International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-6, March 2021

An Optimization of Feature Selection for Classification using Bat Algorithm

V. Yasaswini, Santhi Baskaran

Abstract *Data mining is the action of searching the large existing database in order to get new and best information. It plays a major and vital role now-a-days in all sorts of fields like Medical, Engineering, Banking, Education and Fraud detection. In this paper Feature selection which is a part of Data mining is performed to do classification. The role of feature selection is in the context of deep learning and how it is related to feature engineering. Feature selection is a preprocessing technique which selects the appropriate features from the data set to get the accurate result and outcome for the classification. Nature-inspired Optimization algorithms like Ant colony, Firefly, Cuckoo Search and Harmony Search showed better performance by giving the best accuracy rate with less number of features selected and also fine f-Measure value is noted. These algorithms are used to perform classification that accurately predicts the target class for each case in the data set. We propose a technique to get the optimized feature selection to perform classification using Meta Heuristic algorithms. We applied new and recent advanced optimized algorithm named Bat algorithm on UCI datasets that showed comparatively equal results with best performed existing firefly but with less number of features selected. The work is implemented using JAVA and the Medical dataset (UCI) has been used. These datasets were chosen due to nominal class features. The number of attributes, instances and classes varies from chosen dataset to represent different combinations. Classification is done using J48 classifier in WEKA tool. We demonstrate the comparative results of the presently used algorithms with the existing algorithms thoroughly.*

Index Terms: *Optimization, Meta-heuristic, Feature Extraction, Deep learning*

I. INTRODUCTION

Data Mining [1] is the way of searching important information from the huge present all over in the repository. Data Mining falls in to two ways namely Association and Classification analyzing methods. Optimization algorithm provides a systematic way of developing and leveling new solutions to gain an optimal result. The optimization process must only be used in those problems where there is a specific need of accomplishing a quality or a competitive work. It is expected that the solution obtained through an optimization method is better than other results in terms of the selected objective.

This paper shows the Bat algorithm and Modified Bat algorithm accuracy rates when compared to existing algorithms namely Firefly, Cuckoo search and Harmony Search algorithms that showed almost equal results of the best accuracy rates in existing work. There are various applications with respect to data mining and optimization techniques in different fields. This method proves the better analysis which gives the best results and improved accuracy. The following are the different field of applications.

1. Network Security
2. Computer Vision and Processing
3. Nature Inspired fields.
4. Medical Fields
5. Transition Probabilities for Radio Systems
6. Intrusion Detection
7. Education
8. Financial Banking

II. OVERVIEW ON DATAMINING

Data mining process involves the following stages.

a) Problem Definition. In this stage the analysis of the problem in the business problem is done and tries to get the clear idea of the problem to be solved. This takes some time to make an exact definition of the problem and it does not require any data tools.

b) Exploration of Data. In this stage data is explored by identifying quality problem to understand the metadata meaning. It is next level of problem definition stage which frequently exchanges the data.

c) Preparation of Data. In this stage data model is built after the exploration of data. Data is collects, clear the unwanted data and arrange the data in a format like tables and records.

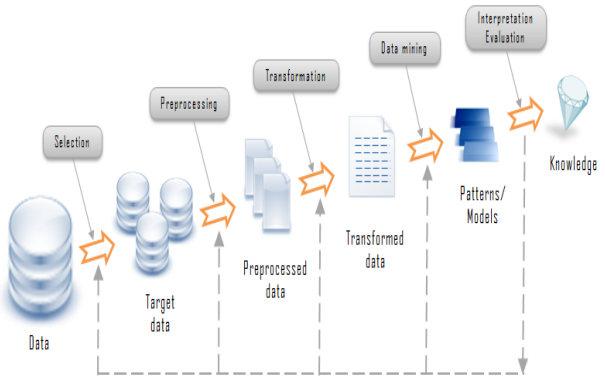
d) Data Modeling. At this stage after preparation of information, different mining functions are applied to the same kind of data. A high quality of mining model is prepared based on the changes in the parameters until we get optimal data model. Finally the good quality model is built and evaluated.

e) Evaluation of the Model. In this stage the evaluated model is checked and tested whether the quality is good or not and objective is satisfied or not?

f) Deployment. In this stage after the evaluation of data, the exporting of the data is done and the results are checked into database tables.

Manuscript received on January 30, 2020.
Revised Manuscript received on February 11, 2021.
Manuscript published on March 30, 2021.

