Cultural Algorithm for Feature Selection

Abbas Keramati
Department of Industrial Engineering
University of Tehran
Tehran, Iran
Keramati@ut.ac.ir

Mohammad Darzi
Advanced Information Systems Research Group
ICT Research Institute - ACECR
Tehran, Iran
Modarzi@ictrc.ir

Mahdi Hosseini
Department of Industrial Engineering
University of Tehran
Tehran, Iran
Mahdi.Hosseini@ictrc.ir

Ali Asghar Liaei
Advanced Information Systems Research Group
ICT Research Institute - ACECR
Tehran, Iran
Liaei@ictrc.ir

***Abstract*—In this paper a new cultural evolution based feature selection method is proposed. The present process contains a wrapper approach based on Cultural Genetic Algorithm (CGA) and naïve Bayes classifiers. Cultural evolutionary algorithms are used for searching the problem space to find all of the possible subsets of features and naïve Bayes classifier is employed to evaluate each subset of features. According to its fast convergence, CGA is expected to show higher performance compared with classical GA. The results show that the proposed approach outperforms GA on different datasets.**

Keywords- Cultural Algorithm, Genetic Algorithm, Feature Selection, Naïve Bayes Classifier.

I. Introduction

Machine learning algorithms attempt to learn a model via training examples. Each training example is defined as a set of attribute-value pair. The accuracy of these algorithms depends on the number of attributes so that it degrades when face with many attributes which some of them may be irrelevant to the mining task or may be redundant [1]. Therefor these features increase the cost of retain and management of data and cause of confusing the task of classification. Generally, they lead to a low learning precision [2, 3, 4].

There are some approaches to cope with this problem that one of them is feature selection [5]. Feature selection task is to choose a subset of the original features present in a given dataset that provides most of the useful information [6]. Feature selection has many advantages such as: facilitating data, reducing the storage requirement, reducing the calculations, reducing training and testing time, defying the curse of dimensionality and etc. [7].

There are two main approaches in feature selection: filter and wrapper. Filter approach uses some techniques such as mutual information to measure the information content of selected subset ignoring classifier algorithm. In wrapper approach, the selected subset evaluates by a classifier algorithm. Each of mentioned approaches has some pros and cons. Filters have faster execution than wrappers because they don’t need to train a machine learning algorithm unlike wrappers. Since filters evaluate the intrinsic properties of data, rather than their interactions with a particular classifier, their results exhibit more generality. Wrappers have an ability to generalize, since they typically use cross-validation techniques to evaluate the accuracy. The advantage of wrappers is their good accuracy on classification unlike filters.

Wrappers require a search strategy to select candidate subset and an objective function to evaluate these candidates. Based on assumption that the number of attributes is n, the all possible combination of features will be 2n-1 that is too large when there is many attributes. So, a search strategy is needed to direct the feature selection process as it explores the space of all possible combination of features. There are different approaches for search strategy such as exponential, sequential and randomized. Randomized algorithms incorporate randomness into their search procedure to escape local minimum. As an example we can point to Genetic algorithm.

GA is a search algorithm that models the natural process biological evolution [8]. This algorithm tries to search on a problem space randomly to find the optimum one as a solution. In feature selection, the problem space is all possible combination of features that is called feature subset. The goal of GA is to find a subset of features that has the best predictive outcome.

In feature selection literature, GA is used in many studies as search strategy. For instance, Sun et al [9] used GA as the search strategy in feature selection for high-dimensional data clustering. Yang et al [10] presented a novel feature selection method that uses GA for global search and Taguchi methods for local search. Also, Seok Oh et al [11] used GA for global search in feature selection.

Cultural algorithms (CAs) are a class of evolutionary algorithms (EAs) that were developed based on the concept of cultural evolution. In this approach, knowledge gained about evolution process is gathered during the search. These information are maintained in a knowledge-base to be used later to lead the search direction in different iterations of the evolution. This approach to use of knowledge makes CA converge faster than other evolutionary algorithms.

In this paper a CA-based feature selection method is presented. The proposed method combines the capabilities of well-known Genetic algorithm (GA) with the advantages of cultural evolution. The resulting method was evaluated on the basis of two different databases. Also a number of experiments were designed to compare the performance of CA against genetic algorithm for feature selection.

