Feature Selection for Prediction using Cuckoo Search Optimization

Aparna A.[1] Dr. Angelina Geetha[2]
[1,2]Crescent University, India

Abstract— Due to emerge of large data repositories in the current world were large data are emerging in terms of millions on daily basis. These emerging data's are to be optimized so that the data can be used in several useful purposes. Optimization methods can be used in feature selection methods to resolve the most related subset of features from the data set with adequate accuracy rate from the original data set of features. Several bio-inspired algorithms which are based on the behavior of nature living beings. In this paper, we have proposed a method of combining feature selection with one of the bio-inspired algorithm for optimizing and further the optimized data with more accuracy are used for prediction.

Key words: Bio-Inspired Algorithms, Feature Selection

I. INTRODUCTION

Data analysis intend at extracting and modeling information content data to identify patterns within the data. As a manner of making simple the amount of information to describe a large set of data, features in the data are extracted, which serves as representing characteristics trait of its contents. Feature selection is an important step used for several tasks, such as image classification, cluster analysis, data mining, pattern recognition, image retrieval and finding accuracy. It is a decisive preprocessing technique for effective data analysis, where only a subset from the original data features is chosen which can eliminate noisy data, irrelevant data or redundant features. This task can be used to reduce computational cost and improve accuracy of the data analysis process.

This paper proposes a feature selection method for data analysis and to find accuracy based on Cuckoo Search Optimization (CSO) approach that can be used in several knowledge domains through wrapper and forward strategies. The CSO Algorithm has been widely used to solve optimization problems; how-ever, there have been few works on feature selection. Our work proposes a binary version of the CSO algorithm, where the number of new features to be analyzed with behavior of cuckoo birds which lays their eggs in other birds nest and will increase its population which was proposed by Yang and Deb.

Experimental results show that a reduced number of features can be achieved classification accuracy better than that using the full set of features. The accuracy has significantly increased even though the number of selected features has drastically reduced. Furthermore, the proposed method presented better results for the majority of the tested data sets.

The paper is organized as follows: Initially, some relevant concepts and work related to feature selection are described. The proposed methodology for feature selection is then presented in detail. Experimental results obtained through the application of the proposed method to several data sets are described and discussed. Finally, the remaining section concludes the paper with final remarks and directions for future work.

A. Related Concepts & Work

The process of feature selection is responsible for choosing a subset of features, which can be described as a search into a state space. One can perform a full search in which all the spaces are traversed; however, this approach is impractical for a large number of features. A heuristic search considers the features, not yet selected at each iteration, for evaluation. A random search generates random subsets within the search space, such that several bio-inspired and genetic algorithms use this approach.

Feature selection can be depict as a search into a space of states, and according to the initialization and behavior during the search steps, we can divide the search into three different approaches forward: the feature subset is initialized empty and features are included in the subset during the feature selection; backward: the feature subset is initialized with a full set of features and the features are excluded from the subset during the feature selection process; bidirectional: features can be inserted or excluded during the feature selection process.

Feature selection methods can be classified into two main categories: filter approaches and wrapper approaches. In filter approaches, a filtering process is performed before the classification process; there-fore, they are independent of the used classification algorithm. A weight value is computed for each feature, such that those features with better weight values are selected to represent the original data set. On the other hand, wrapper approaches generate a set of candidate features by adding and removing features to compose a subset of features. Wrapper methods usually achieve better results than filter methods.

Many evolutionary algorithms have been used for feature selection, which include genetic algorithms and swarm algorithms. Swarm algorithms include, in turn, Cuckoo Search Optimization (CSO) Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), Bat Algorithm (BAT) and Artificial Bee Colony.

The use of Swarm Intelligence for feature selection has increased in the past years. Suguna and Thanushkodi proposed a rough set approach with CSO algorithm for dimensionality reduction using different medical data sets in the area of Dermatology for tests, whereas Shokouhifar and Sabet employed the same algorithm (CSO) for feature selection using neural networks. Particle Swarm Optimization has been proposed for feature selection either as filter method or as wrapper method. Nakamura et al. proposed a wrapper method using a BAT algorithm with OPF classifier.

