A New Feature Selection Method Based on Ant Colony and Genetic Algorithm on Persian Font Recognition

Maryam Bahojb Imani, Tahereh Pourhabibi, Mohammad Reza Keyvanpour, and Reza Azmi

***Abstract*—Dimensionality reduction of a feature set is a usual pre-processing step used for classification applications to improve their accuracy with a small and appropriate feature subset. In this article, a new hybrid of ant colony optimization (ACO) and genetic algorithm (GAs) as a feature selector, and support vector machine as a classifier are integrated effectively. Based on the combination of the fast global search ability of GA and the positive feedback mechanism of ACO, a novel algorithm was proposed in the domain of feature selection. Experiments show that the proposed feature selection can achieve better performance than that the normal GA and ACO does. We tested this method on the extracted features of ten common Persian fonts. The result shows it has affected slightly better on the performance. Furthermore, number of features decreased to almost half of the original feature number after this pre-processing step.**

***Index Terms*—Feature subset selection, ant colony optimization, genetic algorithm, persian font recognition.**

I. Introduction

Pattern classification is the task of classifying any given input feature vector into pre-defined set of classes. Irrelevant and redundant features increase the size of search space which consequently led to increase the time of classification and make generalization more difficult. In addition, the more features increase the risk of over fitting [1]. This large dimensionality (each feature is a separate dimension) is a major problem in machine learning and data mining.

The selection of features which are used for classification could have an impact on the accuracy of the classification and the computational time, training data set requirements, and implementation costs associated with the classification. Selecting the optimal feature subset is proven to be an NP-complete problem [2]. Three approaches of feature subset selection are filter, wrapper and hybrid approaches. One of the sub-categories of wrapper approach is "random search" which include meta-heuristic methods such as genetic algorithm, simulated annealing, ant colony, and hybrid of them [3].

Manuscript received April 15, 2012, revised May 12, 2012. This work was supported by Education & Research Institute for ICT (ERICT) under grants number 8971/500.

Maryam Bahojb Imani is with Computer and Information Technology Engineering Department, Amirkabir University of Technology, Tehran, Iran (e-mail: imani.m@aut.ac.ir).

Tahereh Pourhabibi is with the Computer Engineering Department, Alzahra University, Tehran, Iran (e-mail: tahereh.pourhabibi@student.alzahra.ac.ir).

Mohammad reza Keyvanpour is with Computer Engineering Department, Alzahra University, Tehran, Iran (e-mail: Keyvanpour@alzahra.ac.ir).

R. Azmi is with Computer Engineering Department, Alzahra university, Tehran, Iran (e-mail: r.azmi@alzahra.ac.ir).

Ant colony optimization [4] is used to solve discrete combinatorial optimization problems [5]. ACO has been used as a search procedure for selecting features [6] [7].

Genetic algorithms are search methods that take their inspiration from natural selection. One of the applications of GA optimization is in feature selection.

In this paper, we present a novel hybrid algorithm according to the ACO and GA for selection of optimal feature subsets. Proposed algorithm is applied to the extracted features of Persian common font dataset. The classifier which is used to classify these dataset and to compute feature selection performance is SVM with RBF kernel.

The rest of the paper is organized as follows: a brief overview of related researches in Persian font recognition and feature selection approaches is in Section II. Selection III presents an overview of ant colony optimization and its application in feature selection. Genetic algorithm is addressed in section IV. Section V describes the proposed hybrid algorithm of GA and ACO in feature selection. Feature extraction method for font recognition and experimental result are reported in section VI and VII, followed by conclusions in Section VIII.

II. Related Research

In following sub-section, we'll talk about related researches in font recognition, especially in Persian language. After that, some feature selection approaches will be presented.

A. Font Recognition

Despite of the obvious importance of automatic font recognition, inappreciable font recognition researches in English, Spanish, Korean and Japanese have been done [8-10]. As well as, in Persian font recognition fewer investigations have been done.

In [11], 24 Gabor filters in four scales and six directions were used for classification of 20 common Persian font types. In [12], 10 font types of 7 different sizes and 4 different types were used and feature extraction technique was based on the first and second order moment. In [13], global typographical features based on Gabor filters are used to extract the features. In another recent research, features obtained from Sobel and Roberts matrix. This method attains better result than other methods mentioned in Persian fonts recognition [14].

