Neural Comput & Applic (2017) 28:2915–2930
DOI 10.1007/s00521-017-2959-y

An approach for feature selection using local searching and global optimization techniques

Sadhana Tiwari[1] · Birmohan Singh[2] · Manpreet Kaur[3]

Received: 22 October 2016 / Accepted: 21 March 2017 / Published online: 31 March 2017

Abstract Classification problems such as gene expression array analysis, text processing of Internet document, combinatorial chemistry, software defect prediction and image retrieval involve tens or hundreds of thousands of features in the dataset. However, many of these features may be irrelevant and redundant, which only worsen the performance of the learning algorithms, and this may lead to the problem of overfitting. These superfluous features only degrade the accuracy and the computation time of a classification algorithm. So, the selection of relevant and nonredundant features is an important preprocessing step of any classification problem. Most of the global optimization techniques have the ability to converge to a solution quickly, but these begin with initializing a population randomly and the choice of initial population is an important step. In this paper, local searching algorithms have been used for generating a subset of relevant and nonredundant features; thereafter, a global optimization algorithm has been used so as to remove the limitations of global optimization algorithms, like lack of consistency in classification results and very high time complexity, to some extent. The computation time and classification accuracy are improved by using a feature set obtained from sequential backward selection and mutual information maximization algorithm which is fed to a global optimization technique (genetic algorithm, differential evolution or particle swarm optimization). In this proposed work, the computation time of these global optimization techniques has been reduced by using variance as stopping criteria. The proposed approach has been tested on publicly available Sonar, Wdbc and German datasets.

Keywords Sequential backward selection · Mutual information maximization · Optimization algorithms · Support vector machine

✉ Manpreet Kaur
aneja_mpk@sliet.ac.in

Sadhana Tiwari
sadhanatiwari3@gmail.com

Birmohan Singh
birmohansingh@sliet.ac.in

[1] Bharat Electronics Limited, Ghaziabad 201010, India

[2] Department of Computer Science and Engineering, SLIET, Longowal 148106, India

[3] Department of Electrical and Instrumentation Engineering, SLIET, Longowal 148106, India

1 Introduction

Feature selection is an important step to obtain the desired number of features from the original feature set. A feature set that contains relevant and nonredundant features purely contributes to the correct prediction of test data with the help of a classifier [1]. Theoretically, it is expected that increasing the size of feature vector provides more discriminating power. However, practically when the size of feature vector increases to an extent, the learning process of an algorithm slows down and it results in overfitting. This also compromises the model generalization [2].

Irrelevant features do not affect the target concept in any way, only the processing time of classification is increased. Redundant features might possibly add more noise than the useful information. Both irrelevant and redundant features interfere with useful ones, due to which most of the supervised learning algorithms fail to properly identify those features that are relevant to describe the target concept [3]. Feature selection provides a number of benefits

such as reducing training and utilization time, facilitating data visualization and data understanding, reducing measurement cost and storage requirement, and improving prediction performance [4].

The feature selection framework mainly consists of four parts: a *generation* procedure, which generates features subset, for which a search strategy is required that is used to determine the most promising feature subset candidates, and then, an *evaluation* strategy is used to determine the goodness of candidate feature subset in order to find the best feature subset out of it. Later, a *stopping criterion* needs to be decided and a *validation* procedure to check whether the subset is valid or not [5, 6] as shown in Fig. 1.

For subset generation, a number of methods have been proposed. Subset generation is a process of procreating the subsets of different features from a large set of original features [7]. If the number of features present in the original feature set is N, then the total number of competing candidate subsets to be generated is 2^N. This is an extremely large number even for medium values of N. Selecting optimal subset from such an enormous number of features is a big menace. In order to solve this problem, there exist various search strategies, namely *complete search strategy* (branch and bound [8], beam search [9] and best first search [10]), *heuristic search strategy* [sequential forward search (SFS), sequential backward search (SBS), plus L minus R selection (LRS), bidirectional feature selection (BDS), sequential floating forward selection (SFFS) and sequential floating backward selection (SFBS)] [11, 12] and *random search strategy* (simulated annealing (SA) [13], genetic algorithm (GA) [14, 15], differential evolution-based feature selection (DEFS) [16] and particle swarm optimization (PSO) [17]).

It has been observed from the literature that the search strategies used have certain limitations. Algorithms under the complete search strategy such as branch and bound suffer from the problem of having exponential complexity [1, 8]. Though exhaustive search is a powerful method because it has wide applicability and is known for its simplicity, these algorithms are unexpectedly slow [18]. In heuristic search strategies, algorithms like SFS and SBS are liable to possess the problem of getting trapped in local minima [19]. The algorithms under this category are known to be local optimization algorithms. Though the computational complexities of these algorithms are less, these local algorithms have the capability of locating the local optimum only. These sequential algorithms add or remove feature sequentially; hence, they have a tendency to become trapped in local minima due to the nesting effect [20, 21].

In random search strategies, global optimization techniques like GA, DEFS and PSO though avoid the problem of local minima, but some controlling parameters need to be taken care of, to produce better results [22]. Basically, the success of these global optimization algorithms depends on several factors. The parameter steering the crossover, mutation and survival of the chromosomes has to be carefully chosen so that solution space can be explored by the population and also to prevent the early convergence to homogeneous population occupying a local minimum. The choice of initial population is one of the most important steps in these global optimization techniques for feature selection [23].

In order to overcome the limitations of these strategies, different combinations of subsets generating algorithms have been explored in this work. To eliminate the problem of local minima from SBS to some extent, it can be combined with an optimization technique (GA, DEFS or PSO), because incorporating randomness into their search procedure will help it to escape from local minima. Also, combining SBS with these global optimization techniques will control the population initially, resulting in the generation of an optimal subset. The subset generated needs to be evaluated by some evaluation function, which is the second major step in feature selection.

2 Subset evaluation: correlation and information theoretic criterion

The feature subset generated by using any search strategy must be evaluated with some evaluation function in order to determine the quality of a feature subset. The feature evaluation methods are divided into two broad groups

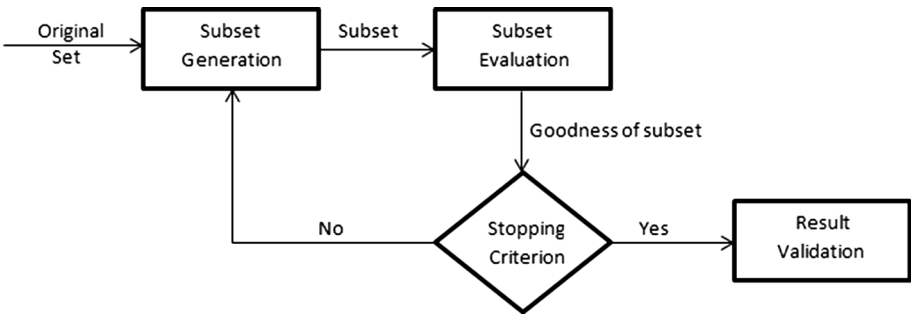

Fig. 1 Four key steps in feature selection [7]

(filters and wrappers) based on the criteria, whether or not feature selection is done independently of the classifier [24]. If the selection of feature is done independently of the classifier, then it is known as filter approach; otherwise, it is called as wrapper approach. The filter approach makes use of intrinsic properties of the data for feature ranking [25]. It is computationally faster than wrapper approach, but it has a limitation that it gives the optimal subset independently of learning algorithm. Wrapper approach is though computationally expensive, it takes classifier into account and gives much better accuracy [10]. But sometimes it leads to overfitting of data when the numbers of instances are less. There is also a hybrid approach (combination of filter and wrapper approach) that attempts to take advantage of two models by exploiting their different evaluation criteria in different search strategies. The process of feature selection using filter approach for ranking the features can be further categorized into two parts build upon the criterion: the one is *correlation criterion* and the second one is *information theoretic criterion* [19, 26].

3 Correlation criterion (redundancy measure)

Feature selection algorithms such as SFS and SBS take correlation into consideration and provide the final subset by considering the dependencies between the features. The objective of the correlation-based algorithm is to eradicate redundancy, as two highly correlated features are redundant in nature and taking any one of these will be efficient. If two variables are perfectly correlated, that means these are redundant and no additional information is gained by adding these features [19].