V. Yasaswini, Research Scholar, Computer Science and Engineering Department, Pondicherry Engineering College, Puducherry, India.
Santhi Baskaran, Professor& Head, Information Technology Department, Pondicherry Engineering College, Puducherry, India



III. DESCRIPTION OF ALGORITHMS

A. EXISTING WORK ALGORITHMS-FIREFLY ALGORITHM:

Firefly Algorithm (FA)[9] being a Nature Inspired algorithm works based on flashing nature of the fireflies. The main reason for its flash is to move as a system which provides signal to absorb the other fireflies towards itself. The algorithm was implemented to perform Feature Selection (FS) for Image Processing and Eigen Value Optimization problem along with other related domains and has been performed so that better results are obtained. In order to achieve best optimal feature subset increases the predictive accuracy of the classifier. In this algorithm, along with considering the brighter firefly to obtain the predictive accuracy of the dataset we have also considered a comparatively brighter firefly and the predictive accuracy of that method have also been calculated.

- **Choosing a Brightest Firefly**
- **Choosing a Comparatively Brighter Firefly**

It includes previously chosen firefly solution and the newly selected brighter firefly solution in our computation. The latest solution is found by solving the below formulation

$$New\ X_i = X_i + \beta_0 \exp\left(-\gamma(r_{ij})^2\right)(X_i - X_j) + \alpha\varepsilon_i$$

Where, **Parameter Settings of Firefly Algorithm**

X_i	The solution pointed by the current firefly (Classification Accuracy)
Xj	The solution pointed by the brightest firefly (First Method) The solution pointed by the comparatively brighter firefly (Second Method)
β_0	Between 0 and 1
γ	Between 0.1 and 10
r_{ij}	Distance is fixed to 1
α	Parameters selected within the range [0,1] randomly

1. First iteration (i)
2. FA arguments are initialized
3. Calculate the Individual feature fitness
4. Iterate it again
5. Build the results from the fireflies (X_i)
   - **Construct the "BRIGHTEST FIREFLY"**
   - Subset of features is assigned for configuration (binary bit string) to each firefly
   - Develop latest feature subset
   - Traverse the developed feature subset to the present classifier
   - Calculate the fitness of the feature subset by computing the new value of current firefly
   - Reset the value of current firefly is based on either consideration of rejection of the "BRIGHTEST FIREFLY"
6. Evaluate the good feature subset in the present iteration
7. Iteration i = i + 1
8. Iterates till it reaches its final range
9. Utilize the same procedure of searching to develop the best optimal feature subset.

Implementation of Firefly Algorithm

B. EXISTING WORK ALGORITHM-CUCKOO SEARCH ALGORITHM:

Cuckoo Search Algorithm (CSA) [10] being idealized by its breeding behavior was tested on engineering optimization and embedded design problems. The results being obtained by this algorithm for engineering optimization problems are quite convincing in its results. Hence forth, Cuckoo Search Algorithm for FS is implemented. In this Cuckoo Search Algorithm addition of three constraints such as Eviction, Abandon and Survival is done. This type of method uses its historical memories for the location and status of the eggs being laid by the cuckoos.

This Algorithm Mainly Concentrates On Replacing Not Good Nests With The Potentially Good Nests. The position of the egg replaces the position of the random new eggs in another nest in three cases. if the eggs have been evicted, if the cuckoos abandon the nest or if the eggs have been hatched resulting to its survival.

- **Hatching of eggs – Survival of the Cuckoo**
- **Abandoning the nest – Host bird abandons its nest and migrates to some other place to build another nest.**
- **Evicting the eggs – The host bird throws the cuckoo bird's eggs.**

Equation has been used as the main formula for the computation of Cuckoo Search Algorithm.

$$F_{ij} = \frac{\{(\propto[I_i(next) - I_i(bp)]) * iter\}}{maxcuckoo}$$

Where, **Parameter settings for Cuckoo Search Algorithm**

F_{ij}	Fitness function used to find the alpha value of the cuckoo
I_i (*next*)	Accuracy of the cuckoo bird being selected
I_i (*bp*)	Accuracy of the host cuckoo bird
Iter	Fixed to 1
$\propto$	Between 0 and 4
Maxcuckoo	Maximum number of cuckoos in a particular dataset

1. Iteration(i) = 1
2. The CSA arguments are initialized
3. Generate the initial population of Host Birds' nests and the Cuckoo Birds' nests
4. Evaluate the status of eggs with its own feature
5. Iterate it again
6. Build the new result based on fitness of the cuckoos eggs (F_{ij})
   - Construct the "STATUS OF THE EGGS – Eviction, Abandon, Survive"
   - Choose the cuckoo according to the alpha value computed in the fitness function by comparing with the alpha value of the other features
   - Subset of features is assigned for configuration (binary bit string) to each egg of the cuckoo
   - Develop latest feature subset
   - Traverse the developed feature subset to the present classifier
   - Calculate the fitness of the feature subset by computing the new value of F_{ij}
   - Compute the new solution of the feature and find the new solution based on its accuracy done through classification
   - Reset the value of F_{ij} based on the rejection either by eviction or abandon or survive
7. Evaluate the good feature subset in the present iteration
8. Iteration i = i + 1
9. Iterates till it reaches its final range
10. Utilize the same procedure of searching to develop the best optimal feature subset.

Implementation of Cuckoo Search Algorithm

3.3 EXISTING WORK ALGORITHM-HARMONY SEARCH ALGORITHM:

The underlying principle behind this HSA algorithm [11] is has been that this algorithm might face the search on the grounds of Pitch, Amplitude and Timbre producing a perfect harmony. The initial random solutions that are considered may be far away from the feasible solutions. To get closer to the feasible and promising solutions, the exact solutions can be obtained by choosing only the amplitude of the tone.

In this method, the best accuracy of Tunes is found with the frequency value of Tunes which varies with time (t) and computation is done for the new solution.

In HSA, Individual music player (variable which takes decision) runs a bit of music (value) which finds the best harmony (global optimum) at the last. Based on the Lambda value two constraints will be considered. Noise (high pitch) and Melody (low pitch).

Equation has been used as the main formula for the computation of Amplitude in Harmony Search Algorithm

$$\boldsymbol{\lambda} = \frac{\boldsymbol{C}}{\boldsymbol{h_i[next]} * \boldsymbol{h_i[music]}}$$

Where,

Parameter Settings for Harmony Search Algorithm

$\boldsymbol{\lambda}$	Lambda Value (Amplitude)
$\boldsymbol{C}$	Velocity of Light
$\boldsymbol{h_i[next]}$	Accuracy of newly selected tune
$\boldsymbol{h_i[music]}$	Accuracy of the existing tune that needs to be compared with the new tune

1. Iteration(i) = 1
2. The HSA arguments are initialized
3. Generate initial population tune for a perfect harmony
4. Evaluate the amplitude of the tone (Lambda) of its own feature
5. Iterate it again
6. Build the result by the Lambda value of the tones ($\boldsymbol{\lambda}$)
   - Construct the "AMPLITUDE OF THE TUNES – Noise or Melody"
   - Compare frequency of each harmonic tune and generate similar frequency of harmonic tunes
   - Choose the tune according to the lambda value being computed through the amplitude by comparing with the lambda value of the other tunes (features)
   - Develop latest feature subset
   - Traverse the developed feature subset to the present classifier
   - Calculate the fitness of the feature subset by computing the new value of $\boldsymbol{\lambda}$
   - Compute the new solution of the feature and find the new solution based on its accuracy done through classification
   - Get the latest harmonics (solutions) if it shows good results
   - Calculate the latest solution and find the new solution based on its accuracy done through classification
   - Reset the value of $\boldsymbol{\lambda}$ based on the rejection either filtering by noise or by melody
7. Evaluate the good feature subset in the present iteration
8. Iteration i = i + 1
9. Iterates till it reaches its final range
10. Utilize the same procedure of searching to develop the best optimal feature subset.

Harmony Search Algorithm

IV. PROPOSED WORK ALGORITHM-BAT ALGORITHM APPLIED IN FEATURE SELECTION.

The primary use of Bat Algorithm (BA)[12] is variant behavior of the micro bats based up on the frequency tuning, velocity v^t_i and its location x^t_i and calculates based on the iteration t with d-dimensional search or solution space.

According to Yang, the following mathematical equation is updated and written as

$$f_i = f_{min} + (f_{max}-f_{min})\beta$$
$$v_i^t = v_i^{t-1} + (x_i^t - x^*)f_i$$
$$x_i^t = x_i^{t-1} =+ v_i^t$$

where β ε [0,1] which uniformly distributes the random vector.is a random vector drawn from a uniform distribution. A direct exploitation for searching local solutions which modifies the present good result.

$$x_{next} = x_{prev} + \varepsilon A^t$$

From the above equation, ε refers to range of [-1,1] which can be any number between -1 and 1. A^t refers to the calculation of overall best loudness mean. During the iterations he emission of pulse from bats and its loudness is variant which is calculated based on the following equation,

$$A_i^{t+1} = \alpha A_i^t$$
$$\gamma_i^{t+1} = \gamma_i^0[1-\exp(-\gamma t)]$$

Where $0 < \alpha < 1$ and $\gamma > 0$ are constants.

