10.48047/jocaaa.2024.33.08.40

Hybrid Firefly Model for Feature Selection in Supervised Learning

Shashikala B[1*], R. Saravana Kumar[2]
[1]Research Scholar, Visvesvaraya Technological University, Belagavi-590018, RC: Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology & Management, Bengaluru, Karnataka, India.
[2]Professor, Department of Computer Science and Engineering, Dayananda Sagar Academy of Technology & Management, Bengaluru, INDIA.
***Corresponding Author:** shashba@gmail.com

Abstract

The research idea presented in this article is to find the pertinent attributes of a dataset for supervised learning by using a Hybrid Firefly Model. The proposed approach is primarily based on swarm optimization namely "Binary Firefly algorithm" (FFA) and evolutionary Genetic algorithm . FFA , a modern meta-heuristic swarm optimization algorithm is collaborated with evolutionary operations for a better exploration of search space. In this proposed hybrid approach ,the movement of fireflies in the solution space is meticulous, hence to avoid reaching non-optimal solutions only the efficient fireflies are exploited to explore the solution space exhaustively in discovering the optimal solution. The fireflies are ranked based on the objective function which incorporates accuracy and correlation parameter. The least effective fireflies are made to evolve by evolutionary genetic algorithm. The optimal set of attributes from each method will be discovered and updated to global solution set. At the end of final iteration, performance evaluation is performed to prove the efficiency of the approach. The performance of this work is expressed in the form of accuracy score, correlation value and the size of selected attributes. Thus, the proposed approach shows a significant improvement in selecting an optimal set of pertinent attributes with good model performance.

Keywords— Feature Selection, Firefly, Accuracy, Classification, Genetic Algorithm, Correlation

I. Introduction

In the technology driven world, every field has shown its own progress using sophisticated scientific techniques. This scenario leads to data centric realm where researchers have a larger scope to play significant roles to utilize the raw data in different dimensions for different purpose by extracting the knowledge using machine learning algorithms. The major problem faced by any machine learning algorithms, which are primarily based on the quality of data, is to find those attributes of dataset which would give the desired output and this process of finding is termed as feature selection. There are many techniques to perform feature selection like evolutionary approach, swarm based approach, conventional deterministic approach and so on. In any of these approaches the primary task is discover those attributes which influence the success of a model in terms of the performance

10.48047/jocaaa.2024.33.08.40

measures like stability , accuracy and training cost . Since conventional methods require enough adaptability to lever datasets having larger number of evidence or tuples and extract efficacious outcomes .The various approaches and the techniques aim at simplification of models to make them easier to interpret by researchers/users, for shorter training times and to avoid the curse of dimensionality.

Swarm optimization algorithm like Particle swarm optimization (PSO), ant colony optimization (ACO), artificial bee colony algorithm (ABC), Simulated Annealing (SA) algorithm, Bacterial Colony Chemo taxis (BCC), Firefly Optimization algorithm etc., incorporate simple search method in the swarm of solution world .

This paper presents a methodology where feature selection is done using an evolutionary approach and swarm intelligent nature inspired Firefly algorithm.

II. Related work

Feature Selection

The researchers [1] have explained how important is feature selection especially in the data sets with many variables and features and experimented using the Random Forest machine learning technique , as it is observed to be a quite useful algorithm that can handle the feature selection issue even with a higher number of variables. Explained how feature selection simplifies the model and decreases the training time leading to better performance.

The authors [2] has proven that feature selection can be used for the data sets in medical field. They have experimented on predicting heart disease and examined the various performance factors to compare feature selection methods' effect on prediction algorithms and it was found with significant improvements in model performance .Also inferred wrapper-based and evolutionary algorithms improved models' performance from sensitivity and specificity points of view.

Evolutionary Algorithm

10.48047/jocaaa.2024.33.08.40

The researcher's [3] focussed on Evolutionary Algorithms (EA) and their types with real-life applications. An overview of different optimization techniques was clearly depicted and explained. Many researchers work towards the modifications of EAs to improve their computational performance and overcoming the challenges posed by it. For complex real-world problems, there is a need to optimize a function in a high-dimensional domain , where EAs in parallel mode will reduce the computational time. The challenges in EA approach was clearly explained like choosing the parameter values, defining good mathematical objective function , the stochastic nature leading to different result in different runs .

The authors [4] have given an insight EA , a population-based stochastic direct search algorithms that in some sense simulate natural evolution. The article explains about the various popular forms of EA like genetic algorithms, evolution strategies, evolutionary programming, and genetic programming. Also discussed about how EAs can be adapted to work well in case of multiple objectives, and dynamic or noisy optimization problems.

Thus it gives a guidance to use EA approach for feature selection where selection depends on various factors namely accuracy and correlation factors with other criteria too.

Swarm Optimization

The author has briefly explained about Swarm Intelligence (SI) and their characteristics [5]. Many researchers of various fields opt SI for many reasons. SI has been described as collective intelligence of groups of simple agents like the group foraging of social insects, cooperative transportation, nest-building of social insects, and collective sorting and clustering. Two fundamental concepts that are considered as necessary properties of SI are self-organization and division of labour which helps us in tackling complex problems .

This survey article was focuses on the performance factor of various Swarm Intelligence (SI) based approaches and aimed to provide a comparison among the well-known SI-based approaches , namely Genetic algorithm (GA), Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Differential Evolution (DE), Artificial Bee Colony (ABC), Glowworm Swarm Optimization (GSO), and Cuckoo Search Algorithm (CSA) . The results have proved variations of PSO and DE have performed well .

Firefly

10.48047/jocaaa.2024.33.08.40

In the article [6], the authors have explained about the Firefly algorithm (FA) as an excellent global optimizer based on swarm intelligence. They have highlighted the drawbacks like slow convergence rate and low precision solutions foe which an approach was proposed with three modified for attraction model , search model and dynamic updation model .The experimental results has proved that modification of the basic FA will result in a better performance.