There are two widely used types of measures for the correlation between two variables: linear and nonlinear dependencies. These linear dependencies are calculated by some correlation coefficient, and the nonlinear dependencies are calculated with the help of entropy. For a pair of variables (A, B), the linear correlation coefficient p is given by the formula:

$$p = \frac{\sum_i (a_i - \overline{a}_i)(b_i - \overline{b}_i)}{\sqrt{\sum_i (a_i - \overline{a}_i)^2}\sqrt{\sum_i (b_i - \overline{b}_i)^2}} \tag{1}$$

where A is the training data and B is the class label, $\overline{a}_i$ is the mean of A and $\overline{b}_i$ is the mean of B. The value of p lies between -1 and $+1$ inclusive. If A and B are completely correlated, p acquires the value $+1$ or -1. If A and B are totally independent, p is zero [27]. The nonlinear correlation is computed by symmetrical uncertainty (SU) with the formula:

$$SU(A, B) = 2\frac{H(A) - H(A|B)}{H(A) + H(B)} \tag{2}$$

It compensates for information gain's bias toward features with more values and normalizes its values to the range [0,1]. A and B both are representing the features. If the value is 1, it indicates that knowledge of the value of either one completely predicts the value of the other, and if the value is 0, it indicates that A and B are independent. In addition, it still treats a pair of features symmetrically [28].

4 Information theoretic criterion (relevancy measure)

An information theoretic ranking criterion is also used by the number of feature selection algorithms. This criterion is also referred as a relevance index or scoring criterion that potentially measures the usefulness of an individual feature [29]. The rank of a feature is calculated with the help of conditional entropy. Various methods depending on this criterion are mutual information maximization (MIM), mutual information-based feature selection (MIFS), joint mutual information (JMI) and condition mutual information maximization (CMIM) [30].

The basic unit of information is the entropy of a random variable A that shows the uncertainty present in the distribution of A. It is defined by:

$$H(A) = -\sum_{a \in A} P(a) \log P(a) \tag{3}$$

where a denotes the possible values that A can adopt. In order to compute this equation, estimation of distribution of $P(A)$ is required [31]. If the distribution is highly biased toward one particular event $\in A$, then uncertainty is less and entropy is low. If all events are equally likely, uncertainty is high and value of entropy is also high. Mutual information shows the amount of information shared by A and B, where A represents feature and B represents the class concept C. The conditional entropy of A is given by:

$$H(A|B) = -\sum_{b \in B} P(b) \sum_{a \in A} P(a|b) \log P(a|b) \tag{4}$$

where $P(a)$ is the prior probabilities for all values of A and $P(a|b)$ is the posterior probabilities of A given the values of B. The amount by which the entropy of A decreases reflects additional information about A provided by B [32]. Mutual Information between A and B is given by:

$$I(A : B) = H(A) - H(A|B) \tag{5}$$

It has been concluded from the literature MIM has been used by researchers for ranking the features in the feature set [27]. So, in this work, MIM has been chosen in order to

assign relevancy score to different features. MIM assumes each feature is independent of all other features and effectively ranks the feature in descending order of their individual mutual information content. However, where features may be independent, this is known to be suboptimal. In general, it has been widely accepted that a useful and parsimonious set of features should not only be individually relevant, but also it should not be redundant with respect to each other [33].

After evaluation of feature subset, a stopping criterion is needed in order to avoid feature selection process to run exhaustively or forever through the search space of the subset. The stopping criteria are influenced by both the generation procedure and the evaluation function. Stopping criteria influenced by the generation procedure include whether a predefined number of features are selected or predefined numbers of iteration are reached. Stopping criteria based on the evaluation function include whether the addition (or deletion) of a feature does not produce any better result or an optimal subset according to some criterion is obtained [5]. It has been observed from the literature that GA, DEFS and PSO generally employ stopping criteria based on the generation procedure such as the number of iterations or a predefined fitness threshold, which is probable of getting stuck in local minima [34]. Some stopping criteria need to be considered in order to avoid local minima and to reduce the processing time while maintaining the accuracy.

The proposed feature selection method is based on a combination of local optimization techniques (MIM, SBS) with global optimization techniques (GA, DEFS and PSO). In this work, the chosen local optimization techniques consider both the information theoretic gain of feature and correlation between the features so as to include relevant and nonredundant features. The stopping criteria of these global optimization techniques have been modified to improve the processing time.

Optimization techniques for feature selection have been proposed by a number of researchers. In 2004, *Oh* et al. had proposed a hybrid genetic algorithm in order to improve the finetuning capability of GA, as GA is weak in finetuning near local optimum points, which results in long execution time. They had basically embedded a problem-specific local search operation in GA. These problems can be traveling salesman, graph partitioning and image compression. GA has many inherent variations and parameters that need to be handled properly for a specific problem. Hereby this makes their approach a problem-specific one. Also, a guided hybrid genetic algorithm had been proposed by Jung and Zscheischler, in 2013, which try to minimize the cost function evaluation. But their approach also needs controlling of parameters like crossover and mutation [35].

A hybrid algorithm based on PSO and artificial fish swarm algorithm (AFSA) had been proposed by Jiang et al. 2012 [36], in order to increase the versatility in population to achieve better accuracy. A hybrid feature selection that combines the particle swarm optimization (PSO) and differential evolution feature selection (DEFS) for minimizing the time complexity had been proposed by Balakrishnan et al. [37].

The proposed hybrid approach in this paper makes the global optimization techniques a generic approach that can be applied to different problems, as it controls the generation of population initially, with the help of local optimization algorithms (MIM and SBS). This approach reduces the probability of getting stuck into local optima from these locally optimized algorithms. Also the stopping criteria, used for these optimization techniques, improve the processing time.

5 Proposed methodology

In the proposed method, initially two subsets are created by applying information theoretic-based algorithm (MIM) and a correlation-based algorithm (SBS) on the original dataset. These subsets are combined so as to get a subset of features that are relevant and nonredundant. Then, a global optimization technique (GA, DEFS or PSO) is applied on the dataset with resultant feature subset separately along with variance as stopping criteria to give efficient subset of features. The procedure of the proposed method is as follows:

a. A subset of features ($S1$) is attained by using SBS algorithm.
b. Another subset is obtained ($S2$) by applying MIM algorithm to the original dataset, at which it gives maximum accuracy
c. Union of $S1$ and $S2$ is taken to get a subset $S3$.
d. Global optimization techniques, with variance as a stopping criterion, have been applied to feature subset $S3$.

In this proposed procedure of feature selection, step (a) is used to find the set of nonredundant feature set ($S1$) and step (b) is used to find the set of relevant features ($S2$). Step (c) simply combines the two features sets ($S1$) and ($S2$) to get another feature set ($S3$). The above three steps are performed so as to get a parsimonious set of features containing relevant and nonredundant features. The problem with the above set S is that it may gravitate toward local minima because of the inability of SBS to re-evaluate the usefulness of features that were initially discarded. Hence, it requires the inclusion of slight randomness in its procedure that can be done by the addition of a single

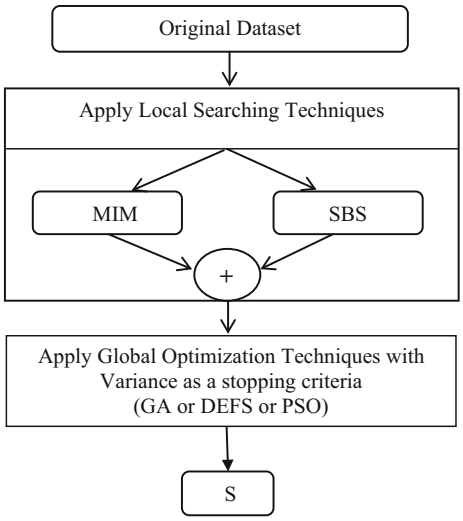

Fig. 2 Steps of feature subset selection in the proposed work

global optimization technique (either GA or DEFS or PSO). Finally, the optimization techniques with variance as stopping criteria have been applied on the reduced feature set (S) so as to get the reduced feature set with GA or with DEFS or with PSO in (d). This is shown in the Fig. 2.

Information-based algorithms (MIM, JMI, CMIM, MIFS) give a relevancy score to every feature [38]. For a class label B, the mutual information score for A_k is given by:

$$F_{\text{mim}}(A_k) = I(A_k : B) \tag{6}$$

This feature scoring criterion is known as 'MIM.' The function F_{mim} simply ranks the features in order of their MIM score and selects top k features, where k is decided by getting some predefined number of features or some stopping criteria that are reached [38]. It assumes that every feature is independent of each other. On the other hand, correlation-based local optimization algorithms (SFS and SBS) are used to remove the redundant features, among the subset to eliminate the unnecessary noise to give better accuracy.

From the literature, it has been observed that a new criterion is required, which takes into consideration both the discriminant ability of individual features and the correlation between these features, to effectively filter out nonessential features. To attain optimal subset of features, the discriminating power of an individual features and correlation between the features has been considered in this work, for which two locally optimized techniques have been chosen.