B. Feature Selection Approaches

The best subset contains the least number of features that most contribute to accuracy. The whole search space for optimization contains all possible subsets of features, meaning that its size is: 2^n, where n is the dimensionality.

Therefore, this search for an optimal set of attributes would be time-consuming and computationally expensive if n is large. Three categories of optimal feature subset selection are as follows [1]:

Filter approaches: In this approach undesirable attributes are filtered out of the data before classification begins.

Wrapper approach: In this approach, feature selection is "wrapped" in a learning algorithm. It generates various subsets of features, and then a specific classification is applied to evaluate accuracy of these subsets. Wrapper methods can broadly be classified into two categories: greedy methods and randomized methods which contain ant colony optimization, genetic algorithm [15], and particle swarm optimization (PSO) [16].

Embedded approach: The idea behind the hybrid method is that filter methods are applied to select a feature pool at first and then the wrapper method is used to find the optimal subset of features from the selected feature pool.

Development of an accurate and fast search algorithm for the selection of optimal feature subset is an open issue.

III. Ant Colony Optimization

Ant colony is a part of swarm intelligence approach that has been successfully used in the general purpose optimization technique. ACO which is a randomized search method based on the foraging behavior of some ant species [17]. These ants deposit same amount of pheromone on their paths in order to mark some favorable path that should be followed by other members of the colony, so shorter paths will receive more pheromone per time unit. In addition, the amount of pheromone on each path decreases as time passes because of evaporation. Therefore, longer paths lose their pheromone intensity and become less favorable over time.

ACO has been successfully applied in the selection of optimal minimal subset of feature. ACO algorithms can be used in any optimization problem, if the following aspects are provided [18]:

Graph Construction: The problem needs to be represented in a shape of graph, i.e. nodes represent features, and the edges between them denote the choice of the next feature. Nodes are fully connected. The search for the optimal feature subset is an ant traversal through the graph in which the minimum number of nodes is visited [19].

Solution construction: Each ant starts randomly chosen start feature and in each iteration adds an unvisited feature to its partial tour. At each step, each ant constructs a solution. Then each ant will share its solution feedback with the entire colony by updating a global data structure called the pheromone matrix. This data structure simulates the pheromone trails. Each entry in the pheromone matrix shows the desirability of each solution component. At the end of each iteration, the pheromone associated with each solution component is reinforced based on the quality of the solution that comprises the particular solution component. In subsequent iterations, ants will use the pheromone intensities of available solution components to guide solution construction. As a result of repeated pheromone reinforcements, a subset of solution components will emerge to have pheromone intensities much higher than the others. At this point, ants will construct identical or nearly identical solutions using these highly desirable components. The solution construction terminates when it can satisfy the stopping criterion (e.g. suitably high classification accuracy has been achieved) [20].

Pheromone trails and heuristic information: Let S be a set of given N features, and $\tau(f)$ be the pheromone level (desirability measure) of feature f to be in the selected subset of features s where $s \subset S$. Initially, the desirability of each feature will be the same, but the desirability of those features which are more important, increases in each step. [21].

Here we used F-score as heuristic information. Eq.(1) defines the F-score of the i th feature.

$$\eta_i = \frac{\sum_{c=1}^{v} (\bar{x}^{(c)} - \bar{x}_i)^2}{\sum_{c=1}^{v} \{\frac{1}{N^{(c)} - 1} \sum_{j=1}^{N_i^{(c)}} (x_{i,j}^{(c)} - \bar{x}_i^{(c)})^2\}} \quad (1)$$

where $i \in \{1,2,..N_F\}$

A larger F-score corresponds to a greater likelihood that this feature is discriminative where v is the number of categories of target variable; N_F the number of features; $N_i^{(c)}$ the number of samples of the i th feature with categorical value c, $c \in \{1,2,3,...,v\}$, $i \in \{1,2,..N_f\}$; $x_{i,j}^{(c)}$ the j th training sample for the i th feature with categorical value c , $j \in \{1,2,3,...,N_i^{(c)}\}$; $\bar{x}_i$ the mean of the i th feature; $x_i^{(c)}$ the mean of the i th feature with categorical value c [22].