The authors of the article [7] have proposed a novel FA to overcome the premature convergence and poor global exploration when used to optimize complex and high dimension engineering problems. The experimental results carried out on real-world complex engineering problems shows acceptable performance in exploration and exploitation, having the optimization capability, robustness, and efficacy.

Ravneil Nand and his team [8] used the idea of Intelligent optimization algorithms based on swarm principles for complex problems. This research has proposed a new modification of FA by introducing a novel and unique stepping ahead parameter hybridising with Covariance Matrix Adaptation Evolution Strategy (CMAES) to improve the exploitation further while maintaining good exploration in the fireflies. The proposed algorithms was found to have outperformed FA algorithm in both benchmark and real-world problems.

The above said research work infers that when basic FA is ensembled with other techniques , an efficient hybrid model can be generated.

Genetic

The researcher [10] has given a comprehensive review of genetic algorithms (GAs), covering their past, present, and future with the pros and cons. This article provides a broad understanding about GAs, including classification, related genetic operators, fitness functions, and hybrid algorithms. The future research directions in the area of genetic operators, fitness function and hybrid algorithms are well defined.

The research study [11] focuses on demonstrating use of genetic algorithm in neural networks. The article presents the research work of modifying the parameters of the genetic algorithm to achieve optimization and acceleration of the neural network learning process. The results which were statistically presented hints at using the genetic algorithm in various domains.

10.48047/jocaaa.2024.33.08.40

Stability of solution

In the article [12] ,details about the need and importance of feature selection algorithm were clearly explained. Since there are many ways of performing feature selection , it arises the need of the interpretability and stability of traditional feature selection algorithms. The high correlation between the features of a data set produces many equally optimal signatures, leading to instability, thereby reducing the confidence of selected features. Stability here it means the sturdiness of the selected feature preferences resulted due to perturbation of training samples. Stability of the feature selection algorithm has to be taken into consideration as the importance is given to high classification accuracy. The authors have given an overview of feature selection techniques and instability of the feature selection algorithm along with the some solutions.

Feng Xiang, Yulong Zhao et.al [13] had proposed an feature selection approach with improved stability. The reason behind focussing on improving the stability is , the uncertainty and complexity of real data increasing the difficulty in data-based knowledge discovering. In this article, authors have proposed an ensemble learning-driven stable feature selection method where the selected features are analysed and filtered ensure its stability.

Thus the above related works guides in proposing a new approach for feature selection using a hybrid model which is clearly explained below.

III. Feature Selection Process

Today, the technology has taken the wheels of the world and is constantly evolving which is becomes a main factor that shapes the upcoming future and helps to be professionally successful. The state-of-the-art technologies sets the emerging trends Every trend possess a lot of potential to allow individuals to stay in the ever-growing technological domain and also help organizations upgrade their business processes in such a way their potential customers can connect with them very easily. With the development that is happening in the technologies in various domains , the chances of generating large scale data becomes more common .

That is data gets generated at a rapid rate from every sphere of human life activities, there by forming a huge chunk of data of different types and categories making the traditional methods fail in processing. Various research findings confirm that these large chunks of data, called as

10.48047/jocaaa.2024.33.08.40

Big data having many features which might be relevant or irrelevant , still will have some important unseen patterns , from which critical information -knowledge can to be grabbed by using appropriate mining and machine learning techniques.To make the model to work efficiently it rises the need for tuning the data set in terms of size and useful attributes. If dataset can be filtered with good set of attributes and records , then it might not hamper the outcome of the machine learning models.There exists many means for filtering out the inappropriate features by using feature selection methods.Feature selection can be performed by deterministic machine learning algorithms and by non- deterministic evolutionary methods also. In the proposed approach both non-deterministic evolutionary approach, that is genetic and swarm based firefly algorithm are used for feature selection.

IV . Evolutionary Algorithm

Non-deterministic approach is often preferred in the scenarios where finding an exactsolution in the solution space for a given problem becomes challenging or impractical ,like in the field of artificial intelligence, machine learning, and optimization problems.The non-deterministic algorithms include the factor of uncertainty along with randomization and parallelism.

The problems can be optimized using heuristic techniques, evolutionary techniques and by using mathematical models.

But for a NP hard problem, in computer science has always been an engaging yet challenging task.Evolutionary methods are more suitable in solving NP-Hard problems.

In the proposed method, the feature selection in larger data set having larger dimensions with respect to number of feature is considered to be a NP-Hard problem and is solved using a hybrid model of both evolutionary and swarm optimization technique.

The term Evolutionary Algorithms (EA) is used to describe systems for solving optimization or search problems based on biological evolution using stochastic search methods.

Evolutionary Algorithms comes under the category of evolutionary computation techniques.

10.48047/jocaaa.2024.33.08.40

It refers to any population-based metaheuristic optimisation algorithm that incorporates the mechanisms similar to biological evolution process namely the reproducing nature , mutation , recombining , natural selection and the persistence of the fittest in the crowd.

Out of many evolutionary algorithms, the most widely accepted Genetic Algorithm (GA) is used in the proposed methodology because of its success witnessed in many researches of different optimization problems.

V. Genetic Algorithm

Most of the problems require optimization, which can be easily solved by using suitable

evolutionary algorithm. Among the different Evolutionary algorithms available , Genetic algorithm (GA) is commonly used because of its solving capability.The working principle of how GA can be used was first introduced by John Holland in 1970s [9].

This GA usually delivers acceptable optimal solutions.It base is mainly on the Darwin's' evolution principles and it is all about chromosome, inheritance, selection, crossover or recombination, mutation and reproduction principles.GA can be Real-coded and binary – coded.

The basic steps involved in GA are ▪ Defining an initial population of chromosomes.

▪ Calculate the fitness of every member inside the population.

▪ Selection of chromosomes based on fitness value.

▪ Formulate the next generation from the current generation by selection, cross-over and mutation

▪ Repeating the steps until suitable solution is found.

VI. Swarm Optimization

Swarm Intelligence is used in many domains for optimization , which refers to collective and collaborative behaviour of distributed and communicating multi-agent systems which can be

10.48047/jocaaa.2024.33.08.40

natural or artificial. SI algorithms are metaheuristic optimizer methods which are inspired by behaviours of swarms of various species[b] which are likely direct or indirectly communicate among each other and collaboratively solve a distributed problem. Natural examples of SI consist of ant colonies, fish schooling, bee colony, fire fly movements etc. In the proposed approach Firefly algorithm is used along with Genetic Algorithm to get combine the benefits of evolution and swarm intelligence in solving the feature selection problem.

a. Fire Fly Algorithm:

Firefly Algorithm (FA) is a meta-heuristic nature inspired swarm intelligence based algorithm developed based on the strategy adopted by the fireflies during their movement . It mainly works on three principles :

• Fireflies irrespective of gender attracts other fireflies based on their attraction power that is the intensity of flashing light.

• The firefly can attract and guide other firefly by the brightness of their flashing light which is inversely proportional to distance. If there exists no flashing light to guide then firefly will move randomly in any direction.

• The fitness function determines the light intensity which in turn gives brightness of a firefly.

In this method, each firefly represent an optimal set of features and will possess a brightness value which is the value obtained by the fitness function which evaluates the firefly content. New solution evolves as the fireflies move towards the brighter firefly.

During the evolution process, better fireflies are retained and carried over to the next successive step. The fireflies are graded based on the objective function output.

The distance between the fireflies are found by using the Euclidean distance formula and the content of firefly can be updated based on the Intensity and light absorption co-efficient.

Suppose that for each solution i , i(Xi) represents the position in solution space at that point of time w.r.t iteration. If the fitness of ith solution is greater than another solution j , the space between them in the solution space can be found by using Euclidean distance formula as mentioned in Equation (1).

rji= SquareRoot (for each position $(Xi - Xj)^2$) ……...(1)

10.48047/jocaaa.2024.33.08.40

The value of Equation (1) will be used in Equation (2) to calculate the new attractiveness value.

$\beta = \beta_0 e^{-\gamma r_{ij}^2}$ ……...(2)

Where β_0 represent the attractiveness when fireflies are at same position and normally set to 1. γ is the light absorption coefficient and normally set to 1 and rand represent the random number.

The new location of i^{th} firefly by using the below equations(3) and (4)

$New X_{ij} = Old\ X_i + \beta.rand.\Delta X_{ij} + \alpha\ (rand - 0.5)$……(3)

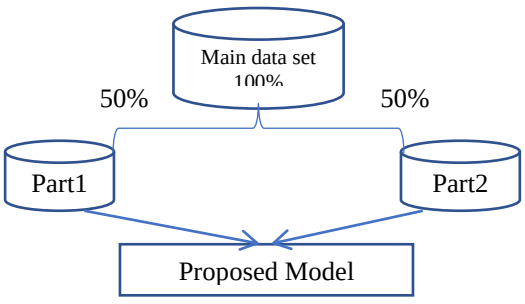
$\Delta X_{ij} = (X_i - X_j)$ ……...(4)

Thus in FA the evolution of solutions in terms of fireflies will repeat until the specified the termination criteria.

V. Proposed Methodology

In this section, the methodology used in the proposed work is discussed. Both swarm optimization and evolutionary methods are being used to retrieve the optimal subset of features from the input dataset by overcoming the main drawback of evolutionary techniques, getting into local optima. The Exploration of solution space is achieved by the partitioning technique. The input database is split into two equal parts leading to two phases ,wherein the first phase will be using the first part of the data set and the second phase using the other part . The reason for splitting data set is to get a stable output from the model by checking the stability of solution under varied training data.

Each part is subjected to the proposed hybrid model scenario Fig.1.1.



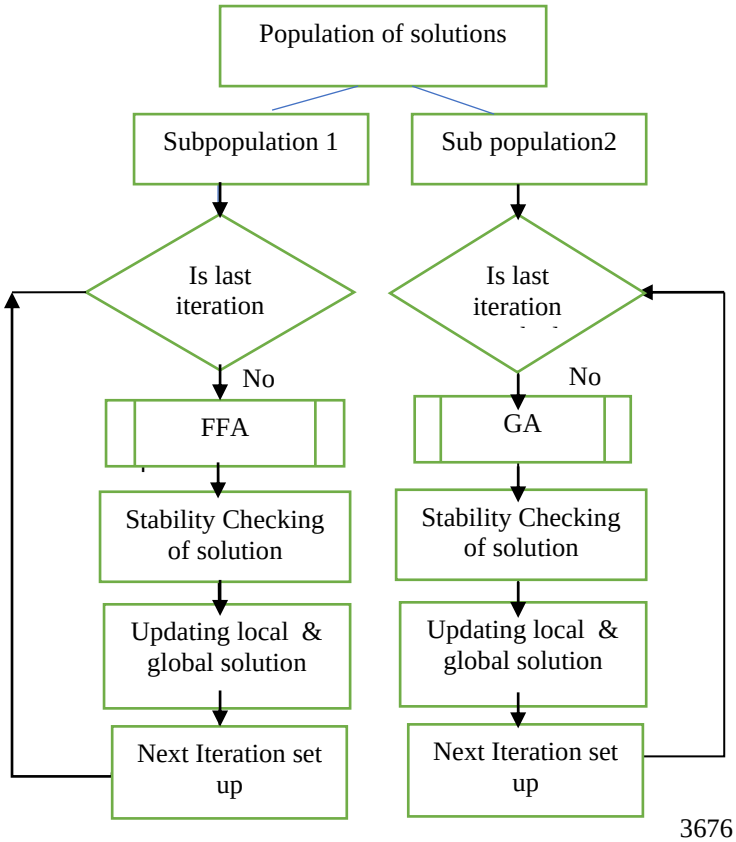
10.48047/jocaaa.2024.33.08.40

Fig.1.1. Partitioning of dataset

In each phase ,the proposed hybrid model is used to find the optimal solution and later integrated .

In the hybrid model , the initial population of solution is divided into two sub population. Each subpopulation is processed twice with 50% of the dataset, first time with first half and next time with the remaining. Each of the subpopulation will be using Firefly algorithm to get optimal solution and also will be processed with the Genetic Algorithm, and thus exploitation of solution space is achieved by both the type of algorithms in the proposal.