In this work, for discriminating ability, one information-based method is considered. MIM has been chosen among MIFS, CMIM and JMI (information-based methods); it is because of its simplicity. MIM is used to rank the features individually. Though it provides the relevant features, it has been proved that an optimal subset of features should not only be individually relevant, but also the features in subset should not be redundant to each other. In 2004, *Fleuret* had concluded that features in subset should not be highly correlated [39]. All the other algorithms such as MIFS, JMI and CMIM try to achieve redundancy relevancy goal. So, purely relevancy-based method MIM has been chosen.

In considering the correlation between features, the proposed approach uses SBS (a correlation-based method). SBS is preferred over SFS, because SFS is supposed to generate weaker subset since importance of feature is not assessed in context of other features not included yet. However, in SBS, the generation of subset is strong, since each feature is assessed in the context of almost all other features [26]. So if two features are highly relevant and redundant, then keeping both the features in the feature subset will not provide any additional information for the classification. It will only summate to the processing time, and it might be possible that accuracy gets reduce, if redundant features add unwanted noise to the important data. Since discrete class data have been chosen for testing, it involves nonlinear dependency between the features that is computed with the help of symmetrical uncertainty. So a set of relevant and nonredundant feature set is obtained from Eqs. (2) and (5) that results in the following equation named as RNR (relevant and nonredundant).

$$\text{RNR}(A, B) = I(A : B) + \text{SU}(A, B) \tag{7}$$

The first term $I(A : B)$ accounts for the relevancy of a feature. Higher the value of $I(A : B)$, the greater is the relevancy of feature and this value will be higher only if the uncertainty of A in the presence of B is very less (the value of $H(A|B)$ should be small). It basically measures the correlation between a feature F_i and the class C (named as C-correlation). A feature F_i is said to be relevant to the class concept C iff $I(A : B) > \alpha$, where α is the threshold relevance value [32].

The second term $\text{SU}(A, B)$ accounts for the removal of redundant features. It measures the correlation between a pair of features F_i and F_j ($\forall i \neq j$). It is also known as F-correlation. S' represents the desired set of features from a whole set of features F, which will be initially empty. The value of $\text{SU}_{j,i}$ (correlation) is zero, until the S' is empty (only one feature F_i, not any other feature F_j). It has to be decided whether the correlation between the pair of features formed by choosing the new feature F_i from whole set of features F and all the features F_j present in S' is high enough to cause redundancy, so that one of these features can be removed. The value $\text{SU}_{j,i}$ can be used to estimate the extent to which F_i is correlated with all the features F_j in

S'. Therefore, it is possible to identify highly correlated features, using some threshold value β. Higher the value of β, more is the redundancy between the features. Thus, the value of $\mathrm{SU}(A, B)$ should not be higher than some threshold value β. Therefore, it has been concluded that a feature F_i is said to be relevant and nonredundant iff $\gamma \leq \mathrm{RNR}(A, B) \leq \mu$, where γ and μ are the threshold value, which are decided for every dataset.

After obtaining the subset of relevant and nonredundant features, the global optimization techniques have been applied so as to avoid these locally optimized algorithms (SBS and MIM) from getting struck into the local optima. There is a possibility that the accuracy of the above obtained subset might get reduced, since when relevant features are combined with some nonredundant feature, it is because inclusion of nonredundant feature may add noise to the feature subset. Hence, the accuracy may get reduced. However, optimization techniques are applied on it, in which a large number of possible combinations of different features have been explored. Thus, this will increase the classification accuracy and also there is a chance that the number of features in the feature subset may also get reduced. There is a need to embed a global optimization technique for the process of feature selection as SBS and MIM suffer from the limitations that:

1. The convergence to an optimal solution depends on the chosen optimal solution, and most of these sequential algorithms tend to get struck into a suboptimal solution [40].
2. An algorithm efficient in solving one optimization problem may not be efficient in solving a different optimization problem. These techniques may be problem specific [40].

Also, to improve the performance of these global optimization techniques, there is need to control the randomness while generating the population initially as well as to improve the processing time. So, sequential algorithms and MIM have been used to reduce the initial features set to a reduced subset that contain relevant and redundant features, for the global optimization techniques.

The optimization techniques have been chosen so as to raise the classification accuracy and to maintain the stability by controlling the randomness. The first step of the optimization algorithms is to generate the constant number of population (possible solutions) randomly. The population varies with the desired number of features.

For a dataset containing 100 features, choosing ten best features from it is a tedious task. The number of all possible solutions is ${}^{100}\mathrm{C}_{10}$ whose value is approximately equal to 10^{14}. Selection of the same set of ten features on different run of the algorithm is not possible with very high probability, hence resulting in unstable classification accuracy. Also, the probability that those ten features out of 100 features will contain only relevant and nonredundant features is very less. So, by any mean, if this randomness is controlled, then the probability of getting higher and stable classification accuracy will increase. The proposed approach has controlled the randomness of GA, DEFS and PSO while generating the initial population. Suppose union of relevant and nonredundant subset given by MIM and SBS has reduced the original feature set of 100 features to 40. Now the number of possible solution is ${}^{40}\mathrm{C}_{10}$ that is approximately equals to 10^9 ($<<<10^{14}$). The probability of getting same ten features from 40 features on the rerun of the algorithm increases when compared with the previous case. Since these ten features will be obtained from a set of important features only, automatically the classification accuracy of these optimization techniques will increase.

Additionally, the stopping criteria of GA, DEFS and PSO have been modified in order to view its effect on execution time. Initially, GA has used N-iteration as the stopping criteria, that is, the algorithm stops when the number of iterations of execution becomes N. Another criterion could be K-iteration, which is if the best fitness value obtained in K consecutive iteration is same, the algorithm is terminated. Similarly, DEFS and PSO have also used N-iteration, which is basically when the number of generation becomes N, the algorithm stops. All the above-mentioned stopping criteria suffer from the problem of local optima. Another stopping criterion could be variance as stopping criterion, which fixes a bound ($\in$) on the variance of best fitness values obtained through a number of iterations, and the algorithm stops when the variance is less than the predefined bound [29].

The proposed approach has embedded variance as stopping criteria for the globally optimization techniques (GA, DEFS and PSO) in order to reduce the processing time of algorithm while maintaining the classification accuracy. This stopping criterion avoids the GA, DEFS and PSO to get struck in local optima. It is preferred over the other two, the iteration-based stopping criteria because they are based on the fact that the algorithm converges as the number of iterations tends to ∞, which seems to be a kind of impossible case, whereas stopping criteria using variance are based on the fact that the difference between global optimal value of the function and the fitness function tends to 0, which is a possible case. It gives the optimal subset of features at which the classification accuracy is maximized. Users need not to input the number of features, which must be contained by the optimal subset, as per the user command.

In order to determine the actual computational cost of a method, an exact analysis of computational complexity is computed. Big-O notation is a prominent approach in terms of analyzing computational complexity. There are four

basic steps in this approach, namely assignment of relevancy score, nonredundant features subset generation, subset construction and evaluation and stopping criteria. The computational complexity of this approach in order to show that inclusion of different types of techniques does not increase the computational complexity in selecting a feature subset is discussed as:

(i) *Assignment of relevancy score* In this step, information gain for each feature is measured by MIM. It assigns relevancy score to every feature as per the value of information gain. If the number of total features for a given dataset is N, then cost of assigning relevancy score to all the features is $O(N)$. It is further mentioning that this cost is required only once, specifically before feature selection process.

(ii) *Nonredundant features subset generation* In this step, a subset of nonredundant features is generated by SBS that is based on correlation criteria. The time complexity for computing this subset is $O(N)$.

(iii) *Subset construction and evaluation* In this paper, population-based optimization techniques have been considered. All three techniques, namely GA, DEFS and PSO, contain population, and each individual of the population represents a subset. To construct the population of constant size k, in which each subset is of size d, where 'd' is the desired number of features, assuming that the upper bound on the number of iteration has been fixed to 'i'. Each of the fitness is evaluated in order to determine its classification ability. The time complexity to perform the operation of subset construction and evaluation is $O(d \times k \times i)$.

(iv) *Stopping criteria* In this approach, for optimization techniques, stopping criteria have been modified by the help of variance. This modification has a positive impact over the computational complexity. The best case complexity for subset construction and evaluation has been reduced to $O(d \times k)$, as the number of iteration will reduce to a constant due to inclusion of variance as stopping criteria, and hence, it can be removed.

