribution of features in all partitions. In case of $F < D$, only one partition is formed that consists of F features. In case of excessive features, features will be assigned to different partitions in sequential fashion. For example, if there are 63 features in a dataset and D is kept as 30, then there will be two partitions, i.e., partition 1 (P1) and partition 2 (P2). First, P1 and P2 will be assigned 30 features each. Then out of the remaining three features, one feature will be allocated to P1 and then to P2 which makes 31 number of features in P1 and P2 in each. For the remaining feature, the excess feature will be assigned to partition P1. Therefore, P1 will have in total 32 and P2 have 31 features. After performing feature partitions, genetic algorithms (GA) are applied on each partition. The chromosome size is equal to the number of features in the corresponding partition. Each chromosome is populated with binary values randomly where zero value means the corresponding feature of the partition is not considered while one indicates that the corresponding feature is considered. Further, Sammon error is computed as fitness function for each chromosome. To update the chromosomes, the proposed method applies tournament selection, crossover operation and mutation operation. Thereafter, elitism is employed to generate a new population. This completes one generation of GA for a partition which is applied for a substantial number of generations to get better population. At the end, the fittest one is the first chromosome in each partition and the value of one in fittest chromosome indicates the features which will be retained for further processing, while remaining features (indicated by zeros) will be dropped. From all the partitions, the retained features are combined. This subset of features is then used for the

classification process [61-64]. The procedure of the proposed method is described in section 3.1.

3.1 Procedure of the Proposed Method

Input: A dataset with number of features as F and number of instances as I.

Output: The feature set with highest classification accuracy.

1. Generate a random but unique numbering to arrange the features of the dataset.
2. Decide a partition size (D). Create N=F/D partitions with F features as P1,P2,…,PN, where N is a natural number and N=1, when F<D. In case ND<F<(N+1)D, the excessive features are re-assigned to P1, P2, …, Pi partitions in sequential fashion. If P_n is the number of features in the P_nth partition, then P_1 + P_2 +…+ P_N =F.
3. On each partition Pn (n=1,2,…N), GA is applied for a sufficiently large number of generations:
   (a) Create a population with chromosome size as P_n and each chromosome is initialize with binary values randomly, where the presence of 1 at mth position indicates that the mth feature of the dataset is retained.
   (b) Compute the fitness of each chromosome for the partition Pn using Eq (1).
   (c) Apply tournament selection, crossover operation and mutation operation to update the current population.
   (d) Apply elitism to generate the population for next generation.
4. At the end of final generation, the fittest chromosome will be at the first position.
5. Combine the feature subset from all the partitions to use them on the classification problem and calculate the classification accuracy.

3.2 Unsupervised Feature Selection (USFS) Criterion

The proposed method selects features in an unsupervised manner. In this paper, Sammon's Stress Function [16] is used as an evaluation criterion for performing feature selection. Let X ={x_k |x_k =(x_{k1}. x_{k2}….. x_{kp})^T. k=1.2…..n} be the set of *n* vectors in the original space and Y ={y_k |y_k =(y_{k1}. y_{k2}….. y_{kq})^T. k=1.2…..n} be the unknown data vectors in the reduced space. Let d^*_{ij} = d(xi. xj), x_i, x_j Є X and d_{ij} = d(y_i. y_j). y_i, y_j Є Y. where d(x_i. x_j) be the Euclidean distance between x_i and x_j. The Sammon error (E) is given by Eq. (1) which is minimized using gradient descent algorithm.

$$E = \frac{1}{\sum_{i<j} d^*_{ij}} \sum_{i<j} \frac{(d^*_{ij} - d_{ij})^2}{d^*_{ij}} \quad (1)$$

Since Eq. (1) preserves the inter-point distances, it is used as an evaluation criterion for feature selection. Clearly, E would be zero when all features are selected. Therefore, the

proposed method selects a subset of q features so that E is minimized. To achieve the same, genetic algorithm is employed in the proposed method.

