FEATURESELECTION IN HIGH-DIMENSIONAL DATASETSUSING CUCKOO OPTIMIZATION ALGORITHM

[1]MAHMOOD MOGHADASIAN, [2]SEYEDEH PARVANEH HOSSEINI

[1]Islamic Azad University, Abadan Branch, Abadan, Iran
[2]Sama technical and vocational training college, Islamic Azad University, Ahwaz Branch, Ahwaz, Iran

Abstract- Feature selection is a process commonly used in machine learning. Based on Binary Cuckoo Optimization Algorithm (BCOA) and information theory, this paper proposes a new filter feature selection method for classification problems. The proposed algorithm is based on BCOA and the Mutual Information (MI) of eachpair of features, which determines the relevance and redundancy of the selected feature subset. Different weights for the relevance and redundancy in the fitness functions of the proposed algorithm are used to further improve their performance in terms of the number of features and the classification accuracy. In the experiments, an Artificial Neural Network (ANN) is employed to evaluate the classification accuracy of the selected feature subset on the test sets of six datasets. The results show that proposed algorithms can significantly reduce the number of features and achieve high classification accuracy in almost all cases.

Keywords- Feature Selection, Mutual Information, Binary Cuckoo Optimization, Filters.

I. INTRODUCTION

Many practical applications of classification involve a large volume of data and/or a large number of features/attributes. Since these datasets are usually collected for reasons other than mining the data (e.g. classification), there may be someredundant or irrelevant features. This is especially important when a large number of features exist and there are comparably few training sample data points, making feature vector dimensionality reduction an imperative.Feature selection involves finding a subset of features to improve prediction accuracy or decrease the size of the structure without significantly decreasing prediction accuracy of theclassifier built using only the selected features. Methods for feature selection have been essentiallydivided into two categories: filters and wrappers. In a wrapper approach,a learning algorithm is used as part of the evaluationfunction to determine the fitness of the selected featuresubset. Wrappers can usually achieve better results than filterapproaches, but the main drawbacks are their computational deficiency and loss of generality. In a filter approach,feature selection is done as a preprocessing procedure and thesearch process is independent of a learning algorithm. Therefore,the performance of a filter approach relies mainly on thegoodness of the evaluation criterion. Many different criteriahave been used in filter approaches, including information measures , dependency measures, consistency measures, and distance measures. Compared with wrappers, filter approaches are argued to be computationally less expensive and more general. Thesearch strategy can significantly influence the results of a feature selection approach. Many search techniques have been applied in feature selection such as greedy search, but most of them usually suffer from the problem of becoming stuck In local optima and/or high computational cost. Therefore, a computationally cheap global search technique is needed to develop a good feature selection algorithm. Evolutionary computation techniques are well-known for their global search ability, and have been applied to feature selection problems. These include particle swarm optimization (PSO), genetic algorithms (GAs) , genetic programming (GP) and COA. Compared with GAs and GP, COA is easier to implement, has fewer parameters, computationally less expensive, and can converge more quickly. Due to these advantages, COA has been used as a promising method for feature selection problems.

II. BACKGROUND

2.1. Cuckoo Optimization Algorithm

The parasite behavior of some species of Cuckoo is extremely intriguing. These birds can lay down their eggs in a host nests, and mimic external characteristics of host eggs such as color and spots. In case of this strategy is unsuccessful, the host can throw the cuckoo's egg away, or simply abandonits nest, making a new one in another place. Based on this context, Yang and Deb have developed a novel evolutionary optimization algorithm named as Cuckoo Optimization Algorithm (COA), and they have summarized COA using three rules, as follows:

- Only one egg at a time is laid by cuckoo. Cuckoo dumps its egg in a randomly chosen nest.
- The number of available host nests is fixed, and nests with high quality of eggs will carry over to the next generations.
- In case of a host bird discovered the cuckoo egg, it can throw the egg away or abandon the nest, and build a completely new nest.

