DOI: 10.24996/ijs.2021.62.8.32

ISSN: 0067-2904

A Survey on Feature Selection Techniques using Evolutionary Algorithms

Nishath Ansari

Computer Science and Engineering, B V Raju Institute of Technolog, Narsapur, India

Received: 17/10/2020 Accepted: 18/2/2021

Abstract

Feature selection, a method of dimensionality reduction, is nothing but collecting a range of appropriate feature subsets from the total number of features. In this paper, a point by point explanation review about the feature selection in this segment preferred affairs and its appraisal techniques are discussed. I will initiate my conversation with a straightforward approach so that we consider taking care of features and preferred issues depending upon meta-heuristic strategy. These techniques help in obtaining the best highlight subsets. Thereafter, this paper discusses some system models that drive naturally from the environment are discussed and calculations are performed so that we can take care of the preferred feature matters in complex and massive data. Here, furthermore, I discuss algorithms like the genetic algorithm (GA), the Non-Dominated Sorting Genetic Algorithm (NSGA-II), Particle Swarm Optimization (PSO), and some other meta-heuristic strategies for considering the provisional separation of issues. A comparison of these algorithms has been performed; the results show that the feature selection technique benefits machine learning algorithms by improving the performance of the algorithm. This paper also presents various real-world applications of using feature selection.

Keywords: Feature Selection, Element, Element Selection, Machine Learning, and Meta-Heuristic Strategies.

1. Introduction

Nowadays, feature selection is consolidated in many fields. Feature selection or element selection is nothing but a selection of appropriate element subsets from the total number of elements. The total number of possibilities of choices from total elements N is said to be 2^N. The primary use of a selection of these subsets is their application for any machine learning strategy to achieve classification on it. The process of selection element subsets involves different techniques. In the past, we noticed that a huge number of features are emerging day by day for each one application. To run every single application, it is challenging to load all elements to identify the accurate result [1]. To avoid this problem, a new approach to element selection is used. Feature selection minimizes the total number of features and selects only efficient features based on input provided by reducing noisy data, which helps in identifying that application quickly. By removing irrelevant data from the input, relevant output features are generated. Different methods, like PCA (Principal Component Analysis), are also used to eliminate redundant data, but using those techniques results in some irrelevant output. To get more accuracy to compare to PCA, this new approach of feature selection gives more accurate and relevant features.

As the data is emerging day by day, numerous optimization techniques have been developed. With the use of the feature selection technique, the efficiency of the algorithms has been improved. In this paper, the various algorithms used for feature selection are filter-based feature selection, wrapper based feature selection, Genetic Algorithm (GA), Support Vector Machine (SVM), Particle Swarm

*Email: nishat.alm@gmail.com

Optimization (PSO), Binary Particle Swarm Optimization (BPSO), Variant of BPSO, Non-Dominated Sorting BPSO (NSBPSO), crowding, mutation and dominance BPSO (CMDBPSO), Decision Tree, Mutation Operator, and Neural Networks.

There are different methods in feature selection. The main methods involved in feature selection are

1. Filter method, and
2. Wrapper method.

1.1 Filter Method

This is the first method used in feature selection to make the choice of relevant features based on input to obtain the output. The process of the filter method mainly uses the ranking technique. Different variables are ranked based on the order of acceptance. The ranking strategy is used because of its simplicity to be applied to any application. The ranking is applied to features before those elements are given to the classification algorithm; i.e., because of the ranking of items, features are filtered. The name filter itself implies that some filtration and reducing process is taken place. Incompatible elements are filtered into compatible elements. Each ranked element should have a unique property to identify that class of application. The feature can be said as compatible if it is conditionally independent of that input provided, but should not be independent of the class label. In this technique, feature correlation is also used to identify the compatible features/elements. The ranking of elements can be categorized into two methods. One is the correlation technique, and the other is the mutual information technique. In the correlation technique, we consider the Pearson correlation coefficient [2]. And for the mutual information technique, we consider Shannons definition for entropy in information-theoretic ranking criteria [3, 4].

1.2 Wrapper Method

This is the second method in feature selection. The wrapper method is simply said as a prognosticator. The performance of the prognosticator is an essential key to choosing subsets. It is noticed that by considering 2^N features, we are undergoing NP-Hard issues. To eliminate these issues, the obtained element subsets are further reduced to subsets by considering some searching techniques, which results in subsets based on heuristical subsets. Some search techniques are used to identify the element subsets so that they can reduce incompatibility and the maximize efficiency and performance of classifiers. In search techniques, the first used strategy is the Branch and Bound method [5]. Also, some emerging sequential search methods can be used in wrapper techniques, like the Genetic Algorithm (GA) [6] or Particle Swarm Optimization (PSO) [7]; these can result in feasible performance and optimized solutions. Wrapper techniques are mainly differentiated into two types, say, I: Sequential Selection techniques, and II: Heuristic Search techniques. In Sequential Selection techniques, initially, a full set of data is given, and by applying the algorithm and based on the need, all the incompatible elements are removed, and the best solution is obtained, which is an optimized subset of features. Sequential Selection techniques include Sequential Feature Selection (SFS) algorithm, Sequential Backward Selection (SBS) algorithm, and Sequential Floating Forward Selection (SFFS) algorithms [8, 9]. In Heuristic Search techniques, they consider an objective function. Based on this objective function evaluation is performed to obtain an optimized subset of the solution. In this technique, searching subsets are used around the space of search or by itself; they generate solutions. Heuristic Search techniques include evolutionary algorithms, like the Genetic Algorithm (GA) [6] and CHC Genetic Algorithm (CHCGA) [10, 11]. Another method that is also used in the wrapper technique is the embedded technique. During the performance of this approach, the time is decreased by reclassifying the subsets. In this step, we consider the training process and apply a greedy selection strategy to obtain optimized subsets.

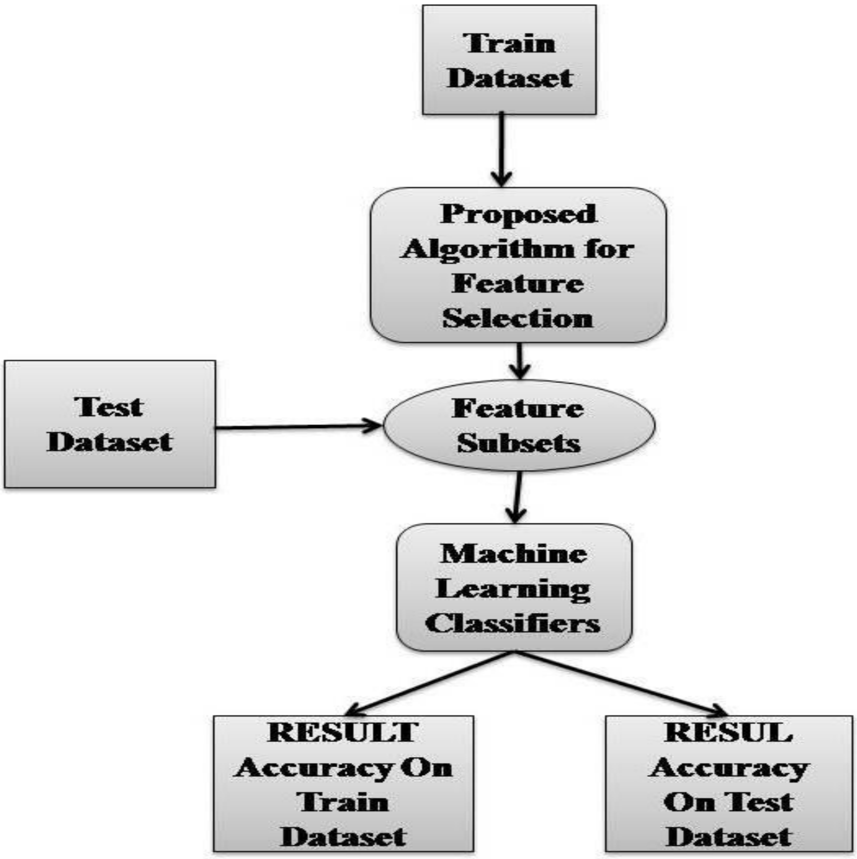
Train Dataset
Proposed Algorithm for Feature Selection
Test Dataset
Feature Subsets
Machine Learning Classifiers
RESULT Accuracy On Train Dataset
RESUL Accuracy On Test Dataset

Figure 1- The process of feature selection to knob feature subsets using ML algorithm.