The goal of this optimization algorithm is to minimize the classification error in predicting the output. In this hybrid approach, the role of each ant is to build a solution subset. The ''ants'' build solutions applying a probabilistic decision policy to choose next node. In this case each subset of feature represents a state. The state transition rules help ants to select features using the pheromone trail and the heuristic value. Each ant chooses a particular feature by maximizing a product of these two parameters [21].

An ant chooses a feature as follows:

$$P_i^k(t) = \begin{cases} \frac{[\tau_i(t)]^{\alpha}.[\eta_i]^{\beta}}{\sum_{u \in J^k} [\tau_u(t)]^{\alpha}.[\eta_u]^{\beta}} & if \quad i \in J^k \\ 0 & otherwise \end{cases} \quad (2)$$

where J^k is the set of feasible features that can be added to the partial solution; $\alpha \geq 0$ and $\beta \geq 0$ are two parameters that determine the importance of the pheromone value and heuristic desirability.

Pheromone update rule: In AS_{rank}, which is used in this research, each ant deposits an amount of pheromone that decreases with its rank. Additionally, the best-so-far ant ($\Delta\tau_i^{bs}$) always deposits the largest amount of pheromone in each iteration.

The best-so-far tour gives the strongest feedback, with weight w (i.e. its contribution $\Delta\tau_i^{bs}$ is multiplied by w); the r -th best ant of the current iteration contributes to

pheromone updating with the value $\Delta\tau_i^r$ multiplied by a weight given by $\max\{0, w-r\}$ [18] and finally ρ is an amount of evaporation. Thus, the AS_{rank} pheromone update rule is:

$$\tau_{ij} \leftarrow \rho * \tau_{ij} + \sum_{r=1}^{w-1}(w-r)\Delta\tau_{ij}^r + w\Delta\tau_{ij}^{bs} \quad (3)$$

where

$$\Delta\tau_{ij}(t) = \sum_{k=1}^{n}(\gamma' \frac{(S^k)}{|S^k|}) \quad (4)$$

The pheromone is updated according to both the measure of the goodness of the ant's feature subset γ' and the size of the subset itself [23].

IV. Genetic Algorithm

GA Techniques based on the very fundamental principles of evolution are known as evolutionary algorithms (EAs). Their search starts from population of solutions (vectors). Here, if a particular attribute is selected in a vector, a '1' is assigned to it. Conversely, for an attribute that is not included in selected subset, a '0' is assigned to it. The vectors of initial population are filled randomly. After that, the fitness value of each vector has to be computed according to appropriate fitness function that reflects its worth in relation to the desired goal. Then the iteration of reproducing the new population will begin and a typical series of operations carried out in these iterations such as parent selection, crossover and mutation. At each step, some condition such as the number of iteration and the performance of the classifier is checked and if none of them is met, the procedure reproduce new population.

V. Proposed Feature Selection Algorithm

The pseudo code of proposed algorithm is presented in Fig. 1.

The best solution may be generated by either ACO or GA. Therefore, ACO can utilize the GA's cross-over and mutation operations. This leads to the exploration of ants around the optimal solution in next iterations. After updating pheromone, the process iterates once more.

Another improvement is to integrate the selection, crossover and mutation operators into the ACO, which leads it converges more quickly.

Furthermore, in this algorithm the better results can be kept for use of GA in next iteration, and the GA operators can be used for them. It helps to have a better population comprising the random chromosomes and k-better results of previous iteration. Furthermore, the GA is much faster than ACO. So that new algorithm can utilize of its speed.

VI. Feature Extraction

For experimental studies we have used ten common Persian fonts database [14]. This database contains 5000 font images from 10 different fonts. So, the number of images for each font in our experiments is 500. Fig. 2 shows some samples images of this database.

Feature extraction technique based on wavelet was used. Applying a gridding approach we divide each texture of size 128*128 into 16 sub blocks of size 32*32 and combined wavelet energy and wavelet packet energy features of each sub block to obtain a feature vector.