The total computational cost of this approach for the worst case will be $O(N + N + d \times k \times i)$. The first and second terms $O(N + N)$ are the cost of operation performed only once. $(N \ll d \times k \times i)$, since the value of $k \gg N$. Thus, these terms can be removed from the total computational cost. It can be concluded that the total cost of this approach is either less or equal to that of other existing GA, DEFS and PSO-based feature selection approach due to the indulgence of variance as stopping criteria. Thus, the incorporation of several techniques in this approach does not increase the computational cost.

6 Results

The proposed approach has been tested with three publicly available datasets, namely Sonar, Wdbc and German. These datasets are obtained from UCI machine learning repository [41]. The specification about various datasets is shown in Table 1. The number of classes in each dataset is two, with a constraint of no missing feature, and all the features contain numeric values. The evaluation of the proposed approach is conducted in terms of three criteria: processing time and classification accuracy and the number of selected features [42]. The accuracy has been measured with the help of a SVM classifier having linear kernel, gamma as 0.001 value, and tenfold crossvalidation is used for validation. A comparison of MIM, SBS, GA and proposed approach over three datasets on maximum accuracy achieved with the number of features (NOF) is shown in Table 2.

Initially, MIM has been applied to Sonar dataset, and the accuracy is calculated for the features from 10 to 45. The maximum accuracy obtained with MIM is 80.2885% with 30 features, which are considered for the subset. SBS is applied to Sonar dataset, which gives 73.5577% as the classification accuracy with 18 features. A union of the feature subsets obtained from both the algorithms is computed, resulting in another set of 38 features shown in Table 2. In order to reduce the computation time of GA, DEFS and PSO, while increasing or maintaining the accuracy, in place of considering the whole dataset, this resultant dataset is used. The results show an increase in the classification accuracy with proposed methodology.

After the initial population of these optimization techniques has been controlled, proposed approach for GA acquires a gain of 1.4423% when compared with MIM, a gain of 8.1731% over SBS and a gain of 5.2885% over GA by achieving a maximum accuracy of 81.7308% with 24 features as shown in Fig. 3. The proposed approach also shows an improvement in the classification accuracy of DEFS and PSO. Proposed approach for DEFS achieves a maximum accuracy of 80.2885% that with 25 features in the optimal subset, which is 1.4423% more than the basic

Table 1 Details of datasets

Datasets	No. of instances	No. of features
Sonar	208	60
Wdbc	569	30
German	1000	24

Table 2 Comparison of MIM, SBS, DEFS, GA and proposed approach for GA, DEFS and PSO on maximum classification accuracy achieved with three datasets (in %)

Datasets and Criteria		MIM	SBS	MIM + SBS	GA	MIM + SBS + (GA)v	DEFS	MIM + SBS + (DEFS)v	PSO	MIM + SBS + (PSO)v
Sonar	No. of features	30	18	38	24	24	17	25	34	25
	Accuracy (in %)	80.2885	73.5577	78.3654	76.4423	81.7308	78.8462	80.2885	79.3269	80.2885
Wdbc	No. of features	16	13	22	18	12	14	10	19	12
	Accuracy (in %)	95.9578	96.3093	95.0791	95.9568	96.3093	94.9033	96.3093	95.9578	96.6608
German	No. of features	22	13	23	22	19	22	20	20	20
	Accuracy (in %)	77.7000	77.2000	77.4000	77.7000	77.7000	77.2000	77.8000	76.8000	78.000

DEFS and 6.7308% more than SBS. Though the maximum accuracy attained by DEFS is same as that of MIM, it has been achieved with five less features. In case of proposed approach for PSO, a gain of 0.9616% has been achieved and that is with a subset of lesser nine features, a gain of 6.7308% over SBS has been achieved, and has achieved a classification accuracy same as that of MIM, but with five less features in the subset.

Figures 3, 5 and 7 show that for some particular number of features, the classification accuracy of proposed methods for global optimization techniques becomes equal to the classification accuracy of GA, DEFS and PSO. However, the processing time of both basic optimization techniques and proposed method are compared for acquiring that particular number of features. Figures 4, 6 and 8 show a significant fall in processing time of proposed approach for GA, DEFS and PSO, when compared with the processing time of basic GA, DEFS and PSO. The basic algorithms for GA, DEFS and PSO have been executed for obtaining 10 to 55 features. Maximum classification accuracy has been attainded with 24 features by GA, with 25 features by DEFS and with 25 features by PSO. It has been observed that the accuracy of these basic algorithms either degrades or remains same. Therefore, in order to maintain the consistency of the classification accuracy and processing time has been shown for 38 features.

Similarly, the proposed method has been tested with Wdbc dataset. MIM has been applied to the original dataset, and it gives maximum classification accuracy of 95.9578% with a feature subset of 16 features that needs to be considered after which SBS has been applied to Wdbc dataset and it provides maximum accuracy of 96.3093% with 13 features. Table 2 shows that after taking the union of both the subsets, the resulting subset contains 22 features, on which GA, DEFS and PSO are applied over it one by one. Though the maximum accuracy obtained with proposed approach for GA was same as the accuracy obtained by SBS, equal to 96.3093%, that is achieved with the reduced number of features. When proposed approach for GA is compared with GA, there is only 0.3525% increase in classification accuracy, but it attains maximum accuracy with 12 features only.

Proposed approach for DEFS and PSO shows an increase in classification accuracy when compared with basic DEFS and Basic PSO. Proposed approach for DEFS attains a gain of 1.406% over basic DEFS and 0.3515% over MIM. Proposed approach for PSO outperforms all the three optimization techniques by attaining an accuracy of 96.6608%.

Figures 9, 11 and 13 show the classification accuracy of GA, DEFS and PSO and proposed method of these techniques for a certain number of features. It has been observed from the above-mentioned figure that the

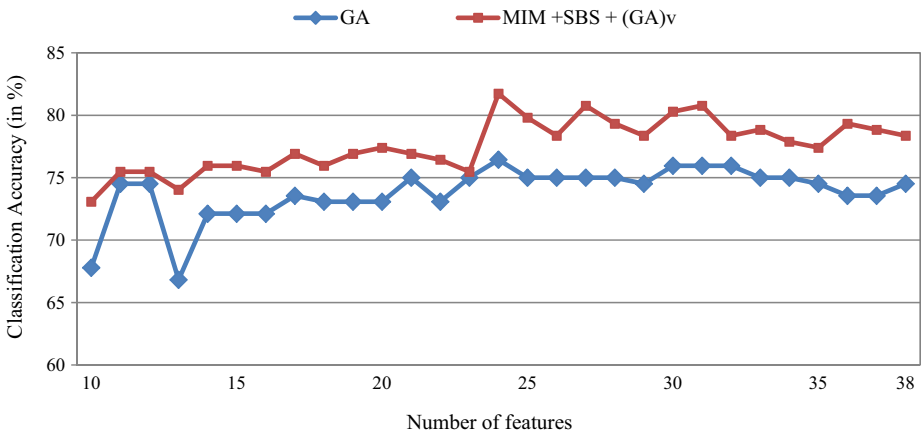

Fig. 3 Classification accuracy of GA and proposed approach for GA (Sonar dataset)

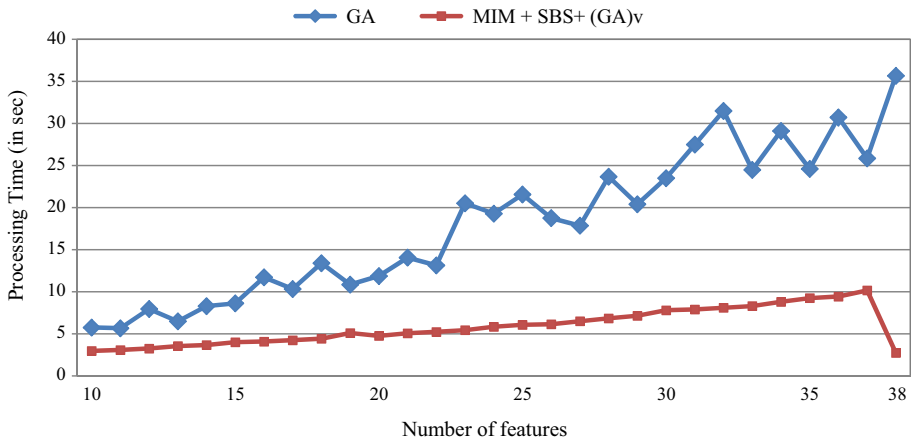

Fig. 4 Processing time of GA and proposed approach for GA (Sonar dataset)

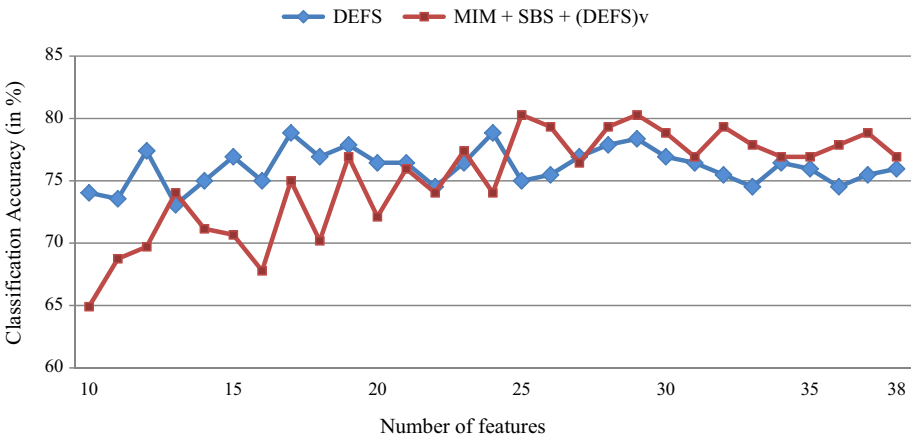

Fig. 5 Classification accuracy of DEFS and proposed approach for DEFS (Sonar dataset)

accuracy of proposed approach for GA, DEFS and PSO becomes approximately equal to accuracy of basic methods for a certain number of features. But for that particular feature, Figs. 10, 12 and 14 show that the processing time of proposed approach for GA, DEFS and PSO has been extremely reduced due to indulgence of variance as stopping criteria, when compared it with basic GA, DEFS and PSO.

Also, the proposed methodology has been tested to German dataset. A subset of 22 features providing the maximum accuracy of 77.7% has been attained by applying MIM on the original dataset. SBS generates a subset of 13 features giving an accuracy of 77.2%. By combining the feature subset obtained from MIM and SBS, the resulting subset contains 23 features shown in Table 2. GA, DEFS and PSO have been applied to the training data with 23

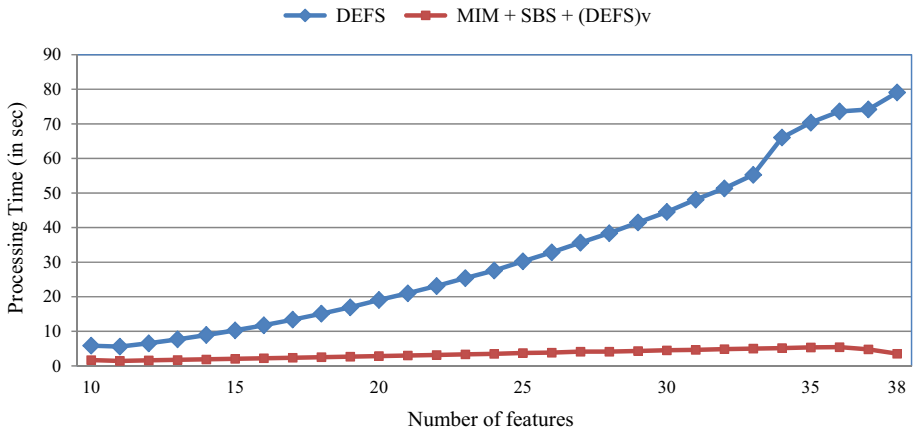

Fig. 6 Processing time of DEFS and proposed approach for DEFS (Sonar dataset)

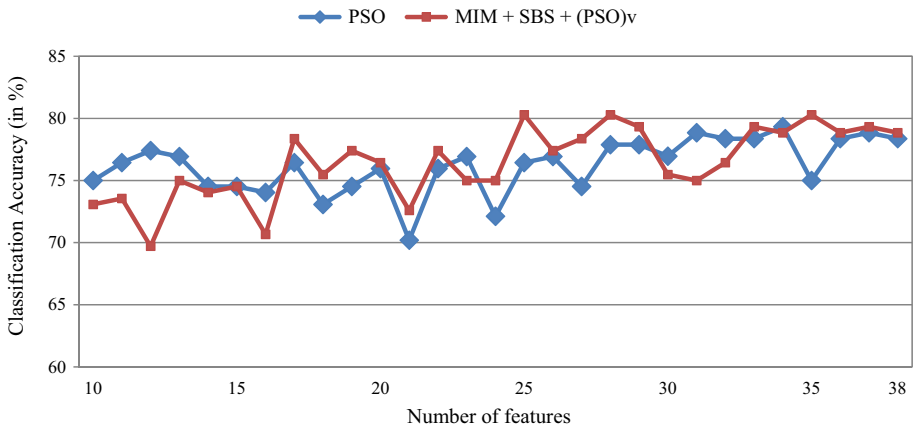

Fig. 7 Classification accuracy of PSO and proposed approach for PSO (Sonar dataset)

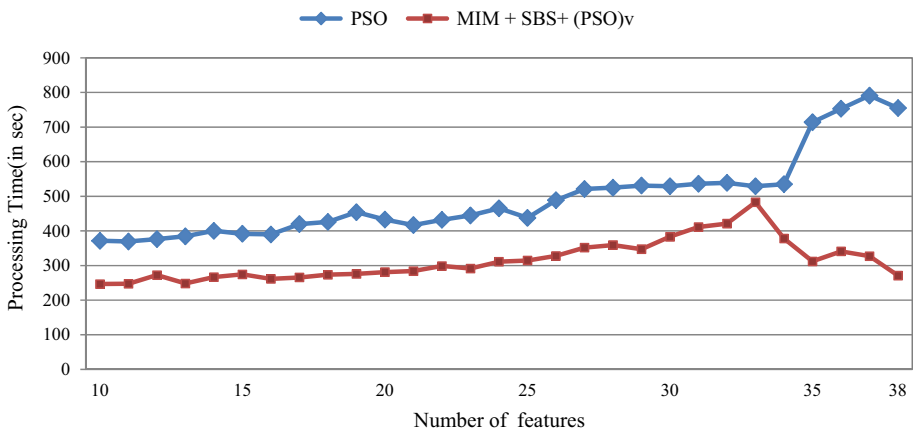

Fig. 8 Processing time of PSO and proposed approach for PSO (Sonar dataset)

features individually. The proposed approach for GA attains the maximum classification accuracy of 77.7% with 19 features, which is equivalent to the maximum classification accuracy of MIM and GA, but these algorithms attain it with 22 features. When DEFS is applied to German dataset, it attains a maximum of 77.8% with 20 features and proposed approach for PSO attains a maximum of 78% classification accuracy. The classification accuracy and the processing time for different numbers of features with basic GA, DEFS and PSO and proposed approach for GA, DEFS and PSO method are shown in Figs. 15, 16, 17, 18, 19 and 20.

Figures 4, 6, 8, 10, 12, 14, 16, 18 and 20 show that there is an abrupt fall in processing time of the proposed method of GA, DEFS and PSO, when all features are considered for classification while the basic of these optimization techniques does not show any fall in processing time with the same number of features. The fall in processing time is

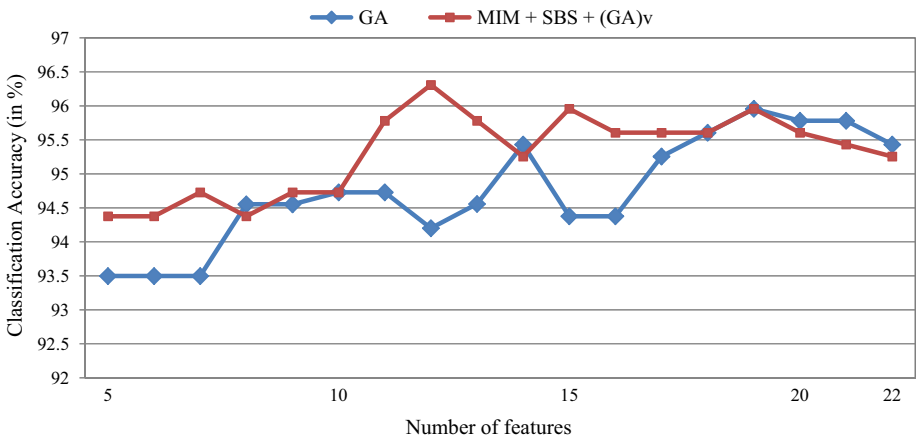

Fig. 9 Classification accuracy of GA and proposed approach for GA (Wdbc dataset)