Cuckoo Search Optimization (CSO) is one of evolutionary techniques. This algorithm is inspired by the lifestyle of a bird called the Cuckoo. This bird didn't make nest for itself a d it be used the nests of other birds for laying eggs. Ability to create eggs like the bird host is reinforced in cuckoo bird. If the bird's host discover eggs that are not mine, it throw away or leave the nest and it makes a nest in other places. Cuckoo eggs are the bigger size of the host bird until

cuckoo brood would hatch soon. When the host bird's eggs throws out of the nest or demand food so much to other broods die of hungry. When the cuckoo brood grows and becomes a mature bird continues the mother's life instinctively.

B. Generating Initial Cuckoo Habitat

In order to solve an optimization problem, it's necessary that the values of problem variables be formed as an array. In GA terminologies this array is called "Chromosome". But in CSO it is called "habitat". To start the optimization algorithm, a candidate habitat matrix is generated. Then some randomly produced number of eggs is supposed for each of these initial cuckoo habitats. In nature, each cuckoo lays between 5 to 20 eggs. These values are used as the upper and lower bounds of egg assigned to each cuckoo at different iterations. Other habit of real cuckoos is that they lay eggs within a maximum distance from their habitat. This maximum area will be called "Egg Laying Radius ELR". Each cuckoo has a egg laying radius ELR which is appropriate with the total number of eggs, number of current cuckoo's eggs a d also variable limits of var_{hi} and var_{low}. So ELR is defined as:

ELR=∞ * number of current cuckoo's / eggs total number of eggs * var_{hi}- var_{low}

C. Immigration of Cuckoos

When young cuckoos grow and become mature, they live in their own area and make society for some time. But when the season for egg laying approaches they move to new habitats with the most similar host eggs and with more food for new young birds. Then the cuckoo groups are formed in different areas, the society with the highest fitness value is selected as the goal point and other cuckoo to move to that point.

When mature cuckoos live in that environment identify cuckoos belong to which groups that are difficult. Now that cuckoo groups are identified their mean benefit value is calculated. The maximum amount of the benefit is determined by the goal group and consequently that group's best habitat is the new destination habitat for moving cuckoos. When moving toward goal point, the cuckoos do not fly all the way to the destination habitat. They only fly a part of the way and also have a deviation.

D. Pseudo Code of Cuckoo Search Optimization for Feature Selection

Begin
N= sample_count;
Selected_features={}
For p=1 to desired_feature_count
For c=1 to feature_count
Data_all=data(selected_features+feature(c));
For i=1 to class_number
Train_data_class(i)=partition(rand(data_all(class==i)),0.75)
Test_data_class(i)=parttion(rand(data_all(class=i)),others);
Objective Function f(x),x=$(x_1 \ldots,x_d)^{T;}$
Initial a population of n host nests x_i (i=1,2,…n),
While (t< maxgeneration) or (stop criterion)
get a cuckoo(say I) randomly by levyflights;
evaluate its quality/fitness Fi;
choose a nest among n(say j) randomly;
if(F_i>F_j)
replace j by the new solution;
end
Abandon a fraction pa of worse nest and built new ones at new locations via Levy's flight
Keep the best solutions and find the current best;
End while
Best_feature(c)= arg max(performance_criteria(feature(c))
Postprocess results and visualization
End

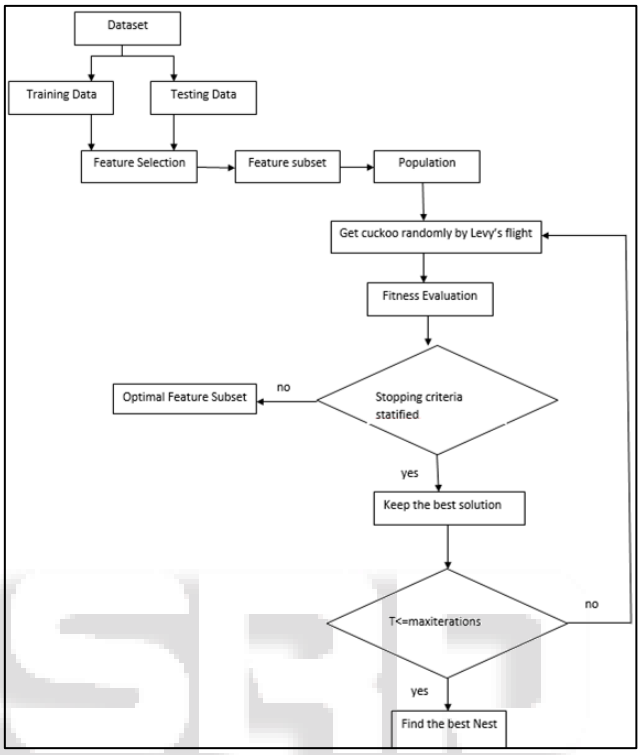

Fig. 1: Cuckoo Search Optimization with Feature Selection

E. Cuckoo Search Optimization for Feature Selection

Unlike optimization problems, where the possible solutions to the problem can be represented by vectors with real values, the candidate solutions to the feature selection problem are represented by bit vectors.