4. EXPERIMENT RESULTS AND DISCUSSION

4.1 Dataset

To evaluate the performance of the proposed method, the proposed method is implemented using MATLAB on a system with Intel i5 processor, 3.5 GHz, and 16 GB of RAM . For extensive experimental analysis, eleven datasets are considered which are taken from a publicly available UCI repository [65].TABLE 1 shows the list of all the considered datasets in the paper. From TABLE 1, it can be observed that some datasets are highly dimensional with the number of feature sets as in the hundreds or thousands.

In the considered datasets, one dataset is a synthetic dataset. The role of synthetic dataset is to see the dummy performance of the method. The parameter setting of the considered method along with different abbreviations are shown in TABLE 2. As considered datasets have variations in the number of instances, features and classes, the population size in GA has been customized accordingly. The classification accuracy (CA) on the considered datasets is computed by using Decision Tree, KNN, and MLP which are tabulated in TABLE 3, TABLE 4, and TABLE 5 respectively. For comparison, eight state-of-the-art feature selection methods have been considered whose results are taken from [20-22]. In each table, there are 12 columns in which first column corresponds to the dataset name, columns 2 to10 (M1 to M9) present the classification accuracies attained by considered methods, column 11 indicates the minimum percentage of selected features by any of the compared methods (M1 to M8), and column 12 contains the percentage of selected features by the proposed method. The value of k in KNN is set to 1, the size of MLP is single hidden layer with 10 neurons, and fit tree is used with default settings for decision tree.

Table 1: Datasets descriptions.

Datasets	**# Samples**	**# Features**	**# Classes**	**Type**
Leukemia1	72	7129	2	Continuous
Colon	62	2000	2	Discrete
Madelon	2600	500	2	Continuous
Yale	165	1024	15	Continuous
WDBC	569	30	2	Continuous
SPECTF	88	44	2	Discrete
Isolet	1559	617	26	Continuous
Lung-Discrete	73	325	7	Discrete
warpAR10P	130	2400	10	Continuous
warpPIE10P	165	2420	10	Continuous
Sonar	208	60	2	Continuous

Table 2: Comparison of Classification Accuracies (in %) and size of reduced features (in %) by various methods for different datasets using Decision Tree as classifier.

Datasets	LSFS M1	MCFS M2	LLCFS M3	UDFS M4	NDFS M5	SPFS M6	EUFSFC M7	Base M8	USFSGA M9	% of Selected F[20]	% of Selected F by M9
Leukemia1	89.17	86.94	91.39	85.83	76.94	71.67	86.11	81.67	**92.85**	3.03	**1.69**
Colon	73.23	**80.65**	76.13	60.65	73.23	76.04	77.10	74.19	77.50	4.45	7.85
Madelon	72.17	50.35	73.16	51.05	52.48	49.73	**78.40**	74.95	66.65	1.20	10
Yale	48.36	**54.91**	44.85	54.30	43.88	48.85	56.12	48.85	49.69	14.74	**10.15**
WDBC	91.42	89.21	92.09	89.77	89.49	93.08	93.29	92.41	**93.53**	40.00	**33.33**
SPECTF	71.54	67.94	74.76	71.39	74.08	75.28	74.41	72.96	**89.18**	29.54	36.36
Isolet	71.38	67.47	60.48	66.67	65.48	54.38	**72.07**	71.45	66.98	23.98	**9.72**
Lung-Small	54.52	60.82	51.78	51.65	52.33	51.23	**62.74**	59.18	59.28	21.85	24.61
warpAR10P	58.62	**75.45**	51.54	64.62	61.32	57.78	76.38	71.85	57.69	20.42	**2.4**
warpPIE10P	72.04	73.48	77.52	77.41	75.58	80.25	**79.24**	74.95	76.42	3.10	**2.02**
Sonar	66.35	71.25	75.48	73.08	70.38	64.33	71.92	69.70	**75.52**	46.67	**30**

Table 3: Comparison of Classification Accuracies (in %) and size of reduced features (in %) by various methods for different datasets using kNN as classifier.