The rest of this paper is organized as follows. Section 2 reviews the background of algorithms that is used in this study. These are cultural algorithm and naïve Bayes classifier. Section 3 presents the architecture of proposed model. In section 4, the results and experiments are explained. Finally, section 5 presents conclusion.

II. BACKGROUND

A. *Cultural Algorithm*

Cultural algorithms (CAs) developed by Reynolds [12-15] are a class of evolutionary algorithms (EAs) that were derived from the cultural evolution phenomenon [16]. There are two levels of evolution in a cultural algorithm. These two levels are micro-evolution and macro-evolution. Micro-evolutionary or population level often is formed of a population based randomized search algorithm like genetic algorithm, genetic programming, particle swarm optimization, etc. the evolution process in the population level is mainly based on the micro level interactions between individuals. Such interactions rely originally on diversification of feasible solution candidates [13,17]. In macro-evolutionary or cultural level, experiences of the individuals extracted from population level can be saved and used to impact the direction of the search process in a more extensive manner [18,19].

The dual inheritance mechanism contained in cultural algorithms benefits the optimization process in different ways. The cultural change can be used to speed up the convergence period of the search. Moreover it can be used to control the ability of the search method in exploration and/or exploitation of better solutions by the use of different conceivable knowledge sources.

The basic framework of cultural algorithms is shown in figure 1[19,20]. As mentioned before, the main components of CA are a population space and a belief space. A set of communication protocols control the interactions between two spaces. These protocols determine which members of the population are acceptable to adjust the knowledge content of the belief space, and how this content should impact populations in different generations.

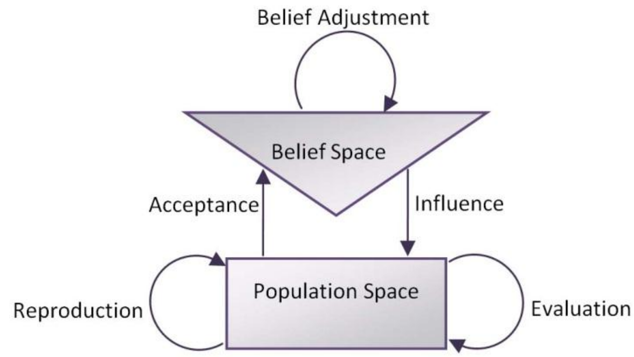


Figure 1. Cultural Algorithms Framework

Five different types of knowledge typically are used to implement the cultural evolution process. They are normative knowledge, topographic knowledge, situational knowledge, historical knowledge, and domain knowledge [21-24].

Situational knowledge stores a number of quality individuals, often the best individual during previous generations, in the belief-space to be used later to influence the next generation. Use of situational knowledge, makes algorithm converges faster.

Historical (or temporal) knowledge consists of the last k important significant changes in the search direction or meaningful movements on the problem landscape. In addition to its capability to prevent the search process from stopping in local optimum, historical knowledge is a good schema to reinforce the exploration of cultural algorithms.

Normative knowledge can be defined as the intervals for the attribute variables of candidates where good solutions have been found. Population members then are led to these intervals. Thus, it is expected to have better solutions.

Topographic knowledge Consists of different regions of the problem landscape, the best solution found in each region during the search process, and probably a list of k best regions found.

Domain knowledge points to concepts, rules and principles that shape the domain of the problem.

B. *Naïve Bayes Classifier*

Naïve Bayes classifier provides the most probable class given input data. Assume a case $x=[x_1, x_2, \dots, x_n]$ *B*elongs to one of C classes. To guarantee minimum classification error, we select the class with maximum posterior probability. The posterior probability of each class can be calculated using Bayes formula [25] as shown in equation (1).

$$P(\omega_j|x) = \frac{P(\omega_j)p(x|\omega_j)}{\sum_{i=1}^{c} P(\omega_i)p(x|\omega_i)}, \quad j = 1, \dots, C \tag{1}$$

In the above equation, $P(\omega_i)$ is the prior probability of class ω_i , and $p(x|\omega_i)$ is the class conditional probability density function (pdf). Calculating pdf is not so simple, but by assuming that the features are independent, pdf is calculated as below:

$$P(x|\omega_j) = \prod_{i=1}^{n} p(x_i|\omega_j)\ , \quad j = 1, \dots, C \qquad (2)$$

So, by replacing the formula 2 into formula 1, the calculations will be easier. The most value of $P(x|\omega_j)$ for $j=1,\dots,C$ determines that the input data **x** belongs to class j.