2. Meta-heuristic Approaches for Feature Selection

Meta-heuristic approaches are defined as high-level methods for solving optimization problems. Initially, this method selects sets of samples from large samples and decides some assumptions to solve optimization problems [12]. With the help of stochastic optimization, the results of Meta-heuristic approaches are computationally best solutions, which results in less time.

Some of the methods included in Meta-heuristic approaches are:

- Genetic Algorithms
- Particle Swarms
- Ant Colony Optimization
- Memetic Algorithms
- Simulated Annealing, and so on.

3. Selection of Feature Using Different Algorithms

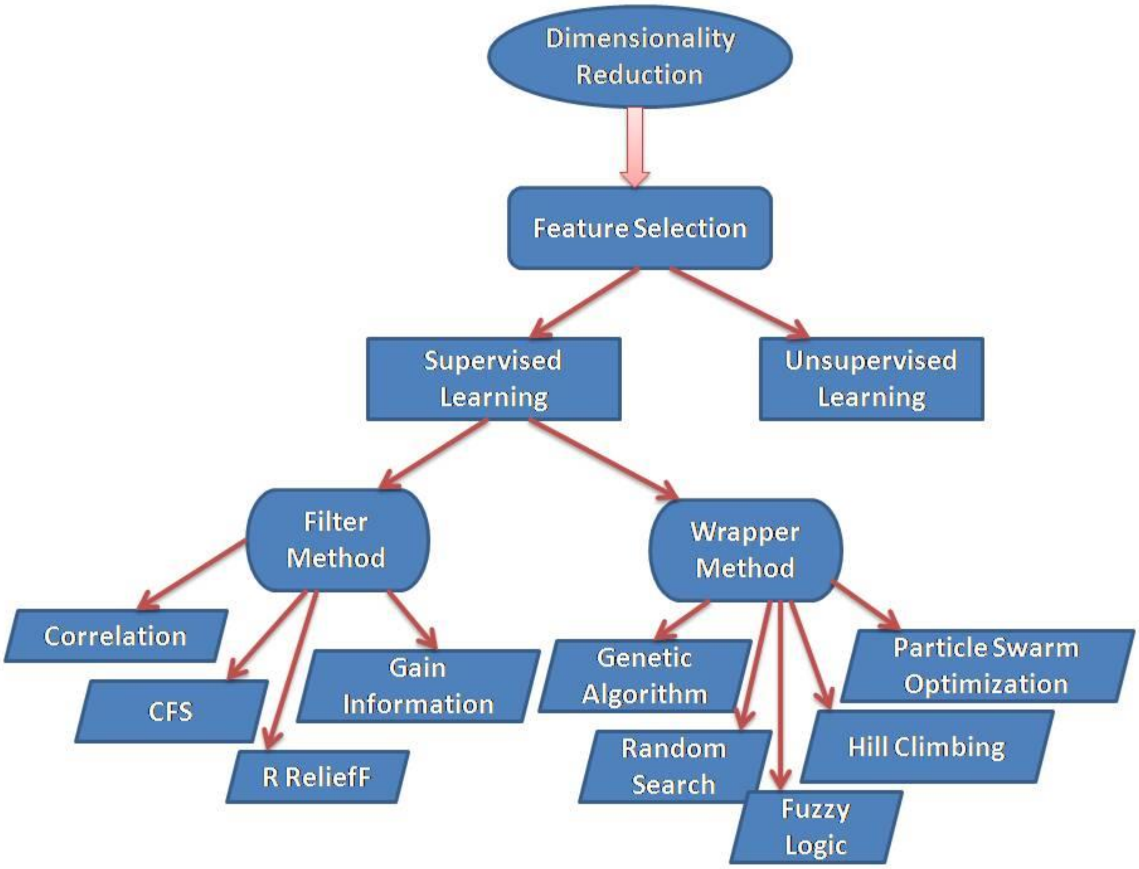

Figure 2- Different algorithms used for feature selection.

3.1. Feature Selection using Filter Technique

Feature selection using a filter technique is another approach. The technique used here is the correlation strategy. Correlation is nothing but a relation between two variables. [13] An element is considered as a functional element if the correlation between that feature and its class is more enough to say it is compatible with that class and the compatibility of other features does not reach that level. Those features can be predicted by other compatible features, and then those features are named as best features for feature selection and for applying in the classification. This technique helps in decreasing the elements and selecting the appropriate ones. Here, the most suitable technique of correlation measures the linear correlation coefficient and the variations of *least square regression error* and *maximal information compression index.* The main benefit of using linear correlation is that it eliminates zero correlative features to class, and then the selected elements redundancy can also be decreased by this method. The limitation of this technique is that all the elements should be linear and should contain numerical values. Using this technique, it is challenging to obtain compatible features from Non-linear features. With correlation, entropy and information gain methods are also used to overcome limitations. Entropy calculates the haphazardness, whereas information gain is calculated as the decreased amount of entropy of a feature R that reflects additional knowledge about the R feature specified by S, where R and S are two different features. Information gain demands two elements. Using the symmetrical uncertainty (*SU*) [14] technique, the best measures of features are calculated and the brink value of SU is obtained. Then, the next F-correlation technique is applied with brink SU value to identify the most compatible features. At some point, it is noticed that correlated variables are related to some other variables at some other aspects of classes. Therefore, to keep features compatible, another concept is introduced as a predominant correlation. A variable is said to be the prevailing variable if feature R_i (with R and S) does not contain any R_i. The algorithm implemented is FCBF (Fast Correlation-Based Filter) to filter the obtained compatible features for classification. In this technique, two steps are involved. The first step calculates the SU and selects features, and the second step is to list the selected features. In FCBF, ranking is performed based on iterations, and time complexity is calculated. The total time complexity of this technique is $O(MN \log N)$, where M defines the number of objects in the dataset. To display the result, in comparison to FCBF, the different techniques used are feature weighting algorithm, ReliefF, and searching strategies. After comparison, it can be said that the accuracy obtained by FCBF is much higher than that obtained by those techniques. FCBF results in high compatibility and exactness of selected features.