Datasets	LSFS M1	MCFS M2	LLCFS M3	UDFS M4	NDFS M5	SPFS M6	EUFSFC M7	Base M8	USFSGA M9	% of Selected F[20]	% of Selected F by M9
Leukemia1	94.44	95.56	**96.94**	92.78	76.11	81.39	92.22	89.47	94.28	3.03	**1.69**
Colon	61.61	76.45	**79.35**	78.42	71.61	74.19	78.10	77.74	78.33	4.45	7.85
Madelon	79.49	50.45	81.28	50.38	54.47	50.06	**83.24**	72.10	62.38	1.20	10
Yale	41.70	52.45	47.88	49.78	57.09	36.85	57.70	57.45	**60.90**	14.74	**10.15**
WDBC	93.39	92.27	93.01	91.28	91.14	91.85	**94.12**	93.29	91.76	40.00	**33.33**
SPECTF	79.85	78.35	77.15	73.56	74.51	79.78	74.76	73.31	**88.64**	29.54	36.36
Isolet	72.66	63.58	68.07	74.82	70.41	66.81	79.64	**80.24**	72.69	23.98	**9.72**
Lung-Small	68.77	84.93	80.82	82.47	77.53	71.78	82.38	81.64	**86.42**	21.85	24.61
warpAR10P	55.85	**60.96**	44.62	46.92	49.86	54.31	56.41	50.31	49.61	20.42	**2.4**
warpPIE10P	72.76	97.14	89.52	95.57	93.46	91.33	95.29	92.29	**99.04**	3.10	**2.02**
Sonar	73.27	78.46	76.92	78.08	77.42	72.69	**83.37**	81.54	81.95	46.67	**30**

Table 4: Comparison of Classification Accuracies (in %) and size of reduced features (in %) by various methods for different datasets using MLP as classifier.

Datasets	LSFS M1	MCFS M2	LLCFS M3	UDFS M4	NDFS M5	SPFS M6	EUFSFC M7	Base M8	USFSGA M9	A% of Selected F[20]	% of Selected F by M9
Leukemia1	**96.08**	95.12	95.68	85.80	87.92	82.71	88.09	69.52	92.14	3.03	**1.69**
Colon	76.28	76.06	76.95	75.48	73.11	76.77	77.94	77.74	**82.50**	4.45	7.85
Madelon	81.69	50.20	75.92	53.17	50.93	52.23	**84.43**	57.14	53.34	1.20	10
Yale	52.35	71.86	65.82	72.36	72.31	63.47	**77.76**	48.56	54.84	14.74	**10.15**
WDBC	87.89	90.44	89.97	92.52	75.96	86.67	96.90	91.19	**96.90**	40.00	**33.33**
SPECTF	76.36	77.21	76.84	77.41	76.95	76.34	79.43	72.56	**92.70**	29.54	36.36
Isolet	65.06	66.26	75.13	68.85	**77.25**	64.26	76.20	77.24	71;41	23.98	**9.72**
Lung-Small	73.55	**84.41**	80.82	82.71	78.65	75.27	83.25	82.71	78.57	21.85	24.61
warpAR10P	85.14	96.00	92.50	97.92	96.98	94.60	**98.26**	67.89	67.30	20.42	**2.4**
warpPIE10P	87.45	95.25	90.78	94.14	96.34	92.45	93.58	92.91	**98.80**	3.10	**2.02**
Sonar	57.45	67.28	61.56	75.92	71.48	62.48	68.86	65.05	**80.48**	46.67	**30**

4.2 Discussions

In TABLE 3, the proposed USFSGA method produces highest CA with only 1.69 % of features for Leukemia dataset. For the colon dataset, M2 gives highest CA of 80.65% while the proposed method gives 77.50% of CA. For Madelon, M4 produces highest 78.40% CA. For Yale, highest CA of 56.12% is attained by M7 in comparison to 49.69% of CA by proposed method. However, the proposed method takes only 10.15% of the features. In WDBC dataset, the proposed method surpasses in CA as well as feature size. Similarly, in SPECTF dataset, the proposed method has better CA in comparison to other methods. On Isolet dataset, M7 has highest CA of 72.07% compared to CA of 66.98% by the proposed method. However, the proposed method uses only 9.72% of the features against 23.98%. The Lung Small Discrete dataset has highest CA of 62.74% for M7. In warpAR10P and warpPIE10P, M7 and M6 produce highest CA respectively. However, the proposed method uses only 2.4% and 2.02% of features for these two datasets respectively. For Sonar dataset, the proposed method yields highest CA and fewer number of features.