III. ARCHITECTURE

The goal of proposed model is to select the best feature subset that has the best results on classification task. Hence, a feature selection model is presented in which cultural genetic algorithm is used as search strategy and naïve Bayes classifier is used as objective function. The pseudo code of the proposed model is presented in figure 2.

The process of proposed model is as below:

1) At first, the initial population is generated based on acceptable points in problem space. Also initial knowledge base is generated. Before starting the evolution process, the initial population is evaluated by fitness function.
2) In each iteration t, the next generation is generated by doing some operation on present population.
3) The generated population is evaluated by fitness function.
4) The belief space is updated based on fitness values of present population.
5) The belief space influences the present population
6) The present population is evaluated by fitness function.
7) Until the stopping criteria is not met, steps 2 to 6 is repeated.

A number of researchers in the field of artificial intelligence have worked on the combination of evolutionary algorithms with cultural approach. In most of these models, a search procedure were used in-which fitness evaluation step takes place only one time in each iteration of the evolution. So the source of knowledge used to lead the search direction would be the information gained through previous iteration. [20]. However this knowledge can be updated by the use of information gathered through the micro-evolutionary reproductions of present iteration.

In fact, randomized interactions between population members in the micro-evolution level may lead to reproduction of new and better candidates. Hence it may improve the capabilities of the algorithm to use the knowledge of present iteration instead of the information gathered at the end of previous iteration. For this purpose, in each generation the information gained through the micro-evolution, should be used to update the knowledge-base that would be the source of cultural variations at the current iteration. In such an approach each individual is needed to be evaluated two times in each generation. Although the processing time will be raised by this change, the pace of convergence is expected to be raised. It would be a meaningful advantage for the proposed model in comparison with other cultural-evolutionary algorithms.

```
Begin
    t=0;
    Initialize Population POP(t);
    Initialize Belief Space BLF(t);
    Evaluate Population POP(t);
    Repeat
        t=t+1;
        Variation(POP(t) ←POP(t-1));
        Evaluate Population POP(t);
        Adjust (BLF(t), Accept(POP(t)));
        Influence (POP(t), BLF(t));
        Evaluate Population POP(t);
    Until (termination condition achieved)
End
```

Figure 2. Cultural Genetic Algorithm Pseudo Code

A. Chromosome Representation

In proposed model each chromosome is a subset of features. So, the chromosome size (number of genes) is equal to number of features. Each chromosome is represented in binary format in which 1 means corresponding feature is selected and vice versa. As shown in figure 3, the number of genes is equal to number of features that and they are equal to n.

B. Population

A population is a set of chromosome. Also a population is called a generation during the evolutionary process. In figure 3 the population size is m. In evolutionary process the initial population is generated randomly and the next generations are obtained through some operations.

C. Fitness Function

For scoring the chromosomes and sort them by their score, all of them are evaluated by fitness function to be determined how fit are them. In feature selection, fitness function is the objective function. This function determines that which chromosome has the best outcome on classification task. In proposed model naïve Bayes classifier is performed as fitness function. For each chromosome and based on the selected features, naïve Bayes classifier trains on train set and tests on test set. The classification error of each chromosome is the scores of it.

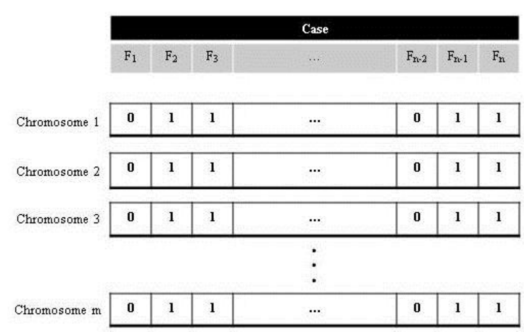


Figure 3. Chromosome Representation

D. Knowledge

In proposed model, situational knowledge is used in cultural genetic algorithm. The best chromosome of each generation influences the current population as knowledge. After reproduction of the next generation, best chromosome of that population is saved as knowledge. So, the belief space is updated.