In the 2-D case, the wavelet transform is usually performed by applying a separable filter bank to the image. The wavelet decomposition of a 2-D image can be obtained by performing the filtering consecutively along horizontal and vertical directions.

```
Input: Dataset
Input for ACO: number of ants and ACO's parameters (α,β,p and r)
Input for GA: number of chromosomes, total iteration, crossover and mutation rate
Output: selected subset features

1: Set the same value for pheromone and heuristic information (Eq. 1)
2:       % Chromosome generation
3: Initialization Chromosome population randomly
4: Do (for each iteration)
5: {
6:     Evaluate population's fitness (SVM classifier accuracy)
7:     Parent selection with roulette wheel policy
8:     Crossover and Mutation ( if their conditions occurs)
9:     Delete extra Chromosomes (according classification error)
10: }
11:      % Ant generation
12: A feature randomly assign to each ant and evaluate it
13: Do (for each Ant)
14: {
15:    Calculate State Rule (Eq. 2)
16:    Choose next feature (according the step 15.)
17:    Evaluate Ant (SVM classifier accuracy)
18:    If three successive steps don't improve accuracy
19:          Continue;
20: }
21: AcoGa ← Merge the Chromosomes and Ants
22: Sort AcoGa according classification accuracy
23: Check the termination condition (iteration number or accuracy)
24:        Return best subset
25: Best_AcoGa ← Select the r-best of AcoGa
26: Use Best_AcoGa in next population of GA beside random population
27: Update the pheromone of all ants (Eq. 3)
28: Go to step 2 and continue.
```

Fig. 1. Proposed algorithm in feature selection algorithm base on ACO and GA

The wavelet energy is the sum of square of detailed wavelet transform coefficients. For an Image of size $N \times N$ the related wavelet energy can be calculated in horizontal, vertical and diagonals direction at $i-$level, respectively as below:

$$E_i^h = \sum_{x=1}^{N}\sum_{y=1}^{N}(H_i(x,y))^2$$
$$E_i^v = \sum_{x=1}^{N}\sum_{y=1}^{N}(V_i(x,y))^2 \quad (5)$$
$$E_i^d = \sum_{x=1}^{N}\sum_{y=1}^{N}(D_i(x,y))^2$$

$(E_i^h, E_i^v, E_i^d)_{i=1,2,...,K}$ is wavelet energy feature vector. Where K is the total wavelet decomposition level and H_i, V_i, D_i are the wavelet coefficient in horizontal, vertical and diagonals direction, respectively.

Wavelet packet transform recursively decomposes the

high-frequency components, thus constructing a tree-structured multiband extension of the wavelet transform. An image of a square local area is decomposed and related wavelet packet coefficients are extracted, then the following averaged energy is calculated:

$$E = \frac{1}{N * N} \sum_{i=1}^{N} \sum_{j=1}^{N} [s(i, j)]^2 \quad (6)$$

Fig. 2. Some samples images of Persian font database

where $s(i, j)$ denotes the wavelet coefficients of a feature sub image in the N×N window centered at pixel (i, j) [24]. Here, the length of this feature vector is 368.

VII. EXPERIMENTAL RESULT

In this step, we have applied proposed algorithm to the extracted feature of previous step. In this algorithm various parameters for leading to better result are tested and the best parameters are as follows:

α=1, β=0.2, ρ=0.2, the initial pheromone intensity of each feature is equal to 1, the number of Ant in every iteration m=10, the maximum number of iterations is100, the number of ant that is selected for AS_{rank} r=10, and parameters for GA are number of population equal to 50, crossover rate is 0.5, mutation rate is 0.02, and the number of iteration of the GA is 100.

This program is run in Matlab 7.6.0 on PC with core 2 duo 2.5 GHz CPU and 3 GB of RAM.

During feature selection, SVM classifier with RBF kernel is used as a feedback of the classifier in selected features.

The results of this step are summarized in a figure 2 and Table I.

In Table I, the classification accuracy without feature selection is an average of 5-fold cross validation method. In addition, it's clear that our proposed feature selection is more efficient than ACO and GA. In this comparison GA and ACO were run separately.

In Table II, we compare the existing methods in Persian font recognition with our feature extraction and feature selection approach. A comparisons show that after our feature selection, Persian classification has higher accuracy and the number of features become almost one second of original feature length. Therefore it affects the total time for feature extraction, which is another contribution of this work.