In TABLE 4, for Leukemia dataset, M3 method produces highest CA of 96.94% in comparison to the proposed method which achieved CA of 94.28% but with only 1.69 % of the features. For colon dataset, M3 gives highest CA 80.65% and proposed method gives 78.33% of CA. On Madelon, M4 produces highest 83.24% CA. For Yale, highest CA of 70.90% is achieved by the proposed method with 10.15% features only. In WDBC, M1 produces 93.39% of CA while the proposed method results in 91.76% of CA with 33.33% feature size. In SPECTF dataset, the proposed method has better CA (88.64%) in comparison to other methods with least number of features. In Isolet, M7 has highest CA (79.64%) compared to

72.69% of the proposed method. However, the proposed method uses only 9.72% against 23.98%. The Lung Small Discrete dataset has highest CA of 86.42% by using the proposed method. In warpAR10P, M2has highest CA (60.96%) while the proposed method uses only 2.4% features. On warpPIE10P, the proposed method outperforms others with CA 99.04 and feature size of 2.02% only. For Sonar dataset, M7 gives highest CA(83.37%)in comparison to 81.95% by the proposed method but later method uses only 30%of the considered features.

In TABLE 5, M1 method produces highest CA 96.08%for Leukemia dataset. However, the proposed method attained CA of 92.14% on the same dataset with only 1.69 % of features. For the colon dataset, the proposed method gives highest CA (82.50%). On Madelon and Yale, M7 produces highest CA. In WDBC, the proposed method produces highest CA of 96.90% with 33.33% feature size. In SPECTF dataset, the proposed method has better CA (92.70%) compared to other methods. On Isolet, M8 has highest CA of 77.24% but the proposed method achieved CA of 71.41 % by using only 9.72% against 23.98%. The Lung Small Discrete dataset has highest CA 84.41 against 78.57% using the proposed method. In warpAR10P, M7 has highest CA (98.26 %), however the proposed method uses only 2.4% features. On warpPIE10P, the proposed method outperforms others with CA 98.80% and feature size as 2.02% only. For Sonar dataset, the proposed method gives highest CA of 80.48% using only 30% features. Therefore, following observations can be drawn for the proposed method according to TABLES 3,4, and 5.

For classification using Decision Tree, the proposed method (M9) produces highest CA on 04 out of 11 datasets, whereas it uses less number of features to achieve competitive CAs against other methods on 07 datasets.

With kNN classifier, the proposed method (M9) produces highest CA on 04 out of 11 datasets, whereas on 08 datasets, M9 uses less number of features to achieve competitive CAs against other methods.With MLP classifier, the proposed method (M9) produces highest CA on 05datasets, whereas in 07 datasets, M9 uses minimum number of features to achieve competitive CAs against other methods.

4.3 Time Complexity of UGAF

The time complexity of the proposed USFSGA method can be formulated as follows.

For genetic algorithm: Suppose that there are P partitions, each partition has approximately Q features, and R is the population size. The tournament selection has time complexity of $O(PQR)$. If the number of instances in the dataset be I, then time complexity of the Sammon error will be $O(I*(I-1)/2)$, thus time taken for one generation will be $O(PQR+I*I)$. For G number of generations, total time complexity will be $O(G(PQR+I*I))$. For kNN classifier: $O(KIS)$ where S is the size of subset of features finally obtained For MLP: $O(N)$ where N is the Number of Neuron. For Decision Tree: $O(SI\log_2 S)$for a CART based DT. Total complexity=$O(PQR+I*I+KIS+N+Silo_2S)$ For E number of runs, the overall complexity is $O(E(PQR+I*I+KIS+N+Silo_2S))$

Table 5: Time Complexity of the Proposed Method.