3.2. Feature Selection using Genetic Algorithm

Variable selection helps in decreasing the estimation cost by minimizing the bigness of data and increasing the forecasting practice and the accuracy of the patterns by eradicating incompatible and noisy features. Day by day, different feature algorithms are emerging just to give the best optimal solutions [15]. Using GA, multiple selections can be performed. The main goals are to apply the algorithm, find a minimal subset, and get more accuracy on any classification algorithm. In this work, a genetic algorithm is used with three different element selection techniques, which are the entropy-based feature elimination [16], *T* –statistics feature elimination, and SVM-recursive feature elimination (RFE) [17]. These techniques help in obtaining candidate elements which are further given to GA to find out the minimal subset. An amalgam technique is used to find out element subsets, and this technique involves GA with some rationalization algorithms, such as decision trees, artificial neural networks, and Naïve Bayes. But in this paper, a SVM classifier is used along with GA [18]. GA results in a combination of multiple elements and finds out minimal element subsets. Initially, by applying the element selection methods, an element pool is created, which consists of subsets of features. Then, next to that pool, GA finds out the optimal solutions and searches elements from the pool that is considered as population. Each element of this population is generated using a fitness function (randomized algorithm). The fitness function consists of two elements, one defines the weighted accuracy from the classification technique, and the second establishes the weight size from element subset or feature subset. With this, a new feature set is achieved by GA and its operations, like selection, crossover, and mutation. In the end, it is modelled to get a decreased feature subset with high accuracy on classification techniques. In this feature selection process, 3 techniques are utilized; the first two are from the filter and the third one is from the wrapper; they are entropy [19], T-statistics [20], and SVM, respectively. With the help of entropy, features are ranked based on randomness. If an element consists of less randomness, then that element is treated as a compatible element for selection. Elements are ranked in a descending order. In *T* -statistics, two samples are considered and the statistical distinctness of both is identified. Then, elements with the higher score are selected. Thereafter, SVM is applied with optimal brain damage (OBD) [21] strategy. Here, data is trained to SVM by considering element eradicating criteria and approximated by OBD, then ranking is done for all elements, and those elements with less ranking are removed. Two samples, named curse of dimensionality and curse of dataset sparsity, are taken to acknowledge the accuracy test. Cross-validation with 5 folds is applied using the Leave-one-out cross-validation. Then classification is tested using SVM. After applying SVM, it is noticed that, from 3 techniques, only the last 2 techniques result in best accuracy with 16 features, with precision of 98.3%, entropy of 64.5%, and T-statistics value of 88.7% are obtained. Then, testing is performed using GA for different population sizes, where the results are obtained with a lesser number of features and 100% accuracy.

3.3. Feature Selection using PSO

Another estimation of element selection is using a rough set approach and PSO. Rough sets are used for the identification of attribute selection. Due to the huge amount of emerging attributes, it is difficult to identify an application. Reducing the total number of incompatible characteristics is a task of feature selection. Rough set approach is a method of identifying and selecting variables [22]. Due to the increase of noisy, abundant, and incompatible variables, which are misleading, the accuracy has become the most noticeable problem in the real world. These noisy, abundant, and contradictory data are eliminated by applying the roughest method. Rough set selects only those variable subsets which can predict the decision of the initial element set. The primary purpose of the Rough Set is to obtain variable subsets, which results in high accuracy on classification [23]. Using a rough set, the minimal subsets were found that also eradicate NP-Hard problems. The rough set strategy is further divided into two types, which are: 1. hill-climbing (or greedy) methods, and 2. stochastic methods [24]. Rough set theory [25] is a new analytical method to treat exaggeration, ambiguity, and unpredictability. Rough set results in the estimation of an ambiguous approach with a couple of definite strategies, said as a lower and upper estimation. Lower estimation deals with the domain object, which gives certainty of belonging to the subset, whereas upper estimation deals with objects which have the circumstance of belonging to subsets. A set is noted as rough when both lower and upper estimations are not alike. The convenience associated with a rough set is that it does not require any additional data, i.e. the original data is enough. Granularity structure of the data helps in searching feature selection in this approach. Particle swarm optimization (PSO) is a metamorphic computing strategy developed by Kennedy and Eberhart in 1995, based on the nature animals, like birds and others. This algorithm is inspired by the movements of natural animals. Using rough set, the variable subsets are calculated, and then this PSO algorithm is initialized by considering a random population. PSO consists of a space where each population is treated as a particle. From all random particles, the efficient particles are identified and named as gbest particles. Different particles consist of different spaces, which determine the velocities of each particle, which is found out by initializing the population randomly with some velocities, and then the fitness positions are identified using a fitness function. Here, the fitness function is nothing but the subsets of a rough set. Those values are included in the PSO strategy. Then, based on PSO, position updating is done by considering particle velocities and flying capacities of particles to find the gbest outputs that should be best and optimal. By considering the maximum velocity limit, the changes in position are done. This algorithm is tested on 27 datasets by applying the LEM2 algorithm, which is considered to show the classification result. 10 folds are enforced to identify the accuracy of the classification technique. Classification differentiation is shown by comparing 5 different algorithms named POSAR, CEAR, DISMAR, GAAR, and PSORSFS. In this 5 hill-climbing or greedy algorithm, the methods employed are POSAR, CEAR, DISMAR, and the stochastic methods, which include GAAR and PSORSFS. The comparison of POSAR, CEAR, DISMAR, GAAR, and PSORSFS shows that the inertia obtained by PSO is much better and more optimized as compared to all other algorithms. Furthermore, the stochastics algorithm of PSO shows better results as compared to the GA.

3.4. Feature Selection using BPSO

This is another level of features selection in large dimensional biological datasets. Gene expression DNA data is considered to the variable selection here. A large amount of data is emerging day by day in the medical science sector. Here, an improved version of PSO, i.e., BPSO (Binary Particle Swarm Optimization), is considered [26]. Hamming distance is also used as a calculation of the distance between features and identification of the best features of the Microarray data set of DNA sequence. A fitness function is also included to know the best solution. Already, PSO based measures are calculated [27]. Hence, the new, improved method is the binary PSO. Initially, micro clusters of data are pre-handled to eliminate incompatible and superfluous data, then discretizing is applied followed by Binary PSO. The strategy is loaded with indiscriminate population swarm particles in an open space to identify the local leading and universal leading solutions. Based on the particle's participation, the finest fit solutions are obtained. PSO is converted to BPSO when the particles are considered as pairs of 0s and 1s. The velocities for BPSO are mapped between the interval of [0, 1]. In this topic of binary conversions, a distinction table is used, with a binary matrix where rows are taken as the objects pairs [28] and columns are elements in E. Using this d-distinction table, a minimal subset of columns is obtained from N set of columns and rows of object pairs. With the help of distinction, table calculation cost is also minimized, and size is also reduced. 0 is assigned for the entry of a pair of attributes, and 1 is assigned to matrix consistency to categorization determination. A fitness function is used with F1 and F2, where F1 says about the total count of elements, and F2 says about which element to take forward based on object pair. The results obtained from the fitness function are as follows: F1 is nothing but the results are prospect of praise or approval for holding the minimized count of features, and F2 results claim the degree to which the prospect can recognize amid pairs of objects from the d-distinction table.