TABLE I: RESULTS OF THE EXPERIMENTS

Methods	Number of Features	Accuracy (%)	Feature slection Time (h)
Classification without FS	368	95.8%	-
Classification with GA	171	89.7%	24.31
Classification with ACO	256	93.1%	26.13
Classification with Proposed algorithm	189	96.3%	13.93

TABLE II: COMPARISON BETWEEN DIFFERENT PERSIAN FONT RECOGNITION METHODS

Feature type	Feature length	Recognition rate (%)
Gabor [11]	256	80.7
Sobel-Robert [14]	512	94.1
Wavelet energy	368	95.6
Wavelet Energy (With FS)	189	96.3%

VIII. CONCLUSION

We have presented a novel algorithm that combines the strengths of two well known algorithms to select the optimal feature subsets from features of Persian font images. This algorithm is a hybrid of a number of ACO and GA, and SVM as a classifier. ACO and GA help each other to improve the speed of their convergence, i.e. ACO uses the better result of GA to update its pheromone and GA applies the better result of ACO beside its initial random population. We have used this algorithm on wavelet packet energy as extracted features of common Persian font recognition. In this application a proposed feature selection could improve speed and classifier accuracy. Furthermore, the number of features becomes nearly half of the original feature number.

REFERENCES

[1] I. A. Gheyas and L. S. Smith, "Feature subset selection in large dimensionality domains", *Pattern Recognition*, Vol. 43, 2010, p. 5-13.
[2] A. A. Albrecht, "Stochastic local search for the feature set problem, with applications to microarray data", *Applied Mathematics and Computation*, Vol.183, 2006, p. 1148–1164.
[3] J. Doak, "Intrusion detection: The application of feature selection a comparison of algorithms, and the application of a wide area network analyzer", M.S. thesis, Department of Computer Science, University of California: Davis, 1992.
[4] M. Dorigo, V. Maniezzo, and A. Colorni, "The ant system: optimization by a colony of cooperating agents", *IEEE Transactions on System, Man, and Cybernetics*, Vol.26 (No.1), 1996, p. 1–13.
[5] M. Dorigo, G.D. Caro, and L.M. Gambardella, "Ant algorithms for discrete optimization Artificial Life", Vol.5 (No.2), p.137-172, 1999.
[6] A. Ani, "Feature subset selection using ant colony optimization", *International Journal of Computational Intelligence*, Vol.2 (No.1), 2005, p. 53-58.
[7] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics. Bioinformatics", Oxford University Press, 2007.
[8] Y. Zhu, T. Tan, and Y. WANG, "Font Recognition Based On Global Texture Analysis", in *IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp. 1192-1200, 2001.

[9] C. B. Jeong, "Identification of font styles and typefaces in printed Korean documents", Lecture Notes in Computer Science, (2003), p. 666-669.
[10] A. Zramdini, "Study of optical font recognition based on global typographical features", PhD thesis, University of Fri- bourg, 1995.
[11] H. Nezam, S. Nezam Abadipour, and V. Saryazdi and Ebrahimi, "Font recognition based on Gabor filters" (in Farsi), in *9th Iranian Computer Conference*, pp. 371-378, 2003.
[12] E. Rashedi, H. Nezamabadi-pour, and S. Saryzadi, "Farsi font recognition using correlation coefficients" (in Farsi), in *4th Conf. on Machine Vision and Image Processing*, (2007): Iran.
[13] A. Borji and M. Hamidi, "Support vector machine for Persian font recognition", *International Journal of Computer Systems Science and Engineering*, Vol.2 (3), 2007.
[14] H. Khosravi and E. Kabir, "Farsi font recognition based on Sobel-Roberts features", *Pattern Recognition Letters*, Vol.31, p. 75-82, 2010
[15] J. Yang and V. Honavar, "Feature subset selection using a enetic algorithm", *IEEE Intelligent Systems and their Applications*, Vol.13, p. 44-49, 1998.
[16] X. Wang, et al., "Feature selection based on rough sets and particle swarm optimization", *Pattern Recognition Letters*, Vol.28, pp. 459-471, 2007.
[17] M. Dorigo, "Optimization, learning and natural algorithms", in Dipartimento di Elettronica, Ph.D. dissertation, Politecnico di Milano: Italy, 1992
[18] M. Dorigo and T. Sttzle, "Ant colony optimization", MIT press, 2004.
[19] L. Wen, Q.Yin, and P. Guo, "Ant colony optimization algorithm for feature selection and classification of multispectral remote sensing image", in *International Geoscience and Remote Sensing*, pp. 923-926, 2008.
[20] C. K. Ho and H.T. