A Novel Hybrid PSBCO Algorithm for Feature Selection

S.Sandhiya[*1], Dr. U. Palani[*2]

[1]*Teaching Fellow, Department of Information Technology, University College of Engineering Villupuram, Tamilnadu, India..*

[2]*Professor, Department of Electronics and Communication Engineering, IFET College of Engineering, villupuram, Tamilnadu, India.*

Abstract

Feature selection is an important process in data mining which enriches the classification efficiency by eliminating irrelevant and redundant features from the original feature set. Feature selection plays a significant role in artificial intelligence and machine learning. In this paper, we propose a hybrid algorithm which combines the Binary cuckoo search algorithm (BCS) and Particle Swarm optimization (PSO) algorithm called Particle Swarm Binary Cuckoo optimization Algorithm (PSBCO). Proposed algorithm comprises of two steps, in first step subset generation is performed by using BCS optimization algorithm it will generate n number of subsets from the original large dataset. In second step subset selection is performed by PSO Algorithm it is used to evaluate all the selected subsets and selects the best subset from the generated n number of subset. PSBCO algorithm efficiency was tested with heart disease data using MOA tool. Through the experimental outcomes, PSBCO Algorithm has greater prediction accuracy with optimal number of selected features. The experimental results proves that the classification accuracy of proposed algorithm is significantly higher than the other classical cuckoo search and PSO algorithms.

Keywords — *Feature selection, PSBCO, Binary Cuckoo Search, Particle Swarm.*

I. INTRODUCTION

Feature selection plays a vital role in many practical and real time application domains such as machine learning, data mining, knowledge discovery and deep learning. Feature selection goal is to select the best discriminating feature from dataset and it simultaneously eliminate the irrelevant, redundant and extraneous features. Feature selection is also necessary in the problem domain of high dimensional data, were the objective of feature selection is to lessen the dimensions of the original data, so that the processing time and computational cost is reduced and the classification precision is increased.

Researchers have proposed large number of feature selection algorithms which are commonly categorized into Wrapper, Filter and Embedded approaches. The Wrapper method generates feature subset by finding the search space between the features using searching algorithms and selects best subset by evaluating all the generated subsets. Filter approach is similar to the wrapper approach but it uses simple filter model for evaluation instead of running all the models. First approach produces accurate result but takes more time for execution second approach is reduce the execution time but the accuracy is lesser than the wrapper approach. The Embedded approach take advantages of both the above approaches, it improves accuracy and reduces execution time. It takes intrinsic characteristics of data set and use predefined mining algorithms for subset generation and evaluation process.

Proposed work uses embedded approach in which two optimal search algorithms are used such as Binary cuckoo search (BCS) algorithm and particle swarm optimization (PSO) algorithm. A novelty of proposed work is to combine these two algorithms and introduce new framework called PSBCO for feature selection process. Which select the best feature subset from the given original large dataset.

The basic concept of these algorithm are explained as follows:

A. Binary Cuckoo Search

The organism behaviour of Cuckoo bird is enormously interesting. These birds are lay down their eggs in a host nests, and parodist external appearances (egg colour and spots) of host eggs. The host bird throws the cuckoos egg if it finds difference among the eggs, or simply abandon its nest and starts make new nest in some other place. Based on this context D. Rodrigues, L. A. M. Pererira and Xin-she Yang [15] have developed a novel Binary Cuckoo Search (BCS) algorithm based on cuckoo search behaviour. They have summarized three rules for BCS as follows:

i. Each cuckoo randomly choose a nest to lay their eggs.
ii. Available number of host nest is fixed, and only the high quality nest will be passed to next generation.
iii. If the host bird finds the cuckoo egg it will throw the egg and builds the new nest.

B. Particle Swarm Optimization

Initially PSO consists of group of random particles and searches for best particle by passing over to next generations [4]. For every generations it performs number of iterations. Each particles are updated at each iteration process using two parameter values such as (pbest) and (gbest). These parameter values are calculated using fitness function. The first parameter is the best solution or fitness value obtained for every iteration and second parameter is global best value obtained throughout the process.

The general pseudocode for PSO is as follows.

```
For each particle
Initialize the particle
Do
Calculate fitness value for each particle
Compare current fitness value with previous best value
(pbest) if it is best
Then set current value as (pbest)
End
Select best particle with optimal fitness value of all the
particles as (gbest)
End
```

The Proposed Feature Selection framework is consists of two step process such as 1.Subset generation and 2.Subset evaluation. In subset generation process n number of subsets are generated from the original large dataset. The subset generation is a searching procedure it uses searching algorithm to produce number of candidate feature subsets by using binary cuckoo search algorithm. In subset evaluation process the each subset is evaluated by the optimization algorithm. We use particle swarm optimization algorithm for evaluating the each candidate subsets. In general the evaluation is performed based on number of search strategies such as complete, random, sequential and heuristic searching strategies. Our proposed algorithm uses Heuristic searching strategy for evaluation process. In Subset evaluation the redundant and irrelevant features are removed. And the best subset is selected among the n number of generated subsets based on the fitness function.

To analyse the efficiency of the proposed framework we used MOA tool for experiments which is most popular free open source framework for data mining, it consists of collection of machine learning algorithms. In experiments we used Heart disease data set as input and we obtained reduced optimal number of feature subset. Also the classification accuracy of the algorithm is tested with three different classifiers such as K-Nearest Neighbors (KNN), Decision tree, and Naive bayes algorithms. To validate the performance of PSBCO Algorithm the output value is compared with the other existing feature selection algorithms.

II. RELATED WORK

In data mining and machine learning feature selection is an important pre-processing step which obtains less correlated and different feature subset from original feature set. The objective of feature selection is to obtain optimal feature set by removing irrelevant and redundant features with improved prediction accuracy using learning algorithms. The main advantages of feature selection is to save the storage space, computational time and cost. Research on feature selection began around the year 1970s. In feature selection step the optimal feature set is obtained from the original feature set by generating and evaluating all possible subsets. The number of subset generated from n number of original feature is 2^n, which requires more computation time and cost even for normal sized datasets. To overcome this problem the meta-heuristic algorithms are required with random searching strategies. Many researchers have introduced meta-heuristic algorithms for feature selection this section lists some of the work related to feature selection algorithms.

In [2] novel feature selection algorithm introduced with learning memory and PSO, the learning memory strategy was designed to process more knowledge from the individuals with higher fitness. Also genetic operator is used to balance the local and global exploration of the algorithm. In this study the classification accuracy is performed using k-nearest neighbors algorithm. Fatima et. al introduced hybrid filter and wrapper approach [3], in this study five filtration methods are used with different weights with PSO in order to produce new hybrid algorithm BPSO. In [4] feature selection is performed with high dimensional data using BPSO algorithm which is combined with C4.5 classifiers called BPSO+C4.5. [5] Proposed feature selection methodology for intrusion detection system using PSO. In [6] hybrid heuristic algorithm proposed which combines genetic algorithm and competitive swarm optimizer, which improved the generation speed and prevents premature population. [7] Proposed modified PSO algorithm in this study the SVM is used for fitness function generation.

Sandhiya S and Palani U [8] proposed a hybrid feature subset selection algorithm called Genetic Binary Cuckoo Optimization Algorithm (GBCOA) which comprises of two steps, in first step subset generation is performed by using Binary Cuckoo search (BCS) optimization algorithm it will generate n number of subset from the original large dataset. In second step subset selection is performed by using Genetic Algorithm it will select the best subset from the generated n number of subset. In this paper the feature selection process is improved by combining the two different algorithms. In [9] a novel semi-supervised embedded feature selection method is proposed which extends the least square regression model with set of scale factors.

The comparative study of BCSA and GA for feature selection performance is analysed using sonar dataset from UCI machine [10]. A novel hybrid algorithm for feature selection combines GWO and PSO called PSOGWO [11], in this study the binary version of the PSO is used for better accuracy this work also used kNN and Euclidean separation matric. The cuckoo bird algorithm is combined with naïve bayes classifier for feature selection [12], in this study the cuckoo search is used for subset generation and naïve bayes classification accuracy is used for evaluation. [13] Proposed hybrid improved quantum-behaviour particle swarm optimization algorithm for feature selection called HI-BQPSO. It combines filtering method with improved PSO which reduces the dimensionality of data with better accuracy. The study of modified cuckoo search algorithm [1] improves the prediction accuracy by building fitness function using rough set theory instead of levy flight. In this work the prediction accuracy is analysed using the KNN and SVM with different benchmark datasets. Umair, sadiq and okyak [14] proposed the engineering of genetic algorithm, in this work the fitness function generated using two terms first is feature selection metric MI, JMI and mRMR and second is overlapping- coefficient.

III. PROPOSED WORK

Our proposed work consists of two parts: Subset generation and subset evaluation. The Subset generation process uses binary cuckoo search algorithm and subset evaluation process uses Particle Swarm optimization algorithm. The objective is to find the best combinations of features i.e best subset of features by evaluating fitness function. In our work binary cuckoo search algorithms used to generate n number of subsets by finding distance between each pair of features. The Root Mean Square Deviation (RMSD) formula is used to find distance between each pair of features. It is a regression function which used to find the difference between predicted values and observed values. BCS finds the distance between the features using (1). The Particle Swarm optimization algorithm is used to select best subset from the n number of generated subsets by evaluating all the subsets. PSO uses the fitness function for evaluation process.

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{n}\left((\text{fi}-\text{fj})^2\right)}{n}} \qquad (1)$$

A. *Proposed Architecture*

The architecture of the proposed framework is shown in fig. 1. which is divided into following two step process.

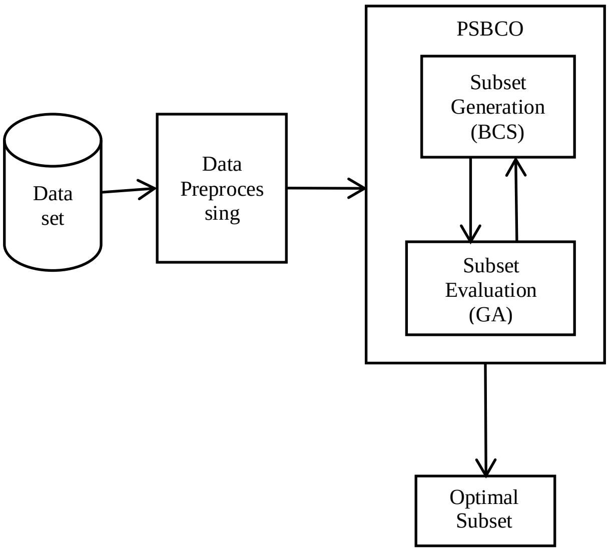

Fig. 1. Proposed Architecture

1) *Subset Generation*

In this step the Binary Cuckoo Search algorithm is used to generate the k number of subsets from the given original large dataset. Subset generation process consists of following steps:

1. K number of clusters (nest) initialized with host egg f_j (features)
2. Cuckoo bird randomly selects nest and lays their egg f_i.
2. The host egg f_j is compared with cuckoo egg f_i using distance formula RMS(f_j,f_i).
3. If obtained distance value is higher than the threshold value then host bird will abandon the egg (feature) otherwise it will be passed to the next generation.
4. The selected eggs are assigned binary value 1 and abandoned egg assigned value 0.
3. For every iteration the features are compared and relevant features are grouped as, k number of clusters.
4. Clusters are considered as subsets S_k[].

2) *Subset Evaluation*

In this step the PSO algorithm used to find the optimal feature subset from the n number of generated subsets. PSO is a population based stochastic optimization technique inspired by that procedures of natural evolution. It functions on a population of individuals to yield better approximations. At each iteration, a new population is generated, the best functions is selected for next generation. Once subsets are generated each subset is evaluated using fitness function. Fitness function is calculated using (2). Fitness function depends on the following three criteria's

1. Classification Accuracy
2. The number of selected features
3. Total feature cost

$$\text{Fitness} = Sk[m] * \text{SVM classification Accuracy} + \sum_{i=1}^{m}(ci*fi)^{-1} \qquad (2)$$

For fitness function generation the classification accuracy (CA) is calculated for every subset using Support Vector Machine (SVM) and the feature cost (FC) is calculated based on computation and execution time of the subsets. The number of feature value 'm' is obtained from every subset. Finally fitness value of all the subsets are compared with each other subset and the subset with best fitness value is selected as optimal subset with reduced number of features.

B. Algorithm Implementation

Our proposed PSBCO Algorithm is shown in fig. 2. in which the parameters are initialized as follows,

Total number of features are considered as eggs (f_1,f_2,…..f_n), number of clusters are initialized as nests (s_1, s_2, …..s_k), each Subset consists of m number of selected features S_k[ ] = { f_1, f_2, …. f_m} , Optimal Feature set OP_f[ ] is the best subset with optimal number of features which has highest fitness value, Threshold limit for maximum distance between every pair of features defined as T_{dist}, obtained distance between the features denoted as F_{dist}, Global best value as Gbest it stores the best subset among all the iteration , Local best value as $Lbest_k$ it stores best fitness value for every iteration, Number of selected features is NF, Classification Accuracy is CA and Total Feature cost denoted as FC.

After initialization of parameters the subset generation is performed. The binary cuckoo search algorithm generates k number of subsets S_k[]={f_1,f_2,f_3,…..f_m), where m is number of selected features in subset k from given original input feature set f_n[]={f_1,f_2,..f_n}, where n is number of input features. The Host egg f_j is randomly assigned to nest S_k[], remaining eggs are considered as source egg f_i. For every iteration distance between source and host egg is calculated using RMSD function, if the dist($host_{egg}$, $source_{egg}$) is less than the threshold distance T_{dist} then that $souce_{egg}$ f_i is added to the nest otherwise the egg will be destroyed by cuckoo search, it assigns binary value 1 to selected $source_{egg}$ and 0 to destroyed $source_{egg}$.

In sudset evaluation the fitness value for all the subset S_k[f_1,$f_{2,…}$$f_m$] is calculated using number of features (NF), The SVM classification accuracy value (CA) and the cost of each features (FC). The fitness value is assigned to $Lbest_k$ then which is compared with Gbest then the best fitness value is assigned to Gbest. Finally the best optimal feature subset is generated OP_k[f f_1,$f_{2,…}$$f_m$].

Particle Swarm Binary Cuckoo Optimization (PSBCO) Algorithm

```
INPUT   : Number of Input Features  f(n)={f1, f2, …. fn}
OUTPUT: Reduced Number of Optimal Feature set
             $OP_f$[ ] = $S_k$[ ]

Step 1:  Initialization

          Input feature set f(n)={$f_1$, $f_2$, …. $f_n$}
          n →    No. of features in input dataset
          Generate Initial population of k host nests
          k → No. of possible subsets (clusters)
          Subset $S_k$[ ] = { $f_1$, $f_2$, …. $f_m$}
          m → No. of features in subset $S_k$[ ]
          $OP_f$[ ]={ }, $T_{dist}$ = 1, Gbest=0, $Lbest_k$=0;
          $F_{dist}$=0, NF=0, CA=0, FC=0;

Step 2:  Subset Generation

          While (k >n) or (stop criterion)
 // Randomly select nest  $S_k$ among k →($S_k$,k)
             For(j=1; j≤n; j++)
               {
                 For(j=1; j≤n; j++)
                   {
 //Host bird abandon the egg(feature ($f_i$)) if it notice difference
   between the eggs ($f_i$ &$f_j$)
                       $F_{dist}$= RMSD($f_i$, $f_j$);
                          If($F_{dist}$ < $T_{dist}$)
                            Set $f_i$=1
//Feature is added to subset $S_k$[ ]
                          Else
                              Set $f_i$=0
//Feature is deleted
                   Subset $S_k$[ ] = {$f_i$};       Where  i=1to n;
                     }
                  }
Step 3:   Subset Evaluation

// Fitness function for randomly selected subset $S_k$[]
             For(k=1; k≤n; k++)
                  {
                     $Lbest_k$ =  $Fitness_k$ (NF * CA * FC);
                         If($Lbest_k$ > Gbest);
                              {
                                 Gbest = $Lbest_k$;
                                 $OP_f$[ ]= $S_k$[ ];
                              }
                    }
```

Fig. 2. PSBCO Algorithm

IV. EXPERIMENTAL RESULTS

We used Heart Disease Data Set from UCI machine learning database for experiment. The experimental result shows that the optimality of the proposed algorithm is better than the existing cuckoo search algorithms. The original dataset consists of 15 features after applying PSBCO algorithm to the original dataset the optimal subset is generated OPk[] = { 0 0 1 1 0 0 1 0 1 1 1 …..} which consists of 10 feasible feature in order to produce optimal result.

Sample input attributes (from UCI dataset)

1 id: patient identification number
2 ccf: social security number (I replaced this with a dummy value of 0)
3 age: age in years
4 sex: sex (1 = male; 0 = female)
5 painloc: chest pain location (1 = substernal; 0 = otherwise)
6 painexer (1 = provoked by exertion; 0 = otherwise)
7 relrest (1 = relieved after rest; 0 = otherwise)
8 pncaden (sum of 5, 6, and 7)
9 cp: chest pain type
-- Value 1: typical angina. -- Value 2: atypical angina
-- Value 3: non-anginal pain. -- Value 4: asymptomatic
10 trestbps: resting blood pressure (in mm Hg on admission to the hospital)

The sample output optimal subset attributes:

1. 3 (age)   2. 4 (sex)   3. 9 (cp)
4. 10 (trestbps)   5. 12 (chol)   6. 16 (fbs)
7. 19 (restecg) 8. 32 (thalach)   9. 38 (exang)
10. 40 (oldpeak)   11. 41 (slope)   12. 44 (ca)
13. 51 (thal)   14. 58 (num)

The original dataset and PSBCO algorithm given as input to MOA tool and we obtained optimal feature set with reduced number of features. The selected optimal features are given as input to the weka tool for performance analysis. The performance is analysed by 3 different classifiers and result value is compared with existing feature selection algorithms such as Binary Cuckoo search (BCS), Genetic Cuckoo Optimization Algorithm (GCOA) and Binary Particle Swarm Optimization (BPSO). The comparison results are shown in fig. 3. and table 1. From the Experimental results the proposed algorithm performs better with minimum number of features and gives better accuracy.

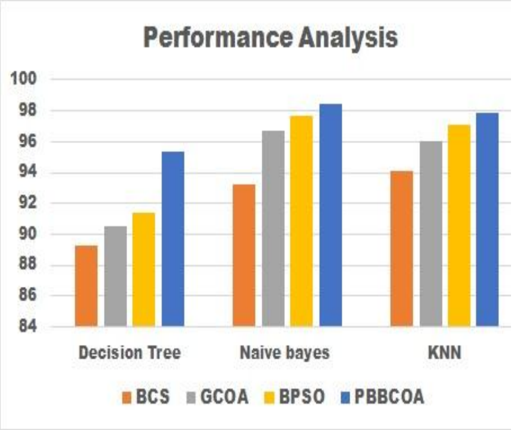

Fig. 3. Performance Analysis

Table 1

Technique	**Execution Time [sec]**	**Number of Features Selected**	**Classification Accuracy**
BCS	764.63476	14	93.22%
GCOA	711.17643	14	96.74%
BPSO	685.91677	12	97.66%
GBCOA	678.49871	11	98.38%
PBBCOA	675.34512	10	98.42%

V. CONCLUSION

In this paper we addressed the task of feature selection as an optimization problem and proposing a hybrid meta-heuristic feature section algorithm called Particle Swarm Binary Cuckoo optimization (PSBCO) algorithm. The experiments were carried out over Heart Disease Data Set aiming to detect presents of heart disease. We have shown how to predict the heart disease from patient's Records with minimum number of features. The proposed algorithm reduces the number of feature which reduce computation time and cost also provides better accuracy. In future work the algorithm can be implemented for different medical datasets.

REFERENCES

[1] Elsayed abd el aziz, Mohamed & Hassanien, Aboul Ella. (2016). "*Modified cuckoo search algorithm with rough sets for feature selection*". Neural Computing and Applications. 10.1007/s00521-016-2473-7.

[2] B. Wei, W. Zhang, X. Xia, Y. Zhang, F. Yu and Z. Zhu, "*Efficient Feature Selection Algorithm Based on Particle Swarm Optimization With Learning Memory*," in IEEE Access, vol. 7, pp. 166066-166078, 2019.

[3] F. Koumi, M. Aldasht and H. Tamimi, "*Efficient Feature Selection using Particle Swarm Optimization: A hybrid filters-wrapper Approach*," 2019 10th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2019, pp. 122-127.

[4] L. Brezočnik, "*Feature selection for classification using particle swarm optimization,*" IEEE EUROCON 2017 -17th International Conference on Smart Technologies, Ohrid, 2017, pp. 966-971.

[5] Ahmad, Iftikhar. (2015). "*Feature Selection Using Particle Swarm Optimization in Intrusion Detection. International Journal of Distributed Sensor Networks*". 2015. 1-8. 10.1155/2015/806954.

[6] Ding, Y., Zhou, K. & Bi, W. "*Feature selection based on hybridization of genetic algorithm and competitive swarm optimizer".* Soft Comput (2020). https://doi.org/10.1007/s00500-019-04628-6

[7] Khushboo Jain and Anuradha Purohit. "*Feature Selection using Modified Particle Swarm Optimization*". International Journal of Computer Applications 161(7):8-12, March 2017.

[8] S.Sandhiya and U.Palani, "*A Novel hybrid genetic binary cuckoo optimization algorithm for feature selection*", International Conference on Recent trends in science, engineering and management (ICRTSEM), 2019.

[9] X. Chen, G. Yuan, F. Nie and Z. Ming, "*Semi-Supervised Feature Selection via Sparse Rescaled Linear Square Regression,*" in IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 1, pp. 165-176, 1 Jan. 2020.

[10] Y. KAYA, "*Comparison of Using the Genetic Algorithm and Cuckoo Search for Feature Selection*," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), Malatya, Turkey, 2018, pp. 1-5.

[11] Q. Al-Tashi, S. J. Abdul Kadir, H. M. Rais, S. Mirjalili and H. Alhussian, "*Binary Optimization Using Hybrid Grey Wolf Optimization for Feature Selection*," in IEEE Access, vol. 7, pp. 39496-39508, 2019.

[12] C. Ruengdetkhachorn and D. Lohpetch, "*Feature Selection using Parallel Cuckoo Algorithm with Naïve Bayes Classifier based on Two Different Strategies*," 2018 22nd International Computer Science and Engineering Conference (ICSEC), Chiang Mai, Thailand, 2018, pp. 1-4.

[13] Q. Wu, Z. Ma, J. Fan, G. Xu and Y. Shen, "*A Feature Selection Method Based on Hybrid Improved Binary Quantum Particle Swarm Optimization,*" in IEEE Access, vol. 7, pp. 80588-80601, 2019.

[14] U. F. Siddiqi, S. M. Sait and O. Kaynak, "*Genetic Algorithm for the Mutual Information-Based Feature Selection in Univariate Time Series Data,*" in IEEE Access, vol. 8, pp. 9597-9609, 2020.

[15] Pereira, Luís & Rodrigues, Douglas & Almeida, T. & Ramos, Caio & Souza, André & Yang, Xin-She & Papa, João. (2014). "*A Binary Cuckoo Search and Its Application for Feature Selection*". 10.1007/978-3-319-02141-6_7.