BPSO is taken, in addition to Hamming distance, to obtain the minimized number of features, and it is said as the BPSO-HD technique. Hamming distance is mainly used for position updating of velocities. This strategy is initialized by a random population to obtain Lbest and Gbest solutions for position updates with velocities of each particle. Negatives numbers are not produced here. That is why it can be said 50% of a comparison is minimized, which leads to optimized computation results. Two binary strings are considered to represent X and Y. The result from the dissimilarity between X and Y is also a binary result. Following the implementation of BPSO-HD, a minimum elements subset based on the Hamming distance measure for big dimensionality gene expression data is produced. The classification result should be evaluated using the acquired feature subsets. In this step, three different cancer datasets are taken with large feature samples. From that data, half is assigned to train the data, and another half is assigned to the test data set—the three datasets named Colon, Lymphoma, and Leukemia. The initial features and the reduced features count that range from 2000 to 1102, 4026 to 1867, and 7129 to 3783, respectively reduced, for the three datasets. The obtained features are applied for classification techniques. In this K-NN classification, the technique is used as a classifier K-NN and results in 100% exact outputs for these three datasets. The accuracy is tested for distinct values of K, say K=1, K=3, K=5, K=7. The results for three datasets for distinct K-values are 90.25%, 92.36%, and 94.74%, respectively. Performance comparisons with the GA and the Non-dominated Sorting Genetic Algorithm (NSGA-II) are shown later. For all K values, the result generated for the colon and leukemia datasets and lymphoma cancer are 100 percent, when K=1 is the next K value, then the result is near to the value of NSGA-II. Finally, it can be said that using BPSO-HD resulted in minimal element subsets that obtained accuracy, as tested using the K-NN classifier on three different datasets of the real-world. Feasibility and effectiveness are obtained for the proposed method.

3.5. Feature Selection using Improved Variant of BPSO

This is another algorithm for element selection with the help of a further improved version of BPSO for gene expression data. Day by day, the emerging of a tremendous amount of gene expression data, especially in the medical field, is creating problems in identifying a particular symptom, with very much time being consumed. To eradicate these obstacles, different strategies are being implemented. Here, we discuss another method of picking elements from a large amount of data to obtain consistency and compatibility of elements [29]. The newly implemented method is introduced to obtain the speed rate of the procession of data, minimizing predictive error count, and avoiding incompatibility of elements that occur in the investigation of a huge number of genes. Particle swarm optimization is the basic version. It is profitably implemented in many fields and has many applications. Eberhart and Kennedy proposed the binary PSO (BPSO) for variables of discrete binary data [30]. The PSO functions based on lbest and gbest as fitness values. As soon as when gbest comes in local optimal condition, each particle search in that same area results in stopping of superior outputs of classification. So, the IBPSO is introduced to eradicate problems brought by the gbest solution, through the resetting of the gbest value, which results in optimized and superior classification outputs. IBPSO is initialized by assigning binary values of 0 for non-selected elements and 1 for selected elements. The utilized measure of the fitness of the subset with LOOCV (leave-one-out cross-validation) strategy and K-NN classifier is taken to find classifying accuracy with k=1 for the nearest neighbor, i.e., 1-NN. The 1-NN values are computed for all the datasets by bestowing the LOOCV strategy. In this method, one object from all initial samples is picked for validation data, and leftover objects are assigned for training data. The process is repeated so as to assign each object once for validation. This method is applied to 11 datasets. For a single-particle value of fitness, pbestp and best fitness values are used. Each particle within a group of pbestp is represented as gbest, i.e. global fitness value. After obtaining pbest and gbest values with velocities, the position updating process starts to find out a minimal subset of elements. The selection of element subset does not only mean to decrease the count of genes; instead, the ones that increase accuracy and optimize the cost of classification are selected. The classification accuracy obtained by this method gives 100% optimized and best results. IBPSO strategy shares similarity measures with many different EC procedures, like BPSO, PSO, GA, and NSGA-II. Comparing IBPSO with EC techniques like GA shows that, unlike GA, IBPSO does not consist of crossovers and mutation techniques for similar populations. The classification accuracy is improved by 2.85% compared to other EC techniques.

3.6. Feature Selection improved variants of BPSO

Feature selection is based on binary particle swarm optimization (BPSO) with two multi-instance elements-picking methods of classification [31]. One is the multi-objective binary PSO using the idea of NSBPSO (non-dominated sorting BPSO) and the second is the multi-objective binary PSO using the ideas of CMDBPSO (crowding, mutation, and dominance BPSO). This feature selection using improved variants of BPSO is the first-ever study of element picking for a filter-based strategy using multi-objective BPSO. Filter methods use analytical aspects for data computation and element picking autonomous of classification or learning techniques. Using this technique has developed 2 measures for information, using entropy and mutual information, and 2 steps for multi-objective BPSO, which are NSBPSO and CMDBPSO.

When there is more than one ambiguous objective, selecting a consistent objective causes a difficult problem in multi-objectives. Optimization of multi-objectives is described by decreasing or increasing multiple objective functions so as to find out all logical objective functions. As an example to illustrate this, consider three objective functions, say a1, a2, and a3, where a1 is influencing both a2 and a3, and a2 is not influencing a3, and a3 is not influencing a2. So, a2 and a3 are called trade-off results of each other. Whenever a result is not affected by any other result, then that result is known as a Pareto-compatible result. In element picking, the main task is a two-objective function; one is to minimize the elements, and the second is to get a high and accurate classification rate. The two information scopes used are mutual information. Entropy describes the ambiguity of initial variables, while mutual information describes the information shared between two initialized random variables, i.e. one variable P can determine how much information about the Q variable is given by mutual information. As mutual information describes the relevance and compatibility of random attributes, it can be used as element selection for the filter-based method. The method based on mutual information (MI) is the BPSO-based filter feature selection technique (BPSOfsMI), which helps in maximizing applicability and reducing repetition among elements. MI is difficult to apply for complicated variables because it only analyzes two variables. As a result, the element that reduces entropy in information is employed here. Entropy can calculate the applicability of variables in a numerous-way approach for complex data, identify the applicable elements, and minimize the repetition elements. Thus, the method proposed as a fitness function is a single objective filter feature selection algorithm (BPSOfsE) [32].

The equations of both BPSOfsMI and BPSOfsE are treated as two fitness functions. To describe repetition and applicability, a new variable is assigned as α, which ranges between 0-1. It is also said that applicability is more important than the reduction of repetition. In BPSOfsMI and BPSOfsE, binary strings considered are 0 and 1, where 1 denotes the selected feature and 0 denotes the unselected feature. After showing that BPSOfsMI and BPSOfsE are sufficient for feature selection, but weights for fitness function should be predefined, new algorithms are proposed based on the PSO technique; i.e. PSO is good for a single instance but it should also perform well for numerous instances. As a result, gbest best leader is found from a set of non-dominated solutions using this criterion. NSGA-II was merged with PSO to evolve multi-instance PSO strategy and gain the best optimization outputs. In that work, binary numerous-instance PSO schema are developed as NSBPSO for filter element picking. By considering NSBPSO, two new numerous-instances feature picking strategies are achieved as NSfsMI and NSfsE. NSfsMI and NSfsE are achieved to find out applicability and repetition on features and their class labels. The main objective behind these techniques is to make use non-dominant sorting techniques by selecting a gbest and updating for every particle. Under every iteration, this technique determines non-dominated results in the swarm and compute cluster distances, and a gbest from the least clustered result is selected randomly. Then, all particles are imitated to a union. After identifying gbest and pbest, the new acceleration and the new location for every particle are computed, and a new location is added to the union. These two instance ranges of each particle are calculated, and then applicability is appraised by CMDfsMI and CMDfsE. Explicitly, the union of the non-dominated front is named as the first non-dominated front, which is later removed from the union, then the non-dominated results are named as the second non-dominated front. In this way, some phases of non-dominated fronts are determined by repeating this process. Then, the updating of the swarm is shown by the next iteration. Explicitly, particles are picked from top phases of non-dominated fronts, i.e. starting from the first front. Based on a certain strategy, e.g. if the count of results required is higher than the count of results in the present non-dominated front, then all results are summed to the next iteration. Diversely, outputs of the present non-dominated front are ranked based on cluster distance, and outputs with more rank have amounted to the next iteration. The process is repeated until completion or the expiry criterion is met.

The achievement of NSBPSO has a limitation of easily failing to diversification of population, all along with the evolutionary method. There are chances of the emergence of many new particles due to iterations, merging, and updating of particles. In order to eradicate this limitation, another groundwork is developed using binary Nemours-instance PSO, said as CMDBPSO, where C is referred to as crowding, M is referred to as mutation, and D is referred to as dominance. Based on the CMDBPSO approach, again, two methods are developed as CMDfsMI and CMDfsE. CMDfsMI is used to compute applicability and CMDfsE is used to calculate applicability between picked elements and class labels. The main use of CMDfsE and CMDfsMI is decreasing the count of elements and increasing the applicability of picked elements and class labels. A crowding equation is added to identify which non-dominate output should be employed in the leader's set, and a bit-flip mutation operator is selected to increase search capability and constant the diversification of the swarm.

To show the result of these algorithms, 8 datasets are picked with a large number of features present in them. From each data, 70% is treated as a training set and the remaining is treated as a test set. The first training is performed to pick element subsets, and then the test is done by a classification strategy to find the accuracy of selected elements. Here, DT classification technique is selected to find accuracy. BPSOfsMI and BPSOfsE are evaluated using 5 distinct weights in the fitness function. The classification result of these 2 techniques is slightly worse than that of using all elements. For different α values, different performance is noticed. When α is high, classification performance is high [33]. Outputs of NSfsMI are described using two datasets that contain elements of a small number and obtain less classification error rate than using all features. For different elements with different fitness functions, the different accuracy rate is obtained in all CMDfsMI, NSfsE, and CMDfsE methods. In one non-dominated output, the picked elements are 11 from 22 features, and it is noticed that the error rate is decreased from 33% to 25%. This also indicates that NSfsMI is an efficient technique for the numerous-instance method which by default evolves the element feature subsets to decrease the number of elements and increase classification performance.

The results obtained from CMDfsMI consist of 2 or more results that pick less count of elements and achieve the best classification result than that obtained using the full elements of all datasets. An example describes a dataset CMDfsMI that selects 1 element. It is noticed that the classification rate of error is reduced from 33% to 28%. This indicates that CMDfsMI, as a numerous-instance technique, is efficiently explored to Pareto front. It decreases the classification rate and the number of elements of classification.

Now, the comparison is made for BPSOfsMI with NSfsMI and BPSOfsMI with CMDfsMI. It is noticed that NSfsMI achieves much better classification rate than the BPSOfsMI, and the elements selected by CMDfsMI achieve very well compared to elements selected by BPSOfsMI. So, it can be said that, when treating mutual information as the fitness function, the best classification result is obtained for more features. It performs well for non-dominate element subsets and gives higher performance than BPSOfsMI.

The results obtained from NSfsE are more than 1 result, which leads to the picking of a lesser count of elements and achieving the best classification achievement. For an example data set, it is noticed that the error rate of classification is decreased from 33% to 25%, just by picking 9 elements of a total of 22 elements. This shows that NSfsE as emerging criteria can, by default, evolve sets of element subsets and likewise decrease the count of elements and best classification achievement for all aspects.

The results obtained from the usage of CMDfsE suggest acquiring element subsets from a smaller number of elements instead of considering all features. CMDfsE increases the achievement of classification just by picking 25% of overall elements. It automatically reduces factors and increases the classification rate.

Now, the comparison is made between NSfsE with BPSOfsE and CMDfsE with BPSOfsE. It is noticed that NSfsE performs a superior classification than BPSOfsE. Moreover, the count of elements is slightly higher. CMDfsE picks a smaller number of features than BPSOfsE and achieves a superior classification rate. Based on this comparison, it can be said that numerous-instance techniques are better than single-instance techniques, i.e., BPSOfsE.

Finally, we can estimate using mutual information and entropy. It can be stated that BPSOfsE, NSfsE, and CMDfsE, which use entropy, show much superior classification results than BPSOfsMI, NSfsMI, and CMDfsMI, which use mutual information. Nevertheless, it can be a fact that numerous-instance techniques are always better than single-instance techniques and they give higher classification result. NSfsMI and NSfsE have some limitations, like the quickly falling diversification of swarm due to position updating. CMDfsMI and CMDfsE have the capability of overcoming this limitation. Hence, finally, it is concluded that the performance of CMDfsMI and CMDfsE is superior to NSfsMI and NSfsE.

3.7. Feature selection using mutation operator and decision tree in BPSO.

This process involves the selection of variables using the BPSO algorithm as an operator, i.e., mutation operator, also in addition to a decision tree. A dataset of features was established by using and organizing 6000 email data [34]. The organized features consist of 3 distinct types, where the first one consists of 48 similar words; the second consists of 6 characters and the third is a 3 capital-run-length alike features. From all the 6000 emails, each email includes all the 3 types, i.e., 57 dimensions of elements. For evaluation, the instance function is used as a wrapper method, since it produces more classification accuracy rate. It is assessed using an instance function and a classification strategy, as there are vast numbers of emerging classifier techniques by using a tree-based classifier, as it is simpler and more understandable with if-then prototypes. The output of the decision tree is a tool which supports decisions with the representation of graphs and their consequences [35]. Mislabeling spam has distinctive symptoms. Spam is set apart as Non-spam will leave a weight on the clients who need to peruse through and erase it. In any case, a non-spam that is set apart as spam is typically trailed by the grouping of programmed erasure, causing the client to lose significant email or, with programmed exchange to the spam box, the client will be resulted in all probabilities. Hence, we have to utilize the disarray framework and cost network to portray extraordinary sorts of blunders and to gauge how genuine they are, separately. Here, a description is provided for element selection for textual data and projected combination of the element selection method with decision tree as an amalgam system. Giving substantial meaning to customers with a huge accuracy rate of classification is the main convenience of this model. Along with this, this method also says about the segregate occurrence of 2 bugs. One bug is shifting a piece of information to the server directly by anticipating spam as a piece of regular information, while the second bug is anticipating a piece of regular information as spam certainly to be destructive. This information is trashed to spam carton without even informing the customer about the transferring. Due to this movement, effective information can be lost and sent to a spam carton. These are the 2 bugs to be considered for elimination. Here, two integers are taken to define spam and non-spam information. Positive integers are treated as spam and non-positive integers are treated as non-spam information. The cost matrix measure is considered to balance these two bugs. This matrix technique helps as best quantity increases in the complete bugs, along with the decrease in the cost of bugs. Training of the dataset of the decision tree is done using an emerging method called a C4.5 strategy, instead of considering ID3. Both of these techniques make use of the entropy strategy. In the C4.5 technique, effective samples are split into subset samples, using variables, and are normalized by making use of the information gain method, which helps in selecting a variable for breaking. Variables with high information gain are considered for making decisions. Now, cross-validation is applied as it is the easiest, understandable, and takes complete data for training and validation. Here K-folds cross-validation is applied by making K partitions of complete data set, which are iterated K times. This process finally gives decreased bug rates after K time analysis. Subsequently, training and validation are performed.