Algorithmically, each host nest $n$is defined as an agentwhich can contain a simple egg x(unique dimension problem)or more than one, when the

problem concerns to multipledimensions. COA starts by placing the nest population randomlyin the search space. In each algorithm iteration, the nests areupdated using random walk via L´evy flights:

$$x_i^{(t+1)} = x_i^{(t)} + \alpha \oplus \text{Lévy}(\lambda) \quad (1)$$

and

$$\text{Lévy} \sim u = t^{-\lambda}, (1 < \lambda \leq 3) \quad (2)$$

Wheret is step size, and$\alpha > 0$is the step size scaling factor/parameter. Here the entry wise product $\oplus$is similar to those used in PSO, and $x_i^{(t+1)}$stands for the$(t+1)$th egg (feature) at nest i(solution), $i = 1, 2, \ldots, m$and$t = 1, 2, \ldots, d$. TheL´evy flights employ a random step length which is drawn from a L´evy distribution. Therefore, the CS algorithm is more efficient in exploring the search space as its step length is much longer in the long run. Finally, the nests which have eggs with the lowest quality are replaced to new ones according to a probability$p_a \in [0, 1]$.

In traditional COA, the solutions are updated in the searchspace towards continuous-valued positions. Unlike, in the BCOAfor feature selection, the search space is modeled as a $n$dimensionalboolean lattice, in which the solutions are updatedacross the corners of a hypercube. In addition, as the problemis to select or not a given feature, a solution binary vectoris employed, where 1 corresponds whether a feature will be selected to compose the new dataset and 0 otherwise. In orderto build this binary vector, we have employed the equation 4, which can provide only binary values in the boole an latticerestricting the new solutions to only binary values:

$$S(x_i^{(t+1)}) = \frac{1}{1 + e^{-x_i^{(t)}}} \quad (3)$$

$$\begin{cases} \text{if } S \leq \text{rand then } x_i^{(t+1)} = 0 \\ \text{if } S > \text{rand then } x_i^{(t+1)} = 1 \end{cases} \quad (4)$$

In which $\text{rand} \sim \text{rand}(0, 1)$and $x_i^{(t+1)}$denotes the new egg'svalue at time step t.

2.2. Entropy and Mutual Information

Information theory developed by Shannon provides away to measure the information of the random variables withentropy and mutual information.The entropy is a measure of the uncertainty of randomvariables. Let X be a random variable with discrete values,its uncertainty can be measured by entropy $H(x)$, which isdefined as

$$H(x) = -\sum_{x \epsilon X} p(x) \log_2 p(x) \quad (5)$$

Where $p(x) = Pr(X = x)$is the probability density functionof X. Note that entropy does not depend on actual values, butjust the probability distribution of the random variable.For two discrete random variables X and Y with theirprobability density function $p(x; y)$, the joint entropy $H(X; Y)$is defined as

$$H(X; Y) = -\sum_{x \epsilon X, y \epsilon Y} p(x, y) \log_2 p(x, y) \quad (6)$$

When a certain variable is known and others are unknown,the remaining uncertainty is measured by the conditionalentropy. Assume that variable Y is given, the conditionalentropy $H(X|Y)$ of X with respect to Y is

$$H(X|Y) = -\sum_{x \epsilon X, y \epsilon Y} p(x, y) \log_2 p(x|y) \quad (7)$$

Where $p(x|y)$ is the posterior probabilities of X given Y. From this definition, if X completely depends on Y, then $H(X|Y)$ is zero, which means that no more other information is required to describe X when Y is known. Otherwise, $H(X|Y) = H(x)$ denote that knowing Y will do nothing to observe X.

The information shared between two random variables is defined as mutual information. Given variable X, how much information one can gain about variable Y, which is mutual information $I(X; Y)$.

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = -\sum_{x \epsilon X, y \epsilon Y} p(x, y) \log_2 p(x|y) \quad (8)$$

According to equation 8, the mutual information $I(X; Y)$will be large if two variables X and Y are closely related. Otherwise, $I(X; Y) = 0$ if X and Y are totally unrelated.

III. METHODOLOGY

In this section, the BCOA filter based feature selection approach is proposed, where the entropy and mutual information are applied to calculate the relevance and redundancy in the feature selected subsets.