Each food source is associated with a bit vector of size N, where N is the total number of features. The position in the vector corresponds to the number of features to be evaluated. If the value at the corresponding position is 1, this indicates that the feature is part of the subset to be evaluated. On the other hand, if the value is 0, it indicates that the feature is not part of the subset to be assessed. Additionally, each food source stores its quality (fitness), which is given by the accuracy of the classifier using the feature subset indicated by the bit vector.

F. Data sets

The proposed method has been evaluated through ten data sets from different knowledge fields. The data sets are available from UCI Machine Learning Repository. Table 1 presents a description of the tested data sets, including the number of instances, number of features, and number of classes for each data set.

UCI data sets have been widely used in the evaluation of data classification since they contain a varied number of features and classes, allowing the analysis of

influence on accuracy and performance when features are selected

Data Set	Number of Instances	Original Number of Features	Selected Number of Features	Average accuracy
Breast Cancer	286	36	19	64
Diabetes	768	146	82	73
Heart Disease	810	110	92	94
Kidney Disease	435	20	12	94

Table 1: Results of Datasets before Implementation of Algorithm

Dataset	Number of Instances	Original Number of Features	Selected Number of Features	Average Accuracy from CSO
Breast Cancer	286	36	15	78
Diabetes	768	146	76	80
Heart disease	810	110	88	96
Kidney Disease	435	20	10	95

Table 2: Results of Datasets after Implementation of Algorithm

G. Computational Environment

All the experiments have been conducted on a computer with Intel Core I5, 4GB RAM. The cuckoo search optimization algorithm was implemented using python programming language with weka and nunmpy, panda's libraries to execute the data classification.

H. Classification Setup

To evaluate the accuracy and performance of the classification process with the original and selected feature sets, a tenfold cross -validation is used. In k-fold cross-validation, the data set is randomly partitioned into k equally sized folds (samples). One partition is retained as the test set, whereas the remaining k − 1 samples are used as the training set. This process is repeated k times, where one of the partitions becomes test data at each time. The average of k results produces an estimation of the accuracy. The accuracy measure employed for evaluating the results is the percentage of instances correctly classified, that is, for which a correct prediction was made.

II. CONCLUSIONS

This work presents a feature selection method based on ABC algorithm. The results show that a reduced number of features can achieve classification accuracy superior to that using the full set of features. For some data sets, the accuracy has significantly increased even though the number of selected features has drastically reduced. The proposed method presented better results for the majority of the tested data sets compared to other algorithms.

For future work, we plan to investigate alternative mechanisms to explore in finding the best nest solutions for best reproduction, parallelize the finding the most accuracy values which are used to implement in the algorithm for predictions in the medical fields using cuckoo search and feature selection with optimized data.

REFERENCES

[1] SamaneAsadi1, Vahid Rafe., International Advances in Engineering and Technology (IAET), Presenting a method for Clustering Using Cuckoo Optimization Algorithm (2014)
[2] Azizah Binti Mohamada, Azlan Mohd Zain, Soft Computing Research Group, Faculty of Computing, Universitiy Teknologi, Cuckoo Search Algorithm for Optimization Problems—A Literature Review and its Applications (2014)
[3] E. F. Shair, S. Y. Khor, A. R. Abdullah, International Review of Automatic Control (I.RE.A.CO.), A Brief Review of Cuckoo Search Algorithm (2014)
[4] N. KAMATCHI & D. SARAVANAN, International Journal of Mechanical and Production Engineering Research and Development (IJMPERD), A Hybrid Algorithm Using Firefly And Cuckoo Search Algorithm For Flexible Open Shop Scheduling Problem, (2017)
[5] Mansaf Alam and Kishwar Sadaf, Indian Journal of Science and Technology, Web Search Result Clustering based on Cuckoo Search (2016)