	Time Complexity
Genetic Algorithm	*O(G(PQR+I*I))*
kNN classifier	*O(KIS)*
MLP Classifier	*O(N)*
Decision Tree	$O(SIlog_2S)$
Total	**$O(PQR+I*I+KIS+N+Silo_2S)$**

5. CONCLUSION

In this paper, a novel efficient USFSGA method is presented to determine significant features of a dataset. In general, Sammon error is a good metric to transform a high dimension dataset to a lower one but if the size of feature set is too high then Sammon error cannot distinct and cannot be taken as reliable to decide the strength of features. The work carried out in this paper focuses on dividing the large set of features into small partitions. With the smaller partitions, Sammon Error can be better distinctive for subsets of features. GA is applied on each partition and an iterative approach computes a substantially reduced feature set. For experimental analysis, three classifiers on eleven datasets are considered. The proposed method attained best CA on many datasets. Although some compared methods achieve higher CAs than the proposed method, yet the proposed method is able to report competitive performance with less number of features in comparison to the considered methods.

In future, the proposed method can be tested on datasets with thousands or millions of features. Moreover, the appropriate parameter settings of GA, kNN, MLP and DT, can be explored which can further improve the performance of the proposed method. Additionally, the proposed method can be extended on real-world data classification problem.

References

[1] Kamber M, Han J. Data mining: concepts and techniques. 2nd ed. CA: Morgan Kaufmann Publishers. San Francisco. 2006.

[2] Duda RO, Hart PE, Stork DG. 'Pattern Classification,' Wiley publications. 2001.

[3] Sammon, Jr. JW. A Nonlinear Mapping for Data Structure Analysis. IEEE Trans Comput. 1969;C-18:401-409.

[4] Schachter B. A Nonlinear Mapping Algorithm for Large Databases. Comput Graph Sammon AImage Process. 1978;7:271-278.

[5] Pykett CE. Improving the Efficiency of Sammon's Nonlinear Mapping by Using Clustering Archetypes. Electron Lett. 1978;14:799-800.

[6] Pal NR. Soft Computing for Feature Analysis. Fuzzy Sets Syst. 1999;103:201-221.

[7] Muni DP, Das Pal NR, J. A Novel Approach for Designing Classifiers Using Genetic Programming. IEEE Trans Evol Comput. 2004;8:183-196.

[8] Saxena NRP, Vora M. 'Evolutionary Methods for Unsupervised Feature Selection Using Sammon's Stress Function'. Fuzzy Inf Eng. 2010;2:229-247.

[9] Pal NR, Eluri VK, Mandal GK. Fuzzy Logic Approaches to Structure Preserving Dimensionality Reduction. IEEE Trans Fuzzy Syst. 2002;10:277-286.

[10] Pal SK, De RK, Basak J. Unsupervised Feature Evaluation: A Neuro-Fuzzy Approach. IEEE Trans Neural Netw. 2000;11:366-76.

[11] Das SK. Feature Selector With a Linear Dependence Measure. IEEE Trans Comput. 2000:1106-1109.

[12] Siedlecki W, Sklansky J. Note on Genetic Algorithms for Large Scale Feature Selection. Pattern Recognit Lett. 1989;10:335-347.

[13] Casillas J, Cordón O, Del Jesus MJ, Herrera F. Genetic Feature Selection in a Fuzzy Rule Based Classification System Learning Process for High Dimensional Problems. Inf Sci. 2001;136:135-157.

[14] Bishop CM. Neural networks for pattern recognition. Oxford: Clarendon Press. 1995.

[15] Chakraborty D, Pal NR. Designing Rule-Based Classifiers With On-Line Feature Selection: A Neuro-Fuzzy Approach. In: Proceedings of the advances in soft computing – AFSS. Springer. 2002: 251-259.

[16] Pal NR, kant CK. A Connectionist System for Feature Selection. Neural Parallel Sci Comput. 1997;5:359-382.

[17] Saxena AK, Dubey VK, Wang J. Hybrid Feature Selection Methods for High Dimensional Multi-Class Datasets. Int J Data Min Modell Manag. 2017;9:315-339.