Mutual information is defined as the information shared between two random variables, which can be used in feature selection to evaluate the relevance between features and class labels. However, in feature selection, because of the interactions between features, the combination of m individually good features may not be the best combination of m features. Therefore, it is necessary to reduce the redundancy among features and select a subset of features with minimal redundancy to each other and maximal relevance to class labels. For this reason, both relevance and redundancy are included in the fitness function to guide BCOA to search for the best

feature subset, which can be represented by Equation 9.

$$\text{Fitness} = \alpha \times D - (1-\alpha) \times R \quad (9)$$

Where

$$D = \frac{1}{|S|}\sum_{i} I(x_i, C) \quad (10)$$

$$R = \frac{1}{|S|^2}\sum_{i}\sum_{j} I(x_i, x_j) \quad (11)$$

Where X is the set of selected features and c is the class label. Each selected feature and class labels are related as discrete random variables. D uses pair wise calculations to calculate labels, which determine the relevance of the selected feature subset to the class labels.R evaluates the mutual information shared by each pair of selected features, which indicates the redundancy contained in the selected feature subset. Fitness is a maximization function to maximize the relevanceD and simultaneously minimize the redundancy Rin the selected feature subset. αis a constant value and $\alpha\epsilon[0,1]$. α shows the relative importance of relevance in fitness function. $(1-\alpha)$ shows the relative importance of the reduction of redundancy. As the relevance is assumed to be more important than the redundancy, α is set to be larger$(1-\alpha)$ than in fitness function.

The proposed algorithm shows the pseudo-code of using BCOA evaluation for feature selection. The representation of a cuckoo in BCOA is a n-bit binary string, where n is the number of available features in dataset and also the dimensionality of the search space. In the binary string "1" represents that the feature is selected and "0" otherwise.

Algorithm: The BCOA based feature selection algorithm

```
begin
  divide Dataset into a Training set and a Test set;
  initialize the population of n host nests x_i (i = 1,2, ..., n);
while (t < MaxGeneration or stop Criterion) do;
move a cuckoo randomly via Lévy flights;
evaluate its fitness according to equation 9 on training set;
randomly choose nest among n available nests(for example j);
If(Fitness_j > Fitness_i)
            Replace j by the new solution
End if
Abandon a fraction (p_d) of worse nests
build new ones at new locations via equations1and 2;
        keep the best solution;
        Rank the solutions and find the current best;
end while;
calculate discrete values according to equation 3, 4;
calculate the classification accuracy of the selected
feature subset on the test set;
return the training and test classification accuracies;
end
```

IV. EXPERIMENTAL DESIGN

The data sets in this study consisted of six gene expression profiles, which were downloaded from http://www.gemssystem.org, whichcan be seen in Table I. The datasets were selected tohave different numbers of features, classes and instances asthe representative samples of the problems that the proposedapproaches can address.

In all datasets instances are randomlydivided into two sets: 70% as the training set and 30% asthe test set. Theproposed algorithms firstly run on the training set to select feature subsets and then the classification performance of theselected features will be calculated on the test set by a learningalgorithm.

There are many learning algorithms that can beused here, such as K-nearest neighbor (KNN), NB, and ANN.AnANN learning algorithm is selected in this study to calculatethe classification accuracy of the selected features accordingto Equation 12:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (12)$$

Where TP, TN, FP and FN stand for true positives, truenegatives, false positives and false negatives, respectively.The parameters of BCOA are set as follows:Number of initial cuckoos=5, Maximum number of iterations=1000, Maximum number of living cuckoos=30, Maximum number of Eggs laid by each cuckoo=4. Five different values for αused in the experiments, which are 0.9, 0.8, 0.75, 0.6and 0.5. For each dataset, the proposed approach has been conductedfor 20 independent runs.

TABLEI Dataset

Dataset	Samples	Classes	Genes
SRBCT	83	4	2308
9_Tumors	60	9	5726
Brain_Tumor2	50	4	10367
11_Tumors	174	11	12533
LungCancer	203	5	12600
14_Tumors	308	26	15009

V. RESULTS AND DISCUSSIONS

Experimental results are shown in Tables II. In this table "Genes selected"compares the number of the genes (features)that selected in different methods. "Classification accuracy" shows the average test accuracy of the selected feature subsets in the 20 runs.