Objective Function: Count of Bats $X_i = X_{i1}$ to X_{iD} to the power of T. where I belongs to the range [1,Np] [13]

1. Initialize the count of Bats in the Search space X_i and V_i
2. Initialize the frequencies (f_i), Pulses (r_i) and the Loudness (A_i)
3. while (t <Max number of iterations)
   Generate new solutions by adjusting frequency,
   Update velocities and locations/solutions
4. if (rand > r_i)
   Select a solution among the best solutions
   Generate a local solution around the selected best solution
   end if
   Generate a new solution by flying randomly
5. if (rand < A_i & $f(x_i) < f(x_*)$)
   Accept the new solutions
   Increase r_i and reduce A_i
   end if
6. Rank the bats and find the current best x_*
   end while

In this algorithm tuning of frequency is compared with the mutation of bats in the local search space if the mutation is large then it leads to global search. In essence, frequency tuning essentially acts as mutation because it varies the solutions mainly locally. For every iteration the current best solution is found and finally we get x_*. Mutation varies due to changes in loudness and Pulse signals from the bats.

V. EXPERIMENTAL SETUP AND RESULT ANALYSIS

Fourteen standard datasets drawn from the UCI collection were used in the experiments. These datasets were chosen due to nominal class features. The number of attributes, instances and number of classes vary in the chosen dataset to represent different combinations. All the features in 10 fold cross validation is run through Weka tool to get classification accuracy. The classifier used for evaluating the feature subsets generated is J48, Naïve Bayes and Logistic. Feature subset (FS) generation by Firefly Algorithm (FA), Cuckoo Search Algorithm (CSA) and Harmony Search Algorithm (HSA) is implemented using Net Beans IDE in the existing work. Feature subset generation by Nature-inspired algorithm named Bat Algorithm (BA)has been implemented using Net Beans IDE while the UCI dataset of Medical field is run through WEKA tool to get the classification rate which is processed through J48 classifier in the proposed work.

Table 1. Dataset Description

DATASETS	INSTANCES	FEATURES	CLASS
Heart-C	303	14	02
Dermatology	366	34	06
Hepatitis	155	19	02
Lung Cancer	32	56	02
Pima Indian Diabetes	768	08	02
Iris	150	04	03
Lymphography	148	18	04
Diabetes	768	09	02
Heart-Stalog	270	13	02
Audiology	226	74	10

Table.1 describes the different types of diseases in medical field and instances, Features and Class of each disease. For example, in this dataset Iris disease can be diagnosed by 4 features and it falls in to 3 classes.

Table 2. Comparison of all the accuracy of Existing with Proposed Algorithms

Datasets	FA–FS FA1(%)	FA–FS FA2(%)	CSA – FS (%)	HSA – FS (%)	BA-FS(%)
Heart-C	83.15	83.07	78.217	79.53	82.983
Hepatitis	69.03	67.00	64.516	67.74	68.128
Lung Cancer	**89.50**	89.33	84.375	87.37	88.872
Pima	78.70	76.10	75.651	77.47	77.561
Iris	**96.00**	**96.00**	96.00	96.00	94.00
Lymphography	**84.10**	82.25	78.378	84.43	81.404
Diabetes	77.47	76.00	75.65	77.21	77.430
Dermatology	**96.72**	96.17	95.901	96.07	95.985
Heart-Stalog	**84.44**	83.39	80.74	84.81	82.775
Audiology	79.051	79.201	79.646	77.87	78.758

It is clear from the table 2 that the accuracy of the Bat algorithm is best when compared to the existing algorithms.

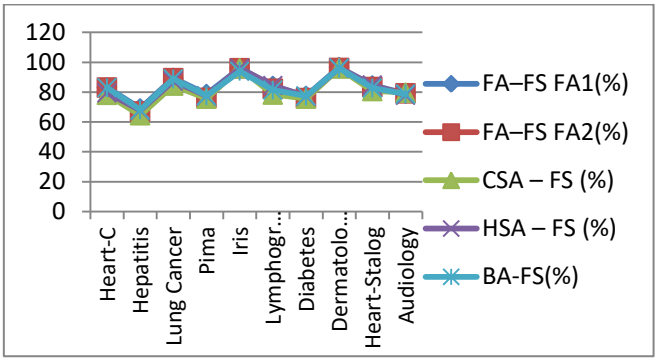


Fig. 1. Graphical Representation of Accuracy of all these algorithms

From Fig.1, we can observe that the Bat algorithm which is obtained for 10 UCI datasets showed better accuracy when compared to other existing algorithms.