The solution obtained from each algorithm is checked for the global best solution criteria and the solution obtained from either part of dataset is checked for similarity with the existing global solution and considered to be part of solution space only if it is not repeated or less similar, thus ensuring the method doesn't get trapped into one particular solution/(s) during iterations. The proposed hybrid model for feature selection is pictorially depicted below in Fig.1.2



1	0	1	1	.....	1	1

10.48047/jocaaa.2024.33.08.40

Fig.1.2. Proposed Hybrid Model Fig.1.3. Firefly / Chromosome

Methodology

In this proposed approach , the data set is partitioned into two sub sets with a scope for parallel processing and subjecting each data sub set to the hybrid model , where the drawbacks for basic version of FFA [13] like less accurate solutions and long duration to converge are overcome by using the population partition strategy and by using the selection and mutation operations of GA, a evolutionary algorithm.

Each sub- population is processed independently and the results are integrated at the end. This proposed approach involves two phases as shown in Fig.1.2 .The population has collection of binary strings , which represents the solution in binary encoded form where “1” indicates the selection of attribute and “0” the non-selection of attribute which is shown in the following Fig.1.3 .This binary encoded string is trateated as firefly in FFA and treated as chromosome in GA.

The size of the binary string relies on the count of attributes of the given input dataset. During the processing of fireflies , a transfer function and a discretization/binarization method are applied to get the result in terms of 0s/1s . First the transfer function is applied to the outcome of the Equation.3 which in turn will result in a real number , then on this real number binarization method is applied so that it results in 0 or 1.

The pool of binary encoded solutions is partitioned into two equal parts resulting into a small sized sub-population of fireflies and chromosomes which are processed separately.

Phase-1

In this phase, the first half of population is used in FFA algorithm for the specified number iterations. In each iteration , the firefly moves towards the brighter firefly and gets modified depending on the light intensity. The light intensity is modelled as the objective function.

The selected attributes, number and position of 1’s of the firefly is used for the dimensionality reduction and tested using Decision Tre classifier model. The accuracy value is considered as the desired fitness. The process is repeated for all the fireflies and sorted based on the light

10.48047/jocaaa.2024.33.08.40

intensity values. Only if the light intensity is greater than the neighbour fireflies' there will be a movement of which is done using the transfer function.

In order to get a stable output ,the stability is checked where the method should produce more similar output under the training data variation since the instability of solution with respect to the input data produces widely different output and will makes the solution untrustworthy maked the model unreliable one.

At the end of each iteration, based on the stability of solution , it gets appended to the local best solution pool.

Fitness of Firefly

The fitness of each firefly is found by considering the number of selected attributes along with accuracy score of the classifier by considering the reduced dataset with selected attributes .

Phase -2

In this phase the second sub-population is treated as the solution pool of chromosomes and is subjected to GA.

Each chromosome is evaluated by the objective function .After mutation followed by cross-over operation the fitness of chromosome is recalculated for all by the objective function as used in FFA and sorted based on the fitness value preparing for the next iteration.

At the end of each iteration , the stabiltity of solution is tested and added accordingly to the solution pool maintain the local best solution.

At the end of both FFA and GA in each iteration , the pool that contains the local best solution is inspected and then the best solution gets added to the global best solution pool.

The outcome of the proposed evolutionary hybrid model is evaluated against other feature selection methods like Univariate selection method (SelectKBest) and Principal Component Analysis by taking into account the count of selected features. The outcomes obtained after experimenting with different datasets show that modified evolutionary firefly algorithm selects better set of features compared to other methods.

A. Transfer Function

10.48047/jocaaa.2024.33.08.40

The binary coded firefly requires a function that transforms continuous search space into the discrete binary search space called transfer functions , which plays vital role in all binary swarm processing algorithms [14].This transfer function that converts the elements of position vector into the interval from 0 to 1 can make the fireflies to wander in the binary space. For a binary space which deals with only two numbers (“0” and “1”), the position updating process, toggling between “0” and “1” values cannot be done by using Equation. (3). This kind of toggling in a firefly algorithm should be done depending on distance and Intensity measure between the fireflies. There are V-Shaped and S-Shaped transfer functions. In this proposed work a V-Shaped transfer function is used for updating the positions of the fireflies as it is proved that the v-shaped family of transfer functions are capable of finding the best solution with good convergence rate for multimodal benchmark functions using swarm optimization algorithms namely BPSO[17] . The V shaped transfer function used is described below.

$T=X / \sqrt{1+X^2}$ ……………………………………………………………………………(5)

Where X represents a position in the firefly and gets updated by the following rule

Xnew = 1 if T > rand (0,1)

0 otherwise ……………………………….…………………………..... (6)

By using the equations 3,5 and 6 every bit in the firefly will be updated for better exploration and exploitation.

B. .Stability of Solution

Basically the stability of feature selection measures are categorized into three groups namely are stability by rank, stability by weight and stability by similarity [12,13].

In the stability by rank approach, stability of feature selection algorithm is evaluated by the correlation between two ranked feature lists. Weight based stability [10] use the weight of features in the subset for measuring the stability.

The stability by similarity handles with the feature subsets containing different number of features, respectively. In this work the stability by similarity only is used.

The stability of solution in an evolutionary approach because vital as the major drawback of evolutionary or Swarm intelligence is getting into local optima .Hence in the proposed

10.48047/jocaaa.2024.33.08.40

approach in every iteration the stability of local best solution of both the Genetic and firefly algorithm is checked using lusgarten's measure and only if the stability index is found to be positive ,the local best solution gets appended to the global solution pool.

3.2.1. Lustgarten's Measure

Lustgarten measure is the modified Kuncheva index. It's measure satisfies the property of correction by chance and is applicable to different cardinality of selected feature subsets [28].