[18] Dubey VK, Saxena AK. A Cosine Similarity- Mutual Information Approach for Feature Selection on High-Dimensional Dataset. J Inf Technol Research. 2017;10:15-28.

[19] Lin CT, Prasad M, Saxena A. An Improved Polynomial Neural Network Classifier Using Real-Coded Genetic Algorithm. IEEE Trans Syst Man Cybern Syst. 2015;45:1389-1401.

[20] Yan X, Nazmi S, Erol B, Homaifar A, Gebru B, et al. An Efficient Unsupervised Feature Selection Procedure Through Feature Clustering. Pattern Recognition Letters. 2020;131:277-284.

[21] Yan X, Homaifar A, Awogbami G, Girma A. Unsupervised Feature Selection Through Fitness Proportionate Sharing Clustering. In: IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE. 2018; 2018:1355-1360.

[22] Yan X, Homaifar A, Nazmi S, Razeghi-Jahromi M. A Novel Clustering Algorithm Based on Fitness Proportionate Sharing. In: IEEE International Conference on Systems, Man, and Cybernetics(SMC). 2017:1960-1965.

[23] Mitra P, Murthy CA, Pal SK. Unsupervised Feature Selection Using Feature Similarity. IEEE Trans Pattern Anal Mach Intell. 2002;24:301-312.

[24] Dash M, Liu H. Feature Selection for Clustering. In: Proceedings of the Asia Pacific conference on knowledge discovery and data mining; 2000: 110-121.

[25] Dy JG, Brodley CE. Feature Subset Selection and Order Identification for Unsupervised Learning. In: Proceedings of the 17th international conference on machine learning. 2000.

[26] Basu S, Micchell CA, Oslen P. Maximum Entropy and Maximum Likelihood Criteria for Feature Selection and Multivariate Data. In: Proceedings of the IEEE international symposium on circuits and systems; 2000:267-270.

[27] Pal SK, De RK, Basak J. Unsupervised Feature Evaluation: A Neuro-Fuzzy Approach. IEEE Trans Neural Netw. 2000;11:366-376

[28] Hall MA. Correlation Based Feature Selection for Discrete and Numeric Class Machine Learning. In: Proceedings of the 17th international conference on machine learning; 2000.

[29] Heydorn RP. Redundancy in feature extraction. IEEE Trans Comput. 1971;C-20:1051-1054.

[30] Das SK. Feature Selection With a Linear Dependence Measure. IEEE Trans Comput. 1971;20:1106-1109.

[31] Kohavi R, John GH. Wrappers for Feature Subset Selection. Artif Intell. 1997;97:273-324.

[32] Devaney M, Ram A. Efficient Feature Selection in Conceptual Clustering. In: Proceedings of the machine learning Fourteenth international conference. Nashville. TN. 1997.

[33] Gluck MA, Corter JE. Information Uncertainty and the Utility of Categories. In: Proceedings of the Seventh Annual Conference of the Cognitive Science Society. Irvine, CA: Lawrence Erlbaum Associates; 1985: 283-287.

[34] Fisher DH. Knowledge Acquisition via Incremental Conceptual Clustering. Mach Learn. 1987;2:139-72.

[35] Talavera L. Dependency-Based Feature Selection for Clustering Symbolic Data. Intell Data Anal. 2000;4:19-28.

[36] Pena JM, Lozano JA, Larranaga P, Inza I. Dimensionality Reduction in Unsupervised Learning of Conditional Gaussian Networks. IEEE Trans Pattern Anal Mach Intell. 2001;23:590-603.

[37] Dash M, Liu H, Yao J. Dimensionality Reduction for Unsupervised Data. Proceedings of the 19th IEEE international conference on tools with AI. ICTAI. 1997.

[38] MacQueen JB. Some Methods for classification and Analysis of Multivariate Observations 5th Symposium on Mathematical Statistics and Probability, Berkeley. University of California Press,; 1967; 1: 281-297.

[39] Yanling Duan QL, Xia S. An Improved Initialization Center K-Means Clustering Algorithm Based on Distance and Density. AIP Conf Proc. 2018:040046-1 - 040046-7.