E. Operators

For reproducing new generation in order to maximize the fitness scores, some operators such as crossover and mutation are used. At first, some of elite chromosomes with the highest fitness score are selected to go to the next generation directly. Afterward, some of remained chromosomes are selected to be changed by crossover. Crossover exchanges substring from pairs of chromosome to generate two new chromosomes. In this study, two-point crossover is employed. Then mutation is applied on the remained chromosomes. In mutation, selected genes are inverted. Mutation prevents the search process from falling into local maxima [26]. After applying the genetic operators, the cultural operators such as knowledge-based crossover and mutation are applied in the current generation. The mutation that is used in cultural step is the same as genetic step. But the crossover is a bit different. In knowledge-based crossover, a copy of the best chromosome (situational knowledge) is used to produce two new offspring. So, after applying the genetic and cultural operation on a generation, the new generation is produced.

IV. RESULTS

As mentioned above, the goal of this study is to check the strength of CAs in feature selection problem. Also, the cultural approach can be considered as a representative of macro-evolution search. In order to achieve the mentioned target, some experiments are designed. According to the use of genetic algorithm in many studies on feature selection, GA is used as base of comparison. So, by designing different experiments through changing the parameters such as population size, the pattern of this algorithm can be obtained. For evaluating the algorithms Holdout Cross-validation is used and for classification naïve Bayes classifier is performed.

Hence, the performance of naïve Bayes classifier is evaluated in below situations:

- Without preprocessing
- With feature selection using genetic algorithm
- With feature selection using CGA

In designed experiments, classification error is evaluated. Classification error is the error rate of naïve Bayes classifier outcome on test set where the inputs are the selected features value according to the cultural genetic algorithm.

TABLE 1. SPECIFICATION OF USED DATASET

Dataset	#Attribute	#Class	#Samples
Spam	57	2	4601

TABLE 2. COMPARISON ACCURACY RATES BETWEEN PROPOSED MODEL AND OTHER MODELS

Method	Error rate (%) in different experiments 1	2	3	4	5	6
Naïve Bayes classifier	13					
CGA+ Naïve Bayes classifier	(A) 6.1	(B) 6.4	(C) 6.5	(D) 6.6	(E) 6.6	(F) 6.7
GA+ Naïve Bayes classifier	(G) 6.3	(H) 7.2	(I) 7.3	(J) 10.8	(K) 10.8	(L) 11.2

For evaluating the proposed model, spam dataset taken from [27] is used. The specification of this dataset is shown in table 1. This dataset is presented for classification. For this purpose, each of GA and CGA were experimented 6 times in the same conditions such as population size, operators, number of generations, and etc.

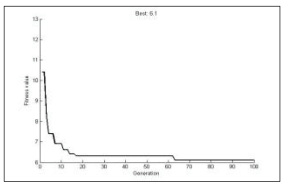
The results (table 2) demonstrate that CGA outperforms GA in five experiments out of six. Also, the best performance of CGA is better than the best performance of GA. It is clear from the results that the naïve Bayes classifier ignoring feature selection is not competitive with two other approaches. For better understanding, the results of all six experiments for both of GA and CGA are depicted in figure 4. The graphs of CGA (left side) and GA (right side) are sorted based on error rate in descending order. As shown in this figure, CGA converges to the optimum solution (optimum error rate) faster than GA. In other words CGA approaches to the optimum in fewer generations.

V. CONCLUSION

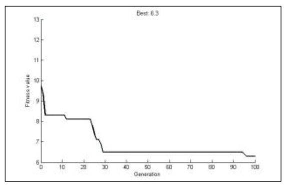
In this paper a new approach based on cultural evolution was proposed for feature selection problem. The goal of feature selection methods is to find the optimum subset of features that result in the best outcome for classification task. For searching on the space of all possible subsets of features, the cultural genetic algorithm was employed. Situational information gathered through the search process was used as the knowledge source of cultural algorithm. To show the effect of adding some knowledge source on evolutionary search process a number of experiments were designed and the performance of CGA was compared with a GA optimization technique. The results illustrate that the use of macro-evolutionary knowledge improves the convergence of the algorithm and reduces the error rate of the resulting classifier.

The experimental results show that although GA is known as a good method for exploring the problem space, it suffers from the weakness in exploitation. The capability of CA in storing and reusing knowledge can be used to improve the exploitation ability of the search algorithm. So, we can expect that the combination of GA and CA leads to a good performance in both exploration and exploitation task. While using only one of the knowledge types presented in the literature for CAs, results show significant difference between CGA and GA. The recent assertion is clear in figure 5 where the two algorithms can be compared based on mean error value over 6 experiments in each generation.

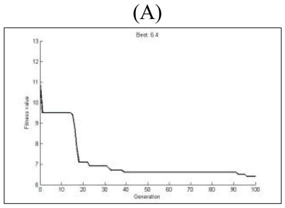
In future work we intend to analyze the proposed algorithm in the presence of other knowledge types including temporal, normative, and topographic knowledge. In addition, to prove the efficiency of cultural algorithm for feature selection, it should be compared with various algorithms based on different datasets.



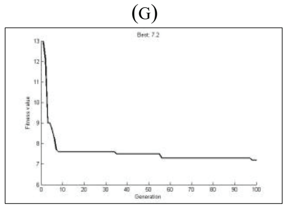
(A)



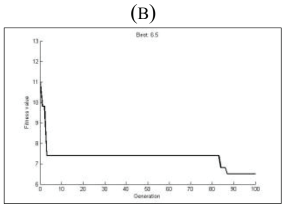
(G)



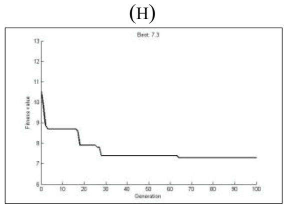
(B)



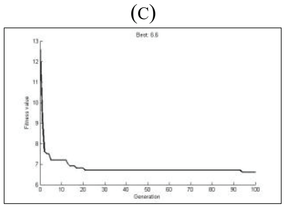
(H)



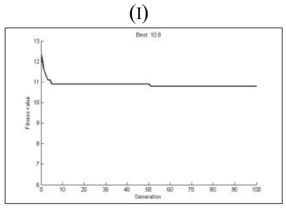
(C)



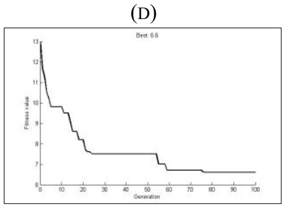
(I)



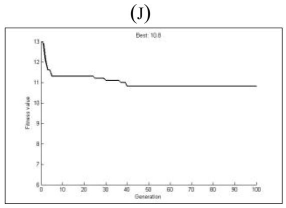
(D)



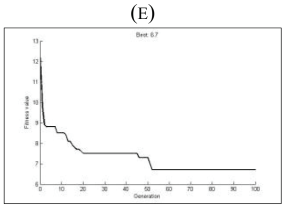
(J)



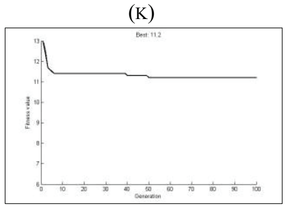
(E)



(K)



(F)



(L)

Figure 4. Error variation during 100 generations on different experiments of cultural genetic algorithm (A-F) and genetic algorithm (G-L)

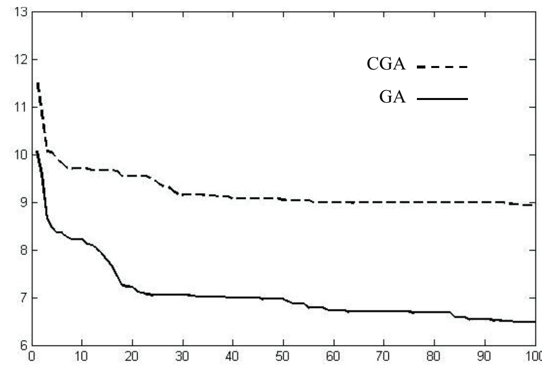


Figure 5. Mean error value (%) over 6 experiments and 100 generations for GA and CGA

ACKNOWLEDGMENT

This study is a part of a multidisciplinary research project conducted jointly by the faculty of industrial engineering in university of Tehran and Intelligent Systems research group in ICT Research Institute – ACECR. We would like to express our warm appreciation for Mr. Habibolah Asghari who helped us through compilation of this paper.