Lustgarten's measure SI_L is defined as:

$$SI_L(Si,Sj) = (r - E[r]) / (max(r-E[r]) - min(r-E[r])) \quad \ldots\ldots(7)$$

If two selected feature subsets S_i , and S_j are of cardinalities k_i and k_j , respectively, $r=|Si \cap Sj|$ is the cardinality of intersection of two selected subsets of features , $E[r]=k_i k_j /n$, $max(r)=min(ki,kj)$ and $min(r) =max(0,ki+kj-n)$, the above equation reduces to:

$$SI_L(S_i,S_j)=r- (k_i k_j /n) / [min(k_i,k_j)-max(0,k_i+k_j-n)] \quad \ldots\ldots(8)$$

This measure has a value in the interval $(-1,1)$. For random feature subset selection, Lustgarten's measure provides a value of 0. It produces a positive value when feature selection method is more stable than random feature selection and produces a negative value when feature selection method is less stable than random feature selection. If Si or Sj or both have no features or Si or Sj or both contain all the feature in the domain, then SI_L is undefined.

The drawbacks of the measure is that it cannot reach maximum $+1$, when the feature subsets are identical and cannot reach its minimum -1 when the cardinality of intersection between feature subsets is zero , is overcome by adding a correction measure as mentioned below.

If $r=ki=kj$ the maximum value for the stability measure occurs i.e 1 , when r is defined within the range $0<r<n$.

When $0<r<n/2$

10.48047/jocaaa.2024.33.08.40

$SI_L(S_i,S_j)=r-(k_ik_j/n)\ /\ [\ min(k_i,k_j)-max(0,k_i+k_j-n)]+$ r/n……..(9)

When $r=n/2$

$SI_L(S_i,S_j)=r-(k_ik_j/n)\ /\ [\ min(k_i,k_j)-max(0,k_i+k_j-n)]+$ ½……..(10)

When $n/2<r<n$

$SI_L(S_i,S_j)=r-(k_ik_j/n)\ /\ [\ min(k_i,k_j)-max(0,k_i+k_j-n)]+$(n-r)/n ……..(11)

When r=0

$SI_L(S_i,S_j)=r-(k_ik_j/n)\ /\ [\ min(k_i,k_j)-max(0,k_i+k_j-n)]+$ (max((k_i,k_j)/n)- 1 ……..(12)

When r=n

$SI_L(S_i,S_j)=0$……..(13)

B. Algorithm

The proposed methodology involves the following algorithm which is made up of portioning and two phases of processing.

Pseudo code:

1. Horizontal partitioning of database into two equal halves.
2. Initial Parameter setting for both Firefly algorithm and Genetic algorithm
3. Creating Initial Population of solution
4. Partitioning the population
5. Applying Firefly algorithm to the first subpopulation and also to the second sub population –Phase1
6. Applying Genetic algorithm to the first subpopulation and also to the second sub population – Phase 2
7. Combining the Population and finding the local best solution of both the methods
8. Integrating the results of both the phases and determining the global best solution based on the stability of solutions.

Phase-1 Firefly Algorithm for Feature Selection

Input: Dataset Output: Global best firefly

Initialization: Attractiveness at zero distance β_0=1; Light absorption coefficient γ=1

10.48047/jocaaa.2024.33.08.40

Objective Function f(x): x=<x1,x2,..xd> [d denotes no. of attributes of dataset]

f(x) is represented in Fig. 3

Max_Iterations=<Integer value> ; Number of fireflies: n=<Integer value>

Step: 1 Consider the pool of " n/2" fireflies[First half of fireflies with first 50% dataset and second half of pool of fireflies with remaining dataset]

Step: 2 Evaluate the light Intensity of fireflies using f(x)

Step: 3 Loop till Max_Iterations

Step:3.a.Loop for i=1 till n

Loop for j=1 till n

If(Intensity(j) > Intensity(i))

Step :3b. Move ith firefly towards jth using Eq.(3) using the transfer function

Step :3c Vary the attractiveness using Eq.(2)

Step :3d Evaluate the fireflies and update light intensity

Step :3e Select the local best from both sub populations and compare with global best for the stability using Lusgarten's measure using the Eq.13 thru Eq.19 and update the global best solution

Step :3f Record the progress of the best firefly

Step:4 Output the global best firefly

Phase- 2 Genetic Algorithm for Feature Selection

Input: Dataset Output: Global best chromosome

Initialization: Mutation Probability pm= Cross Over Probability pc=

Objective Function f(x): x=<x1,x2,..xd> [d denotes no. of attributes of dataset]

f(x) is represented in Fig. 3

Max_Iterations=<Integer value> ; Number of chromosomes: n=<Integer value>

Step: 1 Consider first the pool of " n/2" chromosomes with first 50% dataset and second half with remaining dataset

Step: 2 Evaluate fitness using the f(x)

Step: 3 Loop till Max_Iterations

Step:3.a.Loop for i=1 till n

10.48047/jocaaa.2024.33.08.40

Loop for j=1 till n

If(Intensity(j) > Intensity(i))

Step :3b. Move ith firefly towards jth using Eq.(3) using the transfer function

Step :3c Vary the attractiveness using Eq.(2)

Step :3d Evaluate the fireflies and update light intensity

Step :3e Select the local best and compare with global best for the stability using Lusgarten's Measure using the Eq.13 thru Eq.19 and update the global best solution

Step :3f Record the progress of the best firefly

Step:4 Output the global best firefly

Termination Condition : The proposed algorithm terminates after the specified number of iterations is completed.

Objective function: f(x)

Step1: Scanning the firefly to select those attributes which are set by the corresponding position in the firefly

Step2: Based on selected attributes the dataset dimension reduced by omitting those Unselected attribute values

Step3: Generating a Classifier model using reduced dataset

Step4: Calculating the accuracy score of the classifier and the correlation among the predictor variables and target .