The search strategy used is BPSO with a mutation operator. Dissimilar to canonical PSO, BPSO position updating is barred inside the Hamming distance space. The mutation operator is further merged so that BPSO can explore the in-depth area of search space. This can also be said as MBPSO. All together, to explain the capability of a capital-run-length type of features, the Kolmogorov–Smirnov hypothesis test [36] is used to obtain an instance function.

Comparisons are made for the proposed method with existing methods of identifying spams, like ANN and SVM. The effectiveness values obtained are 91.08% and 97.70%, respectively, for both ANN and SVM. Sensitivity, precision, and effectiveness are obtained. Using MBPSO, the least 7 elements are obtained with 94% effectiveness. Thus, wrapper methods are better than filter methods for classification effectiveness.

3.8. Feature Selection using Neural Networks

This method for selecting elements was developed using entropy and a classifier, named the neural network [37]. Breast cancer datasets are considered nowadays as various types of breast cancers are emerging. It is becoming challenging to identify the type of cancer. For this purpose, element selection criteria are considered to find and select optimal features that describe the total class of that dataset. A different classification technique is already applied to find the accuracy of the algorithm and has also obtained much better results. In this work, ANN (Artificial Neural Network) classifier is used to eradicate features and find the best accuracy results. The dataset used is the Wisconsin breast cancer

(WBCD) dataset. Based on entropy, the Sequential Backward Selection (SBS) technique is also proposed to calculate interdependencies of variables for selection from unsupervised data. Iterations are performed using SBS, and entropy is calculated after eliminating elements from the complete set. The elements to be removed are those that give less entropy. This process of iteration is repeated until the importance of all elements is obtained. The used model is ANN, which is a biological computational neural network model. This technique consists of a complimentary gathering of artificial neurons and process data using a learning technique for calculation. For training purposes, two algorithms are considered, namely LM and PSO, to increase the result of BPNN (back-propagation neural network) learning achievement.

BPNN approach accommodates the abrupt extraction for training and updating the weights; the drawbacks obtained here are eliminated by considering training techniques, i.e. LM, which gives integer outputs for decreasing problem functions. LM results are more efficient than those of the Gauss-Newton algorithm (GNA) and gradient descent methods.

BPNN uses the PSO, which has iterative performance and finds the local and global best solutions. Based on those particles' best solutions in space, the elements start updating their positions and finding the efficiency of using the technique. PSO is applied for different task achievements, like the updating of finite features and the training of ANNs. WBCD [38] is the used dataset, which consists of 2 chunks, where the first is cancerous, and the second is harmless, i.e., non-cancerous. Each tissue of the breast consists of 9 features for each case; there are in total 241 and 456 cases, respectively. In the dataset, 16 missing ranges related to cancerous data and 14 non-cancerous data are taken. The UCI machine learning archive for WBCD was used and it involved BPNN with LM and BPNN with PSO.

After eliminating the missing range, based on importance, the remaining elements are sorted. A ROC graph strategy is taken to show the visualization and achievement of the binary classifier, which shows optimal solutions and eliminates substandard solutions. After applying the binary classifier obtained, 4 possible results are obtained, and the ROC curve is drawn.

The classification technique is trained and tested K times using K folds Cross-Validation. In this, 10-fold validation is applied and accuracy is noted. The effectiveness obtained by BPNN with LM and BPNN with PSO is recorded, and the accuracy obtained is 97% and 99%, respectively. Also, the AROC curve is drawn, showing the accuracy of 98% and 99%, respectively. There is an increase in classification accuracy by 0.32%, which easily demonstrates the identification of breast cancer and other cancers.

3.9. Feature Selection using Ant Colony Bee

An emerging optimization strategy is said as an ant colony bee for feature selection. This technique directs to perform large-quality optimization outputs with the best computing time. It helps in obtaining effective element subsets by performing iterative action. In this algorithm [39], optimal solutions are found by considering some agents, called ants, who talk with each other and based on their personal sharing, this aura works. Every ant picks some elements based on personal experience and iterations. Due to this nature of last stage information and repetitions, this strategy scores overall best effective results. Here, filter-based techniques are used instead of learning techniques. Based on the probability of each element, elements are selected by eliminating redundancy. Here, all the filter-based techniques are is discussed, like Information gain, Gain ratio, Symmetrical uncertainty, Gini index, Fisher score Term variance, Laplacian score, Minimal-redundancy–maximal-relevance, Mutual correlation, Random subspace method, and Relevance-redundancy feature selection.

The proposed method is initialized and started by performing some number of iterations for each node, and then all the ants start replacing their positions randomly. Traversing is done by ants, by considering nodes of graphs using probability technique, said as state conversion rule, until the iteration benchmark is reached and the traversing is carried out. Stopping the iteration is done either when the count of nodes required is picked, or some bugs occur. The state conversion rule helps to pick values that have low similarity and high effectiveness of features. An array is assigned to keep a count of the number of elements that are selected by ants, named as Feature Counter. After iteration is performed, updating is done using a global updating rule. This process is repeated until the required criteria are met and filtered elements are obtained, then selected features are sorted from high to low order. While performing iterations, these ants use both pheromone information and heuristic techniques.

This developed method uses three sectors. The first is to perform computing similarity between features, where each element n is dependent on class c of dataset and complexity is said as O(cn2). The second is to find the probability of each element by performing a maximum count of iterations, where the ants select the next feature based on rules. Also, the ants walk analogously and the complexity is O(ncMaxnAnt). Then, the last section is to select the most appropriate and optimal elements and sort them based on probability ranges.

This proposed method did not use any kind of learning algorithm for the selection of element subsets, which results in obtaining less computing cost compared to wrapper techniques. Also, due to the iterative nature of the method, its complexity increases. It is slightly expensive as compared to filter techniques. Nine datasets are used to analyze the proposed method. These are breast datasets [40] that are computed from digitized images, which consist of 279 features. By applying classifying accuracy of the validation test, missing range elements are removed and the best accuracy is obtained.

This model is independent of classification techniques, but the accuracy of classification is obtained and calculated with three techniques, which are Support Vector Machine (SVM), Decision Tree (DT), and Naïve Bayes (NB). The achievement count is calculated using the classification bug rate. First, the datasets are trained randomly, then testing is performed on the datasets. This technique is evaluated by considering different classification strategies, but it should be acclaimed that filter strategies are independent of the classifier and applying the classification. That is why only the execution of how elements are selected is shown in this work. This developed method is captured as the second shortest bug rate compared to any other unsupervised techniques. From all methods, UFSACO gets the best efficient results with an error rate of 23% and is placed in the first position among all methods. It gives admirable results when applied to SVM classifier. It can be explained by an example of 3 elements selected. The performance of UFSACO is about 11% as bug rate. For other methods like LS it is about 16%, MC is 13%, RSM is 28%, TV is 39%, and RRFS is 17%. Each dataset is examined for each method, and finally, best-optimized results are identified in ant colony optimization (UFSACO) and the best features are selected. UFSACO is applied to three classifier techniques and the best-optimized results are marked. Based on this, we can also say that UFSACO performs best on unsupervised data than on supervised data.

4. Comparison

In this paper, the selection of features is described using different algorithms and two different methods of feature selection, which are the filter-based feature selection and wrapper based feature selection. When considering the past many years, we can notice the use of different algorithms in different fields, with a noticeable growth/increase/decrease being marked. The use of nature-inspired algorithms for feature selection is experienced as a major role. Both filter and wrapper methods were used over years for feature selection. Here, the emerging methods which have been used over the past 10 years for feature selection, like GA and PSO, are compared . From the graph, it can be noticed that the use of both algorithms has faced few fluctuations over time. Genetic Algorithm was popular during the 2010s whereas Particle Swarm Optimization Algorithms was mostly used during 2016, with the highest amount of publications. The use of these methods is marked in the below bar in the graph.

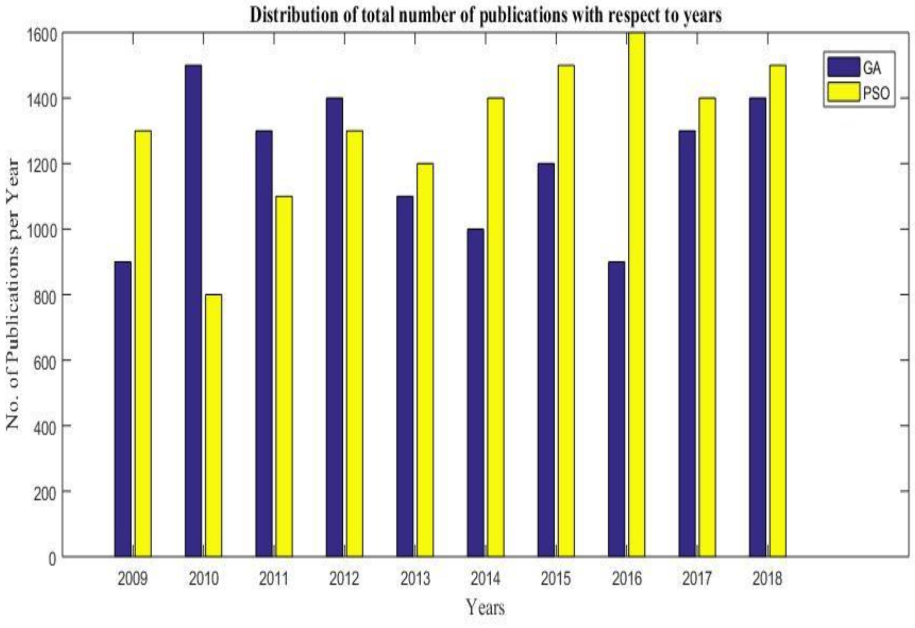

Figure 3- The distribution of the total number of publications with respect to years.

5. Applications of Feature Selection

Feature selection methods can be employed in data pre-processing to achieve effective data reduction. This helps in finding accurate models of data [41]. Because, in most cases, a comprehensive search for an ideal feature subset is difficult, several search strategies have been suggested. The normal applications of FS are clustering, regression, classification, and dimensionality reduction. These applications can be applied to real-world problems, like Computer Vision, Image Processing, Bio-Informatics, Text Mining, and Industrial Applications. Table 1 presents the areas of feature selection applications, the sub-specialty associated with the area of application, feature selection methods that can be used in that particular area, popular datasets, types of evaluation metrics, and the best performance measures in the application area [42].

Table 1- The different applications of feature selection methods.

Area of Application	**Subspecialty**	**FS Methods which can be used**	**Popular Datasets**	**Metrics of Evaluation**	**Best Performance Measures**
Computer Vision	Image Classification	Relief (R), Sequential Floating Backward Selection(B), K-Means (K), Sequential Floating Forward Selection (F), Various Combinations Like R+ K + F/B	Cifar-10, Image Net, Imdb-Wiki Dataset, Ms Coco, Mpii Human Pose Dataset.	Neural Networks.	Depends upon the Size of the subdivision of features, R+K+B / R+K, /R+K+F.
Image Processing	Image Classification, Classification From	Random Search, Genetic Search,	Mnist, Open Images Dataset, Fashion-Mnist,	Based on Accuracy	Best First Forward, Best First Backward

	Mammographic Images	Best First Forward, Best First Backward, Bi-Directional With Naïve Bayesian And K-Nn Classifiers	Visual, Ms-Coco.		And Classifiers
Bio-Informatics	Gene Expression, Microarray Data Classification	Information Gain, Chi-Square, Relief, Twoing Rule, Oner, Sum Minority, Gini Index, Max Minority, T-Statistics, Sum of Variances, One Dimensional SVM, and SVM-Embedded	Bmc Genomics Supplemental Data, Cell Image Library, Ebi Array Express, Freije, Phillips, Ebi Protein Dataset.	Based on Accuracy And Stability	Information Gain, Chi-Square, Svm-Embedded, Relieff, Collaboration of All Strategies
Text Mining	Text Classification	Chi-Square, F1 Measure, Accuracy Balanced, Odds Ration Numerator, Bi-Normal Separation, Information Gain, Document Frequency, Probability Ratio, Odds Ratio, Random, Term Contribution, Term Strength, Entropy-Based Ranking, Iterative Feature Selection	Amazon Reviews Dataset, Enron Email Dataset, IMDB Dataset, Movie Lens Latest Datasets, Opin Rank Dataset.	Based on Accuracy, F-Measure, Entropy, Precision, and Recall	Information Gain, Bi-Normal Separation, and Iterative Feature Selection
Industrial Applications	Fault Diagnosis	SVM Wrapper, Entropy,	Wind Turbine Test Rig Dataset	Based on Accuracy	Global Geometric Similarity

		Neural Network Wrapper, Global Geometric Similarity Scheme			Scheme with Wrapper

6. Conclusions

In many areas, such as statistics, image processing, machine learning, text mining, data mining, pattern recognition, web mining, and gene microarrays analysis, the selection of features is recorded as an evergreen research subject with practical significance. This paper gives a clear image on what is feature selection, methods of feature selection, and most used Meta-heuristic algorithms to find the best fit subsets of feature sets. The results obtained using these optimization techniques lead to defining the best accurate solutions. It can also be concluded that the accuracy levels for feature subset selection using these algorithms are in the order of BPSO>PSO>GA. This work also presents the fact that the performance has been improved from the filter-based method to the wrapper-based method, along with various nature-inspired algorithms which can be applied to real-world applications; for example, medical diagnosis. These techniques are mostly used nowadays in medical fields and different biological areas. The paper sheds a light on the various application areas of feature selection. More emerging technologies are coming into existence to solve problems on feature selection.

References

1. Chandrashekar, G. and Sahin, F., **2014**. A survey on feature selection methods. *Computers & Electrical Engineering*, **40**(1):16-28.
2. Urbanowicz, Ryan J., Melissa Meeker, William La Cava, Randal S. Olson, and Jason H. Moore. **2018**."Relief-based feature selection: Introduction and review." *Journal of biomedical informatics*, **85**: 189-203.
3. Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H. and Nowe, A., **2012**. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **9**(4):1106-1119.
4. Palamidessi, Catuscia, and Marco Romanelli. **2020**. "Feature selection in machine learning: R\'enyi min-entropy vs Shannon entropy." *arXiv preprint arXiv:2001.09654* (2020).
5. Narendra, P.M. and Fukunaga, K., **1977**. A branch and bound algorithm for feature subset selection. *IEEE Transactions on computers*, (9):917-922.
6. Goldberg, D.E., **1989**. Genetic algorithms in search. *Optimization, and MachineLearning*.
7. Shi, Y., **2004**. Particle swarm optimization. *IEEE connections*, **2**(1):8-13.
8. Ding, Jianli, and Liyang Fu. **2018**. "A Hybrid Feature Selection Algorithm Based on Information Gain and Sequential Forward Floating Search①." *Journal of Intelligent Computing*, **9**(3): 93.
9. Reunanen, J., **2003**. Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, **3**(Mar):1371-1382.
10. Rathee, Seema, and Saroj Ratnoo. **2020**. "Feature selection using multi-objective CHC genetic algorithm." *Procedia Computer Science*, **167**: 1656-1664.
11. Rathee, Seema, Saroj Ratnoo, and Jyoti Ahuja. **2019**."Instance selection using multi-objective CHC evolutionary algorithm." In *Information and communication technology for competitive strategies*, pp. 475-484. Springer, Singapore, 2019.
12. Tzanetos, Alexandros, and Georgios Dounias. **2020**. "Nature inspired optimization algorithms or simply variations of metaheuristics?." *Artificial Intelligence Review* (2020): 1-22.
13. Yu, L. and Liu, H., **2003**. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)* (pp. 856-863).
14. Lin, Xiaohui, Chao Li, Weijie Ren, Xiao Luo, and Yanpeng Qi. **2019**. "A new feature selection method based on symmetrical uncertainty and interaction gain." *Computational biology and chemistry,* **83** (2019): 107149.

15. Tan, F., Fu, X., Zhang, Y. and Bourgeois, A.G., **2008**. A genetic algorithm-based method for feature subset selection. *Soft Computing*, **12**(2):111-120.

16. Estrela, Gustavo, Marco Dimas Gubitoso, Carlos Eduardo Ferreira, Junior Barrera, and Marcelo S. Reis. **2020**. "An efficient, parallelized algorithm for optimal conditional entropy-based feature selection." *Entropy*, **22**(4): 492.

17. Rtayli, Naoufal, and Nourddine Enneya. **2020**. "Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization." *Journal of Information Security and Applications*, **55** (2020): 102596.

18. Nafis, Nur Syafiqah Mohd, and Suryanti Awang. **2020**."The Evaluation of Accuracy Performance in an Enhanced Embedded Feature Selection for Unstructured Text Classification." *Iraqi Journal of Science* (2020): 3397-3407.

19. Ahmed, Waseem, MM Sufyan Beg, and Tanvir Ahmad. **2018**."Entropy based feature selection for fuzzy set-valued information systems." *3D Research* 9, no. 2 (2018): 1-17.

20. Wang, J.G. and Liu, G., **2002**. A point interpolation meshless method based on radial basis functions. *International Journal for Numerical Methods in Engineering*, **54**(11):1623-1648.

21. Chaber, Patryk, and Maciej Ławryńczuk. **2018**."Pruning of recurrent neural models: an optimal brain damage approach." *Nonlinear Dynamics* , **92**(2): 763-780.

22. Stańczyk, Urszula, and Beata Zielosko. **2020**. "Heuristic-based feature selection for rough set approach." *International Journal of Approximate Reasoning,* **125** (2020): 187-202.

23. Swiniarski, R.W. and Skowron, A., **2003**. Rough set methods in feature selection and recognition. *Pattern recognition letters*, **24**(6): 833-849.

24. El-Kenawy, El-Sayed M., Marwa Metwally Eid, Mohamed Saber, and Abdelhameed Ibrahim. **2020**. "MbGWO-SFS: Modified binary grey wolf optimizer based on stochastic fractal search for feature selection." *IEEE Access,* **8**(2020): 107635-107649.

25. Das, Asit K., Shampa Sengupta, and Siddhartha Bhattacharyya. **2018**. "A group incremental feature selection for classification using rough set theory based genetic algorithm." *Applied Soft Computing,* **65** (2018): 400-411.

26. Too, Jingwei, Abdul Rahim Abdullah, Norhashimah Mohd Saad, and Weihown Tee. **2019**. "Emg feature selection and classification using a pbest-guide binary particle swarm optimization." *Computation,* **7**(1): 12.

27. Ji, Bai, Xiaozheng Lu, Geng Sun, Wei Zhang, Jiahui Li, and Yinzhe Xiao. **2020**."Bio-inspired feature selection: An improved binary particle swarm optimization approach." *IEEE Access* , **8** (2020): 85989-86002.

28. Xue, B., Cervante, L., Shang, L., Browne, W.N. and Zhang, M., **2013**. Multi-objective evolutionary algorithms for filter based feature selection in classification. *International Journal on Artificial Intelligence Tools*, **22**(04):1350024.

29. Jain, Indu, Vinod Kumar Jain, and Renu Jain. **2018**. "Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification." *Applied Soft Computing,* **62**: 203-215.

30. Kennedy, J. and Eberhart, R.C., **1997**. October. A discrete binary version of the particle swarm algorithm. In *1997 IEEE International conference on systems, man, and cybernetics. Computational cybernetics and simulation* (Vol. 5, pp. 4104-4108). IEEE.

31. Azadifar, S., and A. Ahmadi. **2020**. "A Graph Theoretic Based Feature Selection Method Using Multi Objective PSO." In *2020 28th Iranian Conference on Electrical Engineering (ICEE)*, pp. 1-5. IEEE, 2020.

32. Kumar, Lalit, and Kusum Kumari Bharti. **2019**. "An improved BPSO algorithm for feature selection." In *Recent trends in communication, computing, and electronics*, pp. 505-513. Springer, Singapore, 2019.

33. Kumar, Lalit, and Kusum Kumari Bharti. **2019**. "A novel hybrid BPSO–SCA approach for feature selection." *Natural Computing* (2019): 1-23.

34. Sarhani, Malek, Abdellatif El Afia, and Rdouan Faizi. **2018**. "Facing the feature selection problem with a binary PSO-GSA approach." In *Recent developments in metaheuristics*, pp. 447-462. Springer, Cham, 2018.

35. Lee, Shin-Jye, Zhaozhao Xu, Tong Li, and Yun Yang. **2018**. "A novel bagging C4. 5 algorithm based on wrapper feature selection for supporting wise clinical decision making." *Journal of biomedical informatics,* **78** : 144-155.'

36. Aslam, Muhammad. **2019**."Introducing Kolmogorov–Smirnov tests under uncertainty: an application to radioactive data." *ACS omega,* **5**(1): 914-917.

37. Zhang, Zhihua. **2018**. "Artificial neural network." In *Multivariate time series analysis in climate and environmental research*, pp. 1-35. Springer, Cham, 2018.

38. Showrov, Md Imran Hossain, Muhammad Tanveer Islam, Md Dulal Hossain, and Md Shakil Ahmed. **2019**. "Performance Comparison of Three Classifiers for the Classification of Breast Cancer Dataset." In *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, pp. 1-5. IEEE, 2019.

39. Dorigo, Marco, and Thomas Stützle. **2019**. "Ant colony optimization: overview and recent advances." *Handbook of metaheuristics* (2019): 311-351.

40. Naseri, Norowzani S., And M. A. Shayegan. **2018**."Breast Cancer Staging By Using Tnm System And Ant Colony Algorithm." (2018): 56-70.

41. Venkatesh, B., and J. Anuradha. **2019**."A review of feature selection and its methods." *Cybernetics and Information Technologies,* **19**(1): 3-26.

42. Jović, Alan, Karla Brkić, and Nikola Bogunović. **2015**."A review of feature selection methods with applications." In *2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 1200-1205. Ieee, 2015.