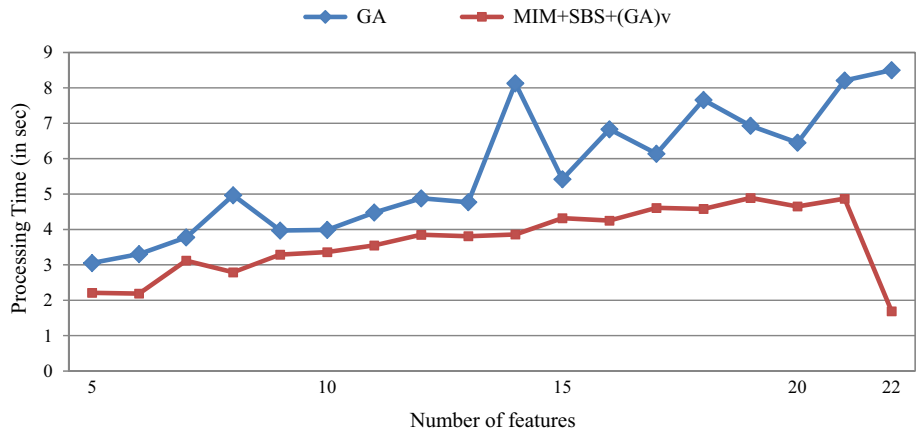

Fig. 10 Processing time of GA and proposed approach for GA (Wdbc dataset)

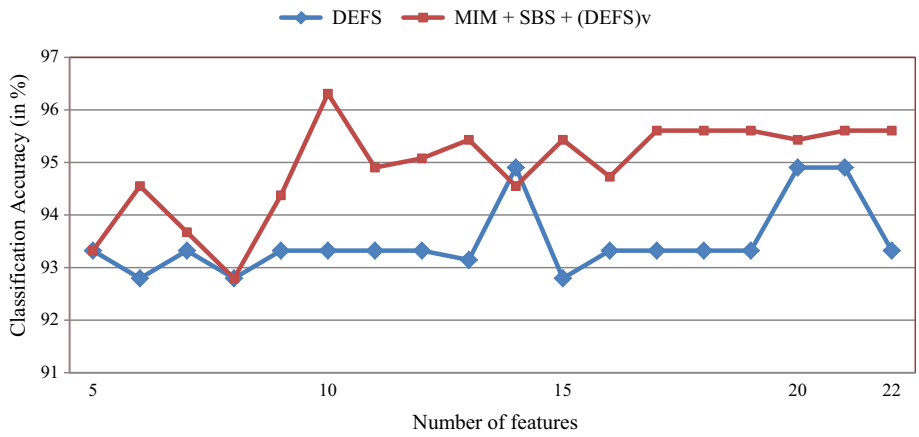

Fig. 11 Classification accuracy of DEFS and proposed approach for DEFS (Wdbc dataset)

due to the reason that the proposed method reduces the size of original feature set from n to k using MIM and SBS. Selecting all k features from the reduced feature set will generate same kind of individuals in the population, resulting in faster convergence while selecting k features from n features will result in different types of individuals in the population, due to which it takes more time to converge. The possible number of individuals when selecting k features from n features will be ${}^{n}C_{k}$. However, Fig. 10 shows that processing time of optimization techniques also decreases along with the proposed method. In this case, the size of original feature set n and the size of a reduced set of features k are approximately same. The value of n is 24 and that of k is 23. Therefore, processing time of both optimized and proposed method decreases in this case.

The comparison of GA, DEFS and PSO with proposed methodology for three datasets based on processing time is

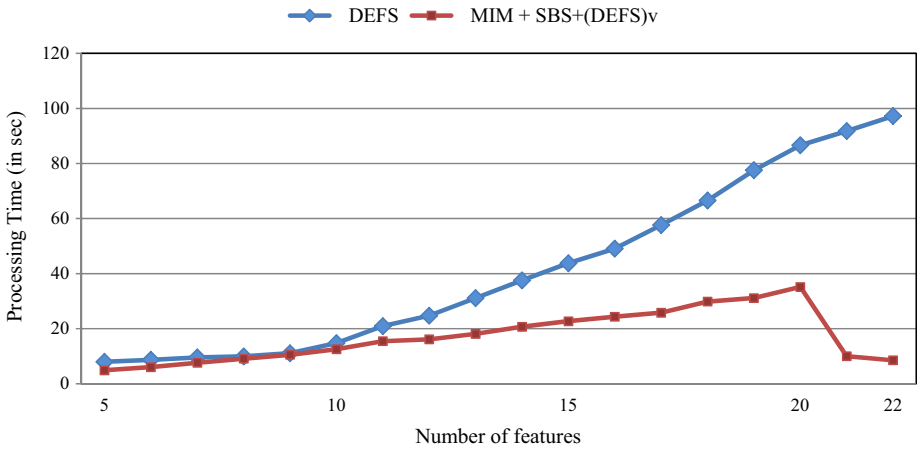

Fig. 12 Processing time of DEFS and proposed approach for DEFS (Wdbc dataset)

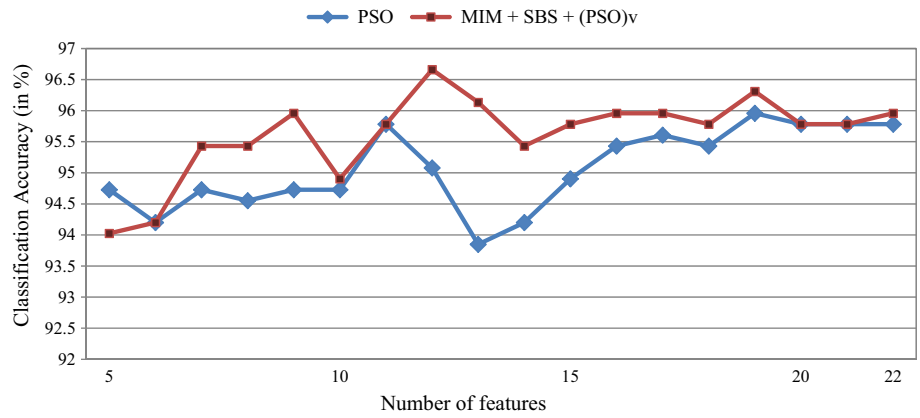

Fig. 13 Classification accuracy of PSO and proposed approach for PSO (Wdbc dataset)

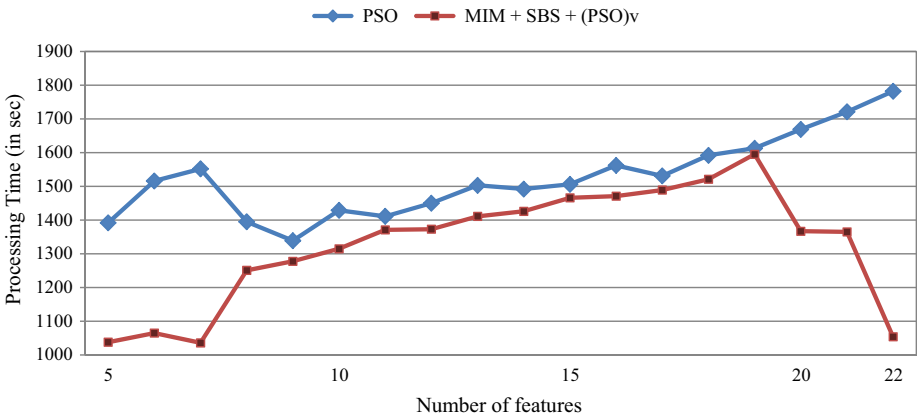

Fig. 14 Processing time of PSO and proposed approach for PSO (Wdbc dataset)

shown in Table 3. Tables 2, 3 show that the proposed approach provides higher accuracy with the reduced number of features and with less execution time. For Sonar dataset, the proposed approach for GA shows an improvement of 66.50% over basic GA, proposed approach for DEFS gains an improvement of 78.65% over basic DEFS, and an improvement of 27.98% is achieved using proposed approach for PSO over basic PSO, when compared these proposed approaches in term of the execution time. Similarly, for Wdbc dataset, when the parameter time has been considered for comparison, an improvement of 40.10, 68.85 and 5.58% is attained with proposed approaches for GA, proposed approaches for DEFS and proposed approaches for PSO over basic GA, DEFS and PSO. The proposed approaches for GA, proposed approaches for DEFS and proposed approaches for PSO achieve a minimum gain of 46.55, 84.66 and 14.98% over basic GA, DEFS and PSO for German dataset. The reduction in