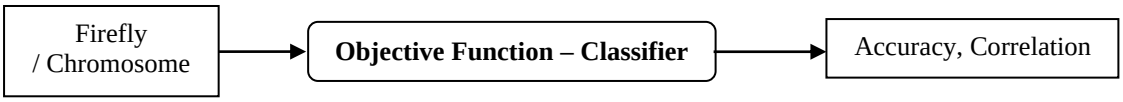


Fig.3. Objective Function outline

The sorting of fireflies is done based on the objective function output and count of selected features. During every iteration locally best firefly is identified and checked with the global best firefly for similarity. If the locally best is found to improved solution than global then global best firefly is updated to local.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

10.48047/jocaaa.2024.33.08.40

The proposed methodology is implemented using python language in Anaconda3 Jupyter Notebook. In the research work several packages are used for randomization, selection and correlation calculation.

Using two dataset as described in Table I, the performance of the algorithm is analyzed. The datasets used were assumed to be fit without noise and missing values. In the experiments conducted the execution time was not making a great difference but showed better results in terms of accuracy and correlation.

The Decision tree classifier model is used in proposed algorithm with the train-test split ratio 70:30 and random State = 10.

TABLE I. DATASET DESCRIPTION

S.No	**Dataset Information**		
	Name	***Instances***	***Attributes***
1	winequality-white.csv Source : https://www.kaggle.com	4898	12
2	WDBC.csv; Source : https://www.kaggle.com/uciml/breast-cancer-wisconsin-data	569	32
3.			
4	Spambase.csv ; Source : https://archive.ics.uci.edu/ml/datasets/spambase	4601	58

In this article SelectKBest available in scikit-learn library is used along with PCA. Principal Component Analysis (PCA) is a another dimensionality-reduction method is used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information found in the large set where data is projected into a subspace (principal components).. The count of selected features from proposed algorithm will be taken as input to PCA which indicates the number of principal components to be generated from original feature space. Then based on the principal components the dataset would be transformed and fit into a Decision Tree Classifier model to check its performance in terms of accuracy.Thus the accuracy score of SelectKBest, Principal Component Analysis and the proposed algorithm is compared .The proposed algorithm has some parameters to be set which controls the movement of fireflies in exploring and exploiting the search space as specified in

10.48047/jocaaa.2024.33.08.40

Table II. PARAMETERS

S.No	Parameter	INITIAL VALUE
1	Number of Fireflies	20
2	MAX_ITERATION	50
3	Betamin	0.2
4	Gamma	1
5	Alpha	0.5
6	Mutation probability	0.15
7	Cross-over probability	0.8

The following tables and graphs depicts the evaluation of the proposed methodology with other methods with respect to each dataset depicting the progress of firefly algorithm along the iterations in terms of accuracy score , accuracy obtained using SelectKBest and the selected attributes, accuracy obtained using Principal Component Analysis – (PCA) method and performance of of Decision Tree classifier considering all the features (i.e., without feature selection)

PERFORMANCE COMPARISON :

Table III : Dataset : **winequality-white.csv** Shape of Dataset : (4898, 12)

<table>
<tr><td rowspan="2">Proposed Approach</td><td colspan="4">Selected attributes : 7+1(target)
: ['volatile acidity', 'citric acid', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'quality']</td></tr>
<tr><td rowspan="2">Other Methods</td><td colspan="2">Accuracy</td><td colspan="2">Precision</td></tr>
<tr><td>All (12)</td><td>Selected (8)</td><td>All(12)</td><td>Selected(8)</td></tr>
<tr><td>Decision Tree Classifier</td><td rowspan="3">59.25</td><td>66.19</td><td rowspan="3">60.47</td><td>66.95</td></tr>
<tr><td>Select K Best</td><td>59.79</td><td>60.30</td></tr>
<tr><td>PCA</td><td>58.29</td><td>59.33</td></tr>
</table>

10.48047/jocaaa.2024.33.08.40

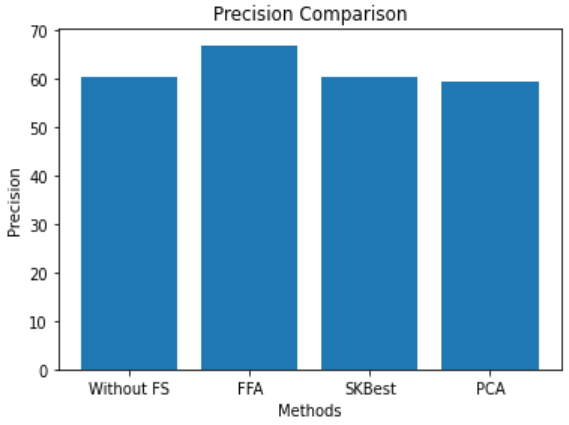


Fig.5. Precision Comparison

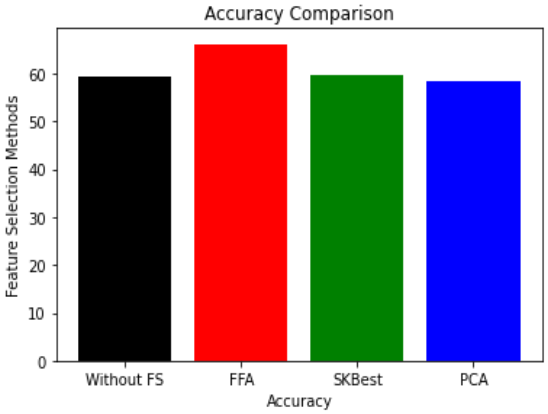


Fig.6. Accuracy Comparison

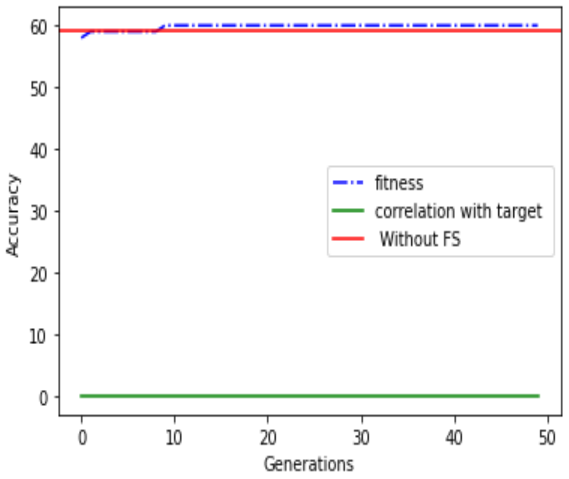


Fig.7. Progress of model

The above e Table III and the figures Fig. 5, Fig.6 and Fig.7 shows that the proposed model has a improved accuracy and precision in the model considering the selected attributes when experimented with cancer data set.The table has the numerical data of accuracy and precision obtained during implementation , where the number of selected attributes of the proposed model is taken as the input for the SelectKbest and PCA models. The figures show how the accuracy has progressed during the iterations.