According to the results shown in Table II, it can be seen that in almost all the datasets, the feature subset selected reduces at least 93% of the available features.

TABLE II
EXPERIMENTAL RESULTS OF PROPOSED ALGORITHM

Dataset	Genes selected			Classification accuracy	
	MI +ANN	BCOA +ANN	MI +BCOA +ANN	ANN	MI +BCOA +ANN
SRBCT	603	521	56	88.90	100
9_Tumors	146	1179	49	53.67	88.00
Brain_Tumor2	3997	998	458	68.80	94.63
11_Tumors	3002	4030	420	79.58	95.57
LungCancer	8159	1634	1906	89.86	98.01
14_Tumors	2023	4543	791	60.20	76.23

CONCLUSION

We present a filter based feature selection method based on BCOA and information theoryto select a smaller number of features of high dimensional datasets and achieve similar oreven better classification performance.This goal was achievedby developing a new feature selection algorithm based onBCOA and two information measures, namely entropy and mutual information. The results suggest that the proposed approach can significantly reduce the number offeatures whilst achieve similar or even better classificationaccuracy in almost all cases.

CONCLUSION

This work has been supported by research deputy of Islamic Azad University, Abadan branch, Abadan, Iran.

REFERENCES

[1] U.M. Fayyad, G. Shapiro and P. Smyth, “From data mining to knowledge discovery: an overview”, Advances in Knowledge Discovery and Data Mining,pp. 1–34, 1996.

[2] I. Guyon, A. Elisseeff, “An introduction to variable and feature selection”. Journal of Machine LearningResearch, pp.1157-1182, 2003.

[3] R. Kohavi, G. H. John, “Wrappers for feature subset selection”. Artificial Intelligence, pp. 273-324, 1997.

[4] P. Estevez, M. Tesmer, C. Perez, and J. Zurada, “Normalized mutual information feature selection,” IEEE Transactions on Neural Networks, vol. 20, no. 2, pp. 189–201, 2009.

[5] L. Yu and H. Liu, “Efficient feature selection via analysis of relevance and redundancy,” Journal of Machine LearningResearch, vol. 5, pp. 1205–1224, 2004.

[6] H. Yuan, S. S. Tseng, and W. Gangshan, “A two-phase feature selection method using both filter and wrapper,” in IEEEInternational Conference on Systems, Man, and Cybernetics(SMC’99), vol. 2, pp. 132–136, 1999.

[7] N. Kwak and C. H. Choi, “Input feature selection for classification problems,” IEEE Transactions on Neural Networks, vol. 13, no. 1, pp. 143–159, 2002.

[8] A. Unler and A. Murat, “A discrete particle swarm optimization method for feature selection in binary classification problems,” European Journal of Operational Research, vol. 206, no. 3, pp. 528–539, 2010.

[9] C. S. Yang, L. Y. Chuang, C. H. Ke, and C. H. Yang, “Boolean binary particle swarm optimization for feature selection,” in IEEE Congress on Evolutionary Computation (CEC’08), pp. 2093–2098, , 2008.

[10] K. Neshatian and M. Zhang, “Dimensionality reduction in face detection: A genetic programming approach,” in 24th InternationalConference Image and Vision Computing New Zealand(IVCNZ’09), pp. 391–396, 2009.

[11] R. Rajabioun, “Cuckoo Optimization Algorithm”, Applied Soft Computing journal, Vol. 11, pp. 5508-5518, 2011.

[12] J. Kennedy, R. C. Eberhart, and Y. Shi, Swarm Intelligence, ser. Evolutionary Computation Series. San Francisco: Morgan Kaufman, 2001.

[13] X.-S. Yang and S. Deb, “Cuckoo search via levy flights,” in Proceedings of the NaBIC 2009 - World Congress on Nature & Biologically Inspired Computing, pp. 210–214, 2009.

[14] C. Shannon and W. Weaver, The Mathematical Theory of Communication. Urbana: The University of Illinois Press, 1949.

★★★