REFERENCES

[1] J. Han, and M. Kamber, “Data mining: concepts and techniques,” 2006: Morgan Kaufmann.
[2] R. Kohavi, and G.H. John, “Wrappers for feature subset selection,” Artificial intelligence, 1997. 97(1-2): p. 273-324.
[3] Y. Yuling, “A Feature Selection Method for Online Hybrid Data Based on Fuzzy-rough Techniques,” 2009: IEEE.
[4] N. Abe, et al., “A divergence criterion for classifier-independent feature selection,” Advances in Pattern Recognition, 2000: p. 668-676.
[5] M. Dash, and H. Liu, “Feature selection for classification,” Intelligent data analysis, 1997. 1(3): p. 131-156.
[6] R. Jensen, and Q. Shen, “Computational intelligence and feature selection: rough and fuzzy approaches,” IEEE Press Series On Computational Intelligence, 2008: p. 340.
[7] Guyon, and A. Elisseeff, “An introduction to variable and feature selection,” The Journal of Machine Learning Research, 2003. 3: p. 1157-1182.
[8] M., Mitchell, “An introduction to genetic algorithms,” 1998: The MIT press.
[9] M. Sun, et al. “A GA-Based Feature Selection for High-Dimensional Data Clustering,” 2009: IEEE.
[10] C.H. Yang, et al., “A Novel GA-Taguchi-Based Feature Selection Method,” Intelligent Data Engineering and Automated Learning–IDEAL 2008, 2008: p. 112-119.
[11] I.S. Oh, J.S. Lee, and B.R. Moon, “Hybrid genetic algorithms for feature selection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004: p. 1424-1437.
[12] Reynolds, R. and W. Sverdlik. Learning the behavior of Boolean circuits from examples using cultural algorithms [1962-18]. 1993: SPIE INTERNATIONAL SOCIETY FOR OPTICAL.
[13] Reynolds, R.G. An introduction to cultural algorithms. 1994: World Scientific.
[14] Reynolds, R.G. and W. Sverdlik. Solving Problems in Hierarchically Structured Systems Using Cultural Algorithms. 1993.
[15] Reynolds, R.G. and W. Sverdlik. Problem solving using cultural algorithms. 1994: IEEE.
[16] Franklin, B. and M. Bergerman. Cultural algorithms: Concepts and experiments. 2000: IEEE.

[17] Kureichik, V., V. Kureichik, and P. Sorokoletov, Analysis and a survey of evolutionary models. Journal of Computer and Systems Sciences International, 2007. 46(5): p. 779-791.
[18] Gao, F., G. Cui, and H. Liu. Integration of genetic algorithm and cultural algorithms for constrained optimization. 2006: Springer.
[19] Reynolds, R.G. Cultural algorithms: Theory and applications. 1999: McGraw-Hill Ltd., UK.
[20] El-Hosseini, M.A., et al. Cultural-Based Genetic Algorithm: Design and Real World Applications. 2008: IEEE.
[21] Iacoban, R., R.G. Reynolds, and J. Brewster. Cultural swarms: modeling the impact of culture on social interaction and problem solving. 2003: IEEE.
[22] Reynolds, R.G. and B. Peng, Cultural algorithms: computational modeling of how cultures learn to solve problems: an engineering example. Cybernetics and Systems: An International Journal, 2005. 36(8): p. 753-771.
[23] Reynolds, R.G., B. Peng, and J. Brewster. Cultural swarms: Knowledge-driven problem solving in social systems. 2003: IEEE.
[24] Reynolds, R.G., B. Peng, and J. Brewster. Cultural swarms II: Virtual algorithm emergence. 2003: IEEE.
[25] Hoare, Z., Landscapes of naïve bayes classifiers. Pattern Analysis & Applications, 2008. 11(1): p. 59-72.
[26] H. Ahn, and K. Kim, “Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach,” Applied Soft Computing, 2009. 9(2): p. 599-607.
[27] Blake, C. L. and Merz, C. J.: UCI Repository of Machine Learning Databases. [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, University of California. Department of Information and Computer Science.

.