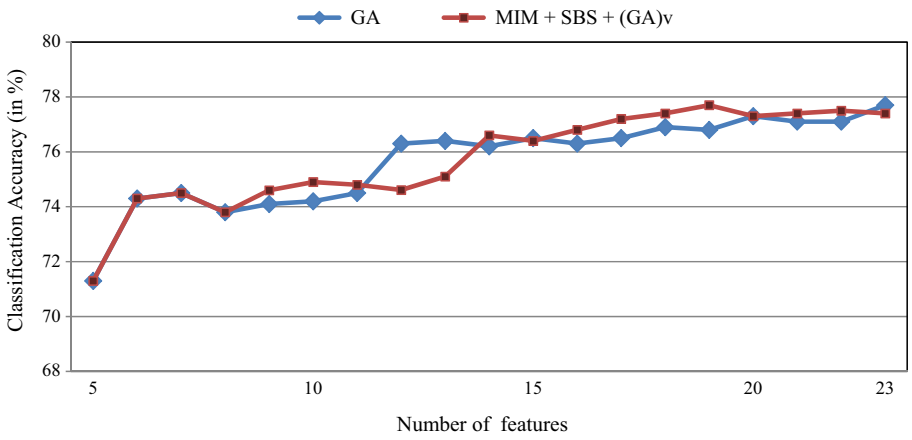

Fig. 15 Classification accuracy of GA and proposed approach for GA (German dataset)

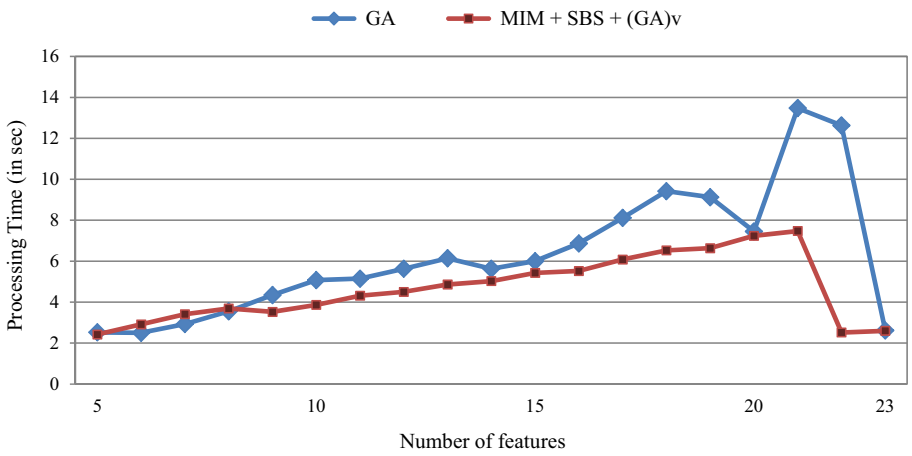

Fig. 16 Processing time of GA and proposed approach for GA (German dataset)

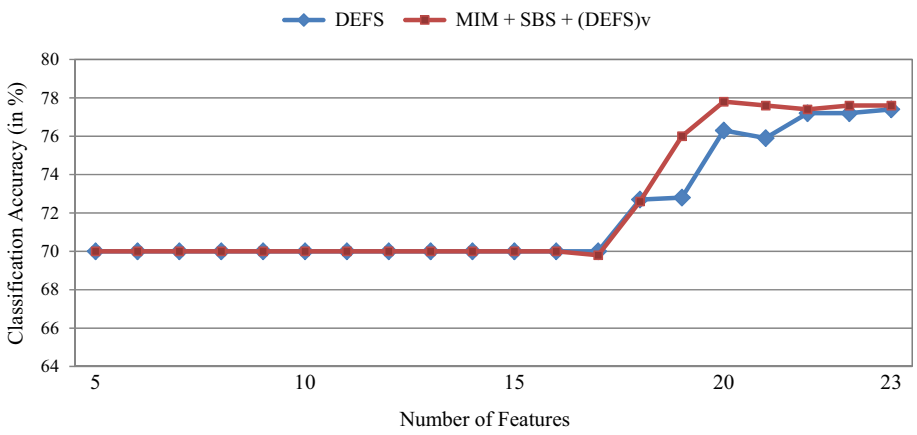

Fig. 17 Classification accuracy of DEFS and proposed approach for DEFS (German dataset)

execution time of the proposed approaches is due to the inclusion of variance as stopping criteria in these global optimization techniques.

Thus, the proposed approach for feature selection improves the performance in terms of the classification accuracy as well as the processing time as compared the original global optimization techniques. The subset of relevant and nonredundant features which is generated by using local searching methods is considered for the initial population improves the classification performance and variance as stopping criteria reduce the execution time.

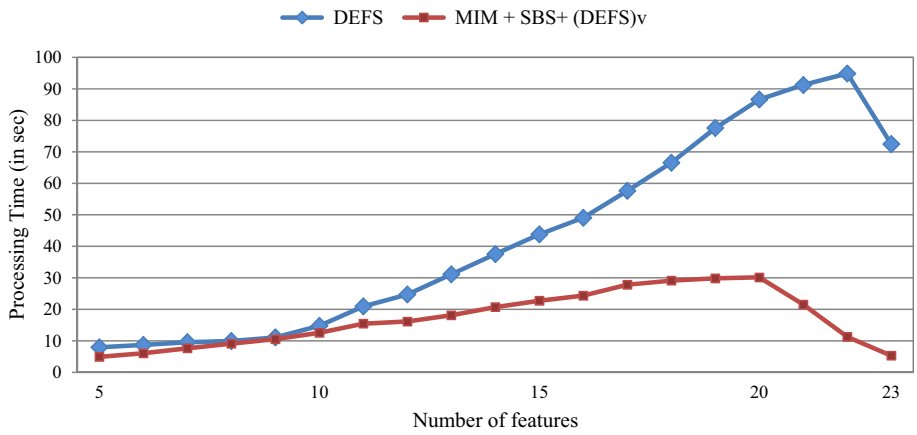

Fig. 18 Processing time of DEFS and proposed approach for DEFS (German dataset)

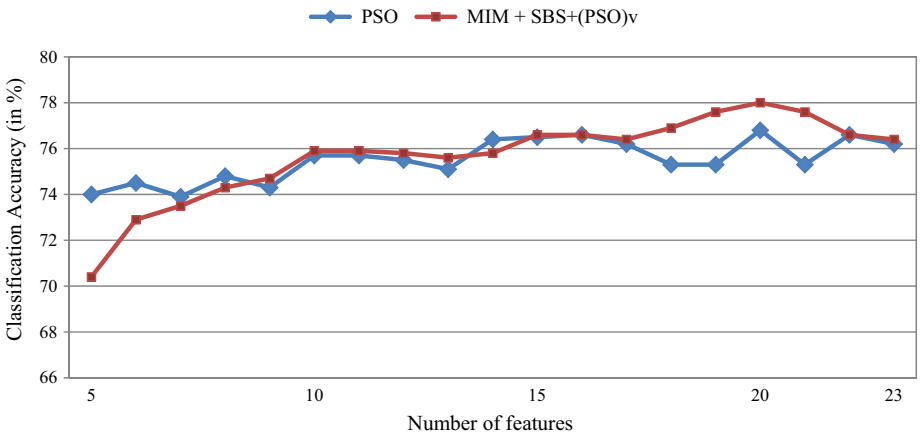

Fig. 19 Classification accuracy of PSO and proposed approach for PSO (German dataset)

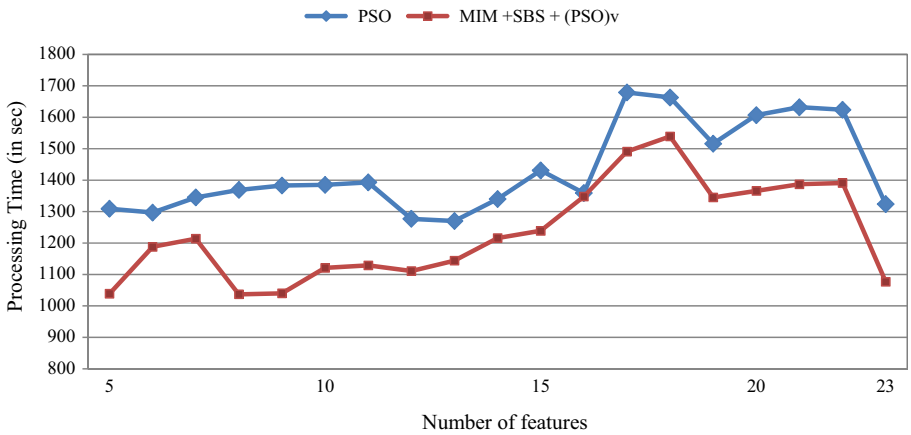

Fig. 20 Processing time of PSO and proposed approach for PSO (German dataset)

Table 3 Comparison of GA, DEFS and PSO with proposed approach for GA, DEFS and PSO on processing time for maximum accuracy with three datasets (in seconds)

Datasets	MIM	SBS	GA	MIM + SBS + (GA)v	DEFS	MIM + SBS + (DEFS)v	PSO	MIM + SBS + PSO_v
Sonar	0.5658	0.0529	19.2836	6.4587	30.2600	4.3500	437.73	315.23
Wdbc	0.1786	0.1219	6.9300	4.1505	23.4100	7.2900	1450.00	1373.30
German	0.0518	0.0689	12.6300	6.7564	86.6200	13.2807	1607.00	1366.12

7 Discussion and conclusions

This paper reviews existing feature selection methods and highlights their common limitations, and therefore, an approach has been proposed based on the existing methods, used for feature reduction, to achieve improvements in terms of classification accuracy and processing time. The proposed approach is a combination of local searching techniques (SBS, MIM) and global optimization techniques (GA, DEFS or PSO). The local searching techniques may trap in local minima, so there is a need of combining these with global optimization techniques. The global optimization techniques explore various combinations of features, thereby providing an opportunity to escape a subset from trapping into some local minima. The global techniques have the ability to converge to a solution quickly, but the selection of initial population is very important. Random initialization of population results into output that may suffer from lack of consistency in classification results and also a high time complexity. In this work, the original feature set has been reduced to contain relevant and nonredundant features using local searching techniques. The stopping criteria of global optimization techniques have also been modified by the help of variance. This approach is designed to resolve the problem of eliminating irrelevant and redundant features from the original feature set, so as to raise the classification accuracy and to reduce the processing time.

This approach has been evaluated using three publicly available datasets. The experimental results show that the proposed approach provides an improvement in terms of both, the classification accuracy and the computation time. The statistical significance of the reported results shows that the consistency of these optimization techniques has been increased and there is a significant reduction in the processing time, when the proposed methodology is compared with the basic GA, DEFS and PSO. The consistency has been attained by controlling the randomness while generating the population randomly, and the processing time of these optimization techniques has been reduced due to embedding of stopping criteria based on variance. The accuracy has also been increased due to the selection of feature subset from a set of relevant and nonredundant features, which has been obtained by using MIM and SBS. The presented work demonstrates that the proposed method can be used in any application areas of expert system and intelligent systems with very high dimensionality such as gene expression array analysis, text processing of Internet document and combinatorial chemistry.

Future work includes more experiments on the modification of stopping criteria for search strategies such as ACO and ABC. These improvements can be made by studying the information shared between features and class labels, classifying the features into strongly relevant, relevant, weakly relevant, and redundant feature based on the information that the features add to the selected subset.

Compliance with ethical standards

Conflict of interest The authors do not bear any financial or personal relationships with other people or the organizations that could inappropriately influence their work.

References

1. Jain A, Zongker D (1997) Feature selection: evaluation, application, and small sample performance. IEEE T Pattern Anal 19:153–158
2. Kotsiantis S (2011) Feature selection for machine learning classification problems: a recent overview. Artif Intell Rev 42:157
3. Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5:1205–1224
4. Peng Y, Wu Z, Jiang J (2010) A novel feature selection approach for biomedical data classification. J Biomed Inform 43:15–23
5. Dash M, Liu H (1997) Feature selection for classification. Intell Data Anal 1:131–156
6. Sutha K, Tamilselvi JJ (2015) A review of feature selection algorithms for data mining techniques. Int J Comput Sci Eng 7:63–67
7. Liu H, Yu L (2005) Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data Eng 17:491–502
8. Narendra PM, Fukunaga K (1977) A branch and bound algorithm for feature subset selection. IEEE Trans Comput 100:917–922
9. Gupta P, Doermann D, DeMenthon D (2002) Beam search for feature selection in automatic SVM defect classification. Proc Int Conf Pattern Recogn 2:212–215
10. Kohavi R, Sommerfield D (1995) Feature subset selection using the wrapper method: overfitting and dynamic search space topology. In: Proceedings of international conference of knowledge discovery and data mining, pp 192–197
11. Pudil P, Novovicova J, Kittler J (1994) Floating search methods in feature selection. Pattern Recogn Lett 15:1119–1125
12. Liu Y, Zheng YF (2006) FS_SFS: a novel feature selection method for support vector machines. Pattern Recogn 39:1333–1345
13. Yusta SC (2009) Different metaheuristic strategies to solve the feature selection problem. Pattern Recogn Lett 30:525–534
14. Chaikla N, Qi Y (1999) Genetic algorithms in feature selection. Proc IEEE Int Conf Syst Man Cybernet 5:538–540
15. Dash M, Liu H (2003) Consistency-based search in feature selection. Artif Intell 151:155–176
16. Price K, Storn RM, Lampinen JA (2006) Differential evolution: a practical approach to global optimization. Springer, Berlin, pp 37–130
17. Ahmad I (2015) Feature selection using particle swarm optimization in intrusion detection. Int J Distrib Sens Netw. doi:10.1155/2015/806954
18. Christensen J, Marks J, Shieber S (1995) An empirical study of algorithms for point-feature label placement. ACM Trans Gr 14:203–232
19. Hall MA (1999) Correlation-based feature selection for machine learning. Dissertation, University of Waikato
20. Burrell L, Smart O, Georgoulas GK, Marsh E, Vachtsevanos GJ (2007) Evaluation of feature selection techniques for analysis of

functional MRI and EEG. In: Proceedings of international conference on data mining, pp 256–262
21. Vafaie H, Imam IF (1994) Feature selection methods: genetic algorithms vs. greedy-like search. Proc Int Conf Fuzzy Intell Control Syst 51:39–43
22. Ladha L, Deepa T (2011) Feature selection methods and algorithms. Int J Adv Trends Comput Sci Eng 3:1787–1797
23. Oh IS, Lee JS, Moon BR (2004) Hybrid genetic algorithms for feature selection. IEEE Trans Pattern Anal 26:1424–1437
24. Blum AL, Langley P (1997) Selection of relevant features and examples in machine learning. Artif Intell 97:245–271
25. Yuan Huang, TsengSS Gangshan W, Fuyan Z (1999) A two-phase feature selection method using both filter and wrapper. Proc IEEE Conf Syst Man Cybernet 2:132–136
26. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
27. Brown G, Pocock A, Zhao MJ, Luján M (2012) Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. J Mach Learn Res 13:27–66
28. Bennasar M, Hicks Y, Setchi R (2015) Feature selection using joint mutual information maximisation. Expert Syst Appl 22:8520–8532
29. Bhandari D, Murthy CA, Pal SK (2012) Variance as a stopping criterion for genetic algorithms with elitist model. Fundam Inform 120:145–164
30. Yu L, Liu H (2003) Efficiently handling feature redundancy in high-dimensional data. In: Proceedings of international conference on knowledge discovery and data mining, pp 685–690
31. Kwak N, Choi CH (2002) Input feature selection by mutual information based on parzen window. IEEE Trans Pattern Anal 24:1667–1671
32. Yu L, Liu H (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. Proc Int Conf Mach Learn 3:856–863
33. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal 27:1226–1238
34. Zhuo L, Zheng J, Li X, Wang F, Ai B, Qian, J (2008) A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine. In: Proceedings of geoinformatics 2008 and joint conference on GIS and built environment: classification of remote sensing images, pp 71471J–71471J
35. Jung M, Zscheischler J (2013) A guided hybrid genetic algorithm for feature selection with expensive cost functions. Proc Int Conf Comput Sci 18:2337–2346
36. Jiang J, Bo Y, Song C, Bao L (2012) Hybrid algorithm based on particle swarm optimization and artificial fish swarm algorithm. Int Symp Neural Netw 607–614
37. Balakrishnan U, Venkatachalapathy K, Marimuthu SG (2015) A hybrid PSO-DEFS based feature selection for the identification of diabetic retinopathy. Curr Diabet Rev 11:182–190
38. Brown G (2009) A new perspective for information theoretic feature selection. In: Proceedings of international conference on artificial intelligence and statistics, pp 49–56
39. Fleuret F (2004) Fast binary feature selection with conditional mutual information. J Mach Learn Res 5:1531–1555
40. Venter G (2010) Review of optimization techniques. Encycl Aerosp Eng. doi:10.1002/9780470686652.eae495
41. Lichman M (2013) UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml
42. Wang G, Song Q, Sun H, Zhang X, Xu B, Zhou Y (2013) A feature subset selection algorithm automatic recommendation method. J Artif Intell Res 47:1–34