Dataset : WDBC.csv Shape of Dataset : : (569, 32)

Proposed Approach	**Selected attributes** : 10+1 ['radius_mean', 'concave points_mean', 'radius_se', 'texture_se', 'perimeter_se', 'compactness_se', 'concavity_se', 'concave points_se', 'fractal_dimension_se', 'fractal_dimension_worst', 'diagnosis']			
Other Methods	Accuracy		Precision	
	All(32)	Selected	All(32)	Selected
Decision Tree Classifier	91.22	95.3	91.75	95.3
Select K Best		91.8		91.9
PCA		95.2		95.32

10.48047/jocaaa.2024.33.08.40

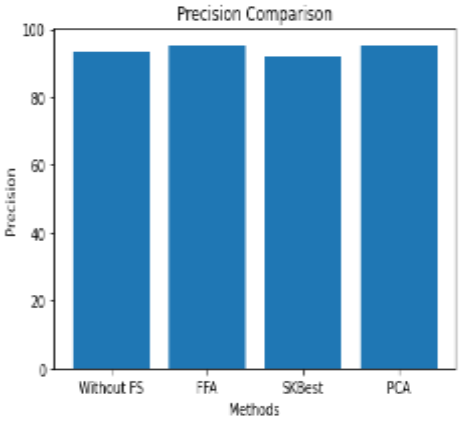


Fig.8. Precision Comparison

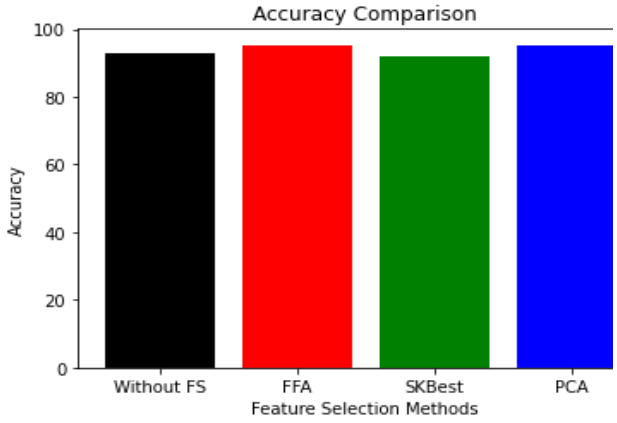


Fig.9. Accuracy Comparison

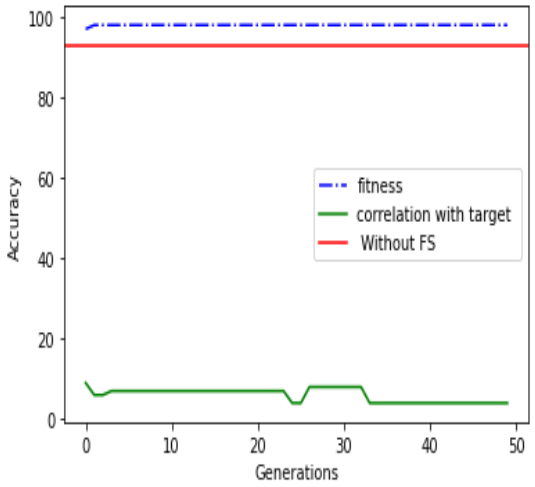


Fig.10. Progress of model

The above e Table IV and the figures Fig. 8, Fig.9 and Fig.10 shows that the proposed model has a improved accuracy and precision in the model considering the selected attributes when experimented with cancer data set.The table has the numerical data of accuracy and precision obtained during implementation , where the number of selected attributes of the proposed model is taken as the input for the SelectKbest and PCA models. The figures show how the accuracy has progressed during the iterations.

Table VI : Dataset Name : spambase.csv Shape of Dataset : (4601, 58)

Proposed Approach	**Selected attributes** : 24+1 ['a1', 'a2', 'a3', 'a5', 'a6', 'a7', 'a8', 'a9', 'a10', 'a11', 'a12', 'a13', 'a17', 'a22', 'a24', 'a26', 'a32', 'a35', 'a37', 'a43', 'a44', 'a51', 'a54', 'a56', 'a58']			
Other Methods	Accuracy		Precision	
	All(58)	Selected	All(58)	Selected
Decision Tree Classifier	90.51	93.7	90.49	93.69
Select K Best		90.87	90.87	
PCA		87.18	87.24	

10.48047/jocaaa.2024.33.08.40

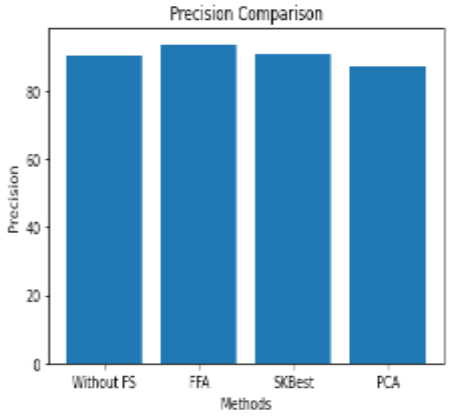


Fig.11. Precision Comparison

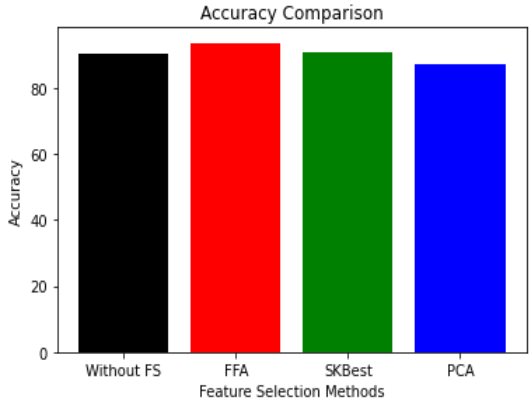


Fig.12. Accuracy Comparison

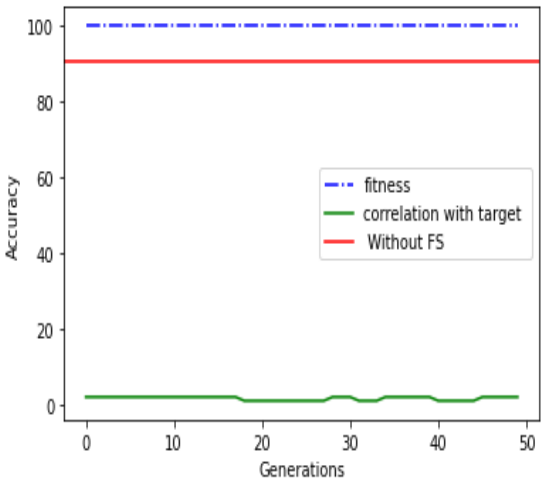


Fig.13. Progress of model

The above e Table V and the figures Fig. 11, Fig.12 and Fig.13 shows that the proposed model has a improved accuracy and precision in the model considering the selected attributes when experimented with cancer data set.The table has the numerical data of accuracy and precision obtained during implementation , where the number of selected attributes of the proposed model is taken as the input for the SelectKbest and PCA models. The figures show how the accuracy has progressed during the iterations.

Conclusion

In this paper , a Metaheuristics swarm intelligent algorithm – modified firefly algorithm is coupled with genetic algorithm to determine a set of significant features for the given datasets.The proposed methodology exploits the benefits of swarm intelligence along with the evolutionary operations like flip mutation and simple cross over on the partitioned population and data set. To demonstrate the success of the algorithm's performance, three different datasets were used for analyzing and found to outperform in selecting features leading to the near to same or better accuracy when considered with all the features. Therefore, it can be

10.48047/jocaaa.2024.33.08.40

employed as a better tool for attribute reduction . It was noticed that during experimentation the results found to vary given the stochastic nature of the algorithm. Thus this study has given a sign that swarm intelligence when combined with evolutionary operations will help in developing a better model. To get a consistent result the solution space has to be exploited , hence in the proposed method has adopted checking the stability of solution at every iteration . The approach has to incorporate larger number of iteration and fireflies for overcoming the time complexity issue, providing a scope for further research.

References

[1] Chen, RC., Dewi, C., Huang, SW. et al. Selecting critical features for data classification based on machine learning methods. J Big Data 7, 52 (2020). https://doi.org/10.1186/s40537-020-00327-4.

[2] Noroozi, Z., Orooji, A. & Erfannia, L. Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. Sci Rep 13, 22588 (2023). https://doi.org/10.1038/s41598-023-49962-w

[3] Slowik, A., Kwasnicka, H. Evolutionary algorithms and their applications to engineering problems. Neural Comput & Applic 32, 12363–12379 (2020). https://doi.org/10.1007/s00521-020-04832-8

[4] Bartz-Beielstein, Thomas & Branke, Juergen & Mehnen, Jorn & Mersmann, Olaf. (2014). Evolutionary Algorithms. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 4. 10.1002/widm.1124.

[5] Mohd Nadhir Ab Wahab ,Samia Nefti-Meziani ,Adham Atyabi ,A Comprehensive Review of Swarm Optimization Algorithms , May 18, 2015.https://doi.org/10.1371/journal.pone.0122827

10.48047/jocaaa.2024.33.08.40

[6] Pan, X., Xue, L. & Li, R. RETRACTED ARTICLE: A new and efficient firefly algorithm for numerical optimization problems. Neural Comput & Applic 31, 1445–1453 (2019). https://doi.org/10.1007/s00521-018-3449-6

[7] Mojtaba Ghasemi, Soleiman kadkhoda Mohammadi, Mohsen Zare, Seyedali Mirjalili, Milad Gil, Rasul Hemmati, A new firefly algorithm with improved global exploration and convergence with application to engineering optimization, Decision Analytics Journal, Volume 5, 2022,100125,ISSN 2772-6622,https://doi.org/10.1016/j.dajour.2022.100125. (https://www.sciencedirect.com/science/article/pii/S277266222200056X)

[8] Ravneil Nand, Bibhya Nand Sharma, Kaylash Chaudhary,Stepping ahead Firefly Algorithm and hybridization with evolution strategy for global optimization problems, Applied Soft Computing,Volume 109,2021,107517,ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2021.107517.(https://www.sciencedirect.com/science/article/pii/S1568494621004403)

[9] J. H. Holland, "Genetic algorithms and the optimal allocation of trials ", SIAM Journal on Computing, vol. 2, no. 2, pp. 88–105, 1973

[10] Katoch, S., Chauhan, S.S. & Kumar, V. A review on genetic algorithm: past, present, and future. Multimed Tools Appl 80, 8091–8126 (2021). https://doi.org/10.1007/s11042-020-10139-6

[11] Kotyrba, M.; Volna, E.; Habiballa, H.; Czyz, J. The Influence of Genetic Algorithms on Learning Possibilities of Artificial Neural Networks. Computers 2022, 11, 70. https://doi.org/10.3390/computers11050070

[12] Utkarsh Mahadeo Khaire, R. Dhanalakshmi, Stability of feature selection algorithm: A review,Journal of King Saud University - Computer and Information Sciences,Volume 34, Issue 4, 2022, Pages 1060-1073, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2019.06.012.

[13] Feng Xiang, Yulong Zhao, Meng Zhang, Ying Zuo, Xiaofu Zou, Fei Tao, Ensemble learning-based stability improvement method for feature selection towards performance

10.48047/jocaaa.2024.33.08.40

prediction, Journal of Manufacturing Systems, Volume 74, 2024, Pages 55-67, ISSN 0278-6125., https://doi.org/10.1016/j.jmsy.2024.03.001.