[40] Saxena A, Prasad M, Gupta A, Bharill N, Patel OP, Tiwari A et al. A Review of Clustering Techniques and Developments. Neurocomputing. 2017;267:664-681.

[41] http://www.leg.ufpr.br/ leonardo/artigos/tese$_m ahfoud.pdf DebK.Multi – ObjectiveOptimizationUsingEvolutionaryAlgorithms.Wiley$; 2014.

[43] Bandyopadhyay S, Bhadra T, Mitra P, Maulik U. Integration of Dense Subgraph Finding With Feature Clustering for Unsupervised Feature Selection. Pattern Recognit Lett. 2014;40:104-112.

[44] . Mandal M, Mukhopadhyay A. Unsupervised Non-redundant Feature Selection: A Graph-Theoretic Approach. In: Proceedings of the international conference on frontiers of intelligent computing: theory and applications (FICTA); 2013:373-380.

[45] Peng H, Long F, Ding C. Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Minredundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27:1226-1238.

[46] Saxena A, Wang J, Sintunavarat W. An Empirical Study on Initializing Centroid in K-Means Clustering for Feature Selection. Int J Softw Sci Comp Intell. 2021;13:1-16.

[47] Xu J, Tang B, He H, Man H. Semisupervised Feature Selection Based on Relevance and Redundancy Criteria. IEEE Trans Neural Netw Learn Syst. 2017;28:1974-1984.

[48] Sheikhpour R, Sarram MA, Gharaghani S, Chahooki MAZ. A Survey on Semi-supervised Feature Selection Methods. Pattern Recognit. 2017;64:141-158.

[49] Saxena A, Pare S, Meena MS, Gupta D, Gupta A, et al. A Two-Phase Approach for Semi-supervised Feature Selection. Algorithms. 2020;13:215-222.

[50] He X, Cai D, Niyogi P. 'Laplacian Score for Feature Selection,' in Proc. 18th Adv. Neural Inf Process. 2005:507-514.

[51] Goldberg D. Genetic Algorithms in Search Optimization and Machine Learning. Reading. MA: Addison-Wesley. 1989.

[52] Holland JH. Adaptation in Nature and Artificial Systems. Ann Arbor: University of Michigan Press. 1975.

[53] Eberhart RC, Shi Y. Particle Swarm Optimization- Developments, Applications and Resources. In: IEEE Congress on Evolutionary Computation. 2001;1:81-86.

[54] Dorigo M, Maniezzo V, Colorni A. Ant System: Optimization by a Colony of Cooperating Agents. IEEE Trans Syst Man Cybern B Cybern. 1996;26:29-41.

[55] Fix E, Hodges J. Dicriminatory analysis, Non Parametric Discrimination: consistency Properties, Technical Report 4. USA, School of aviation medicine Randolph field. TX; 1951.

[56] Altman NS. An Introduction to Kernel and Nearest-Neighbor Non Parametric Re Gression. Am Stat. 1992;46:175-185.

[57] Bishop CM. Neural Networks for Pattern Recognition. Oxford: Clarendon Press. 1995.

[58] Haykin SS. Neural networks: A comprehensive foundation. Prentice Hall. 2nd ed; 1999.

[59] Rumelhart DE, Hinton GE, Williams RJ. Learning Representations by Back-Propagating Errors. Nature. 1986;323:533-536.

[60] Murthy SK. Automatic Construction of Decision Trees From Data: A Multidisciplinary Survey. Data Min Knowl Discov. 1998;24:345-389.

[61] Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. Wadsworth Int Group. 1984.

[62] Fielding A. Binary Segmentation: The Automatic Interaction Detector and Related Techniques for Exploring Data Structure. The Analysis of Survey Data. John Wiley Sons,pp. 1977; 1: 221-257.

[63] Quinlan JR. Induction of Decision Trees. Mach Learn. 1986;1:81-106.

[64] Quinlan JR. C4.5: programs for machine learning. San Mateo, CA: Morgan Kaufmann Publishers; 1993.

[65] http://www.ics.uci.edu/$\sim$mlearn/MLRepository.html