Table 3. Comparison of all the features selected in Existing with Proposed Algorithms

Datasets	FA–FS	CSA – FS	HSA – FS	BA-FS
Heart-C	6	7	**5**	6
Hepatitis	**8**	10	9	**8**
Lung Cancer	**14**	20	18	16
Pima	**3**	**3**	4	4
Iris	**2**	3	**2**	**2**
Lymphography	8	**7**	10	8
Diabetes	**4**	5	6	5
Dermatology	23	22	25	**21**
Heart-Stalog	**7**	6	**7**	8
Audiology	53	55	49	56

In this table 3, it is clear that the number of features selected for Bat algorithm is best when compared to existing algorithms.

VI. INFERENCE

- From the table.2, We can infer that Bat Algorithm applied for UCI dataset for FS with respect to Accuracy gives better as well equal accuracies
- By comparing Firefly Algorithm, Cuckoo Search Algorithm and Harmony Search Algorithm for FS with respect to Features: we can infer that the features for Bat Algorithm gets reduced as well gets increased in some of the datasets in Bat Algorithm.

VII. CONCLUSION

With the emerging trends of Information Technology, Data Mining throws out its purpose to each individual spotting out only the relevant features by the Feature Selection process along with Classification from eons of database to reach the final outcome. The proposed system for FS Optimization to perform Classification is applied and the results have been obtained using UCI datasets. The datasets are taken from UCI repository and Table 1 describes the 10 datasets that we have used. The UCI datasets are used in continuing works in literature; such has Classification, FS and Classifier Ensemble, Stacking Ensemble, so we have adopted these datasets for our work. This paper has attempted to give accuracy results of the several popular Meta-Heuristic algorithms. Finally analyzing the proposed algorithms compared to existing algorithms we found that Bat algorithm showed better results in some particular data in the dataset.

REFERENCES

1. Tan, Steinbach, Kumar. (2005). "Introduction to Data Mining".
2. Hassan AbouEisha et.al, (2018) "Extensions of Dynamic Programming for Combinatorial Optimization and Data Mining"
3. Sunil Kawale,*"Datamining and Optimization Techniques"* International Journal of Statistika and Mathematika*"*, ISSN. 2277-2790, E-ISSN. 2249-8605, Volume 6, Issue 2, 2013 pp 70-72
4. Nidhi Tomar and Prof. Amit Kumar Manjhvar "*A Survey on Data mining optimization Techniques"* International Journal of Science Technology & Engineering | Volume 2 | Issue 06 | December 2015 ISSN (online). 2349-784X
5. Basturk B, Karaboga D (2006) "*An artificial bee colony (ABC) algorithm for numeric function optimization"*. IEEE Swarm Intelligence Symposium, 12–14 May, Indianapolis
6. Bergh F, Engelbrecht AP (2006) "*A study of particle swarm optimization particle trajectories"*. Inf Sci 176. 937–971.
7. Rao, R. Venkata. "Teaching Learning Based Optimization Algorithm. And Its Engineering Applications". Springer, 2015.
8. Rao, R. Venkata, and V. D. Kalyankar. *"Parameter optimization of modern machining processes using teaching–learning-based optimization algorithm"*. Engineering Applications of Artificia Intelligence 26, no. 1 (2013). 524-531.
9. Shunmugapriya .P and Kanmani S, P.Sindhuja, G.Koperundevi, V.Yasaswini, "*Firefly Algorithm Approach for the Optimization of Feature Selection to Perform Classification*", International Conference on Advances in Engineering & Technology, IEEE-ICAET 2014.
10. Xin-She Yang, Suash Deb, "*Cuckoo Search Via Levy Flights*", World Congress On Nature and Biologically Inspired Computing (NaBIC 2009)
11. Xin-She Yang and X. He. "*Bat algorithm: Literature review and applications*". International Journal of Bio-Inspires Computation, 5(3):141-149, 2013.
12. Richardson, P.: The secrete life of bats. http://www.nhm.ac.uk

AUTHOR PROFILE

V Yasaswini, Research Scholar, Working in the data mining field to optimize feature selection using nature inspired algorithms.I am doing my research from the past 3 years and obtained aggregate results. Pondicherry Engineering College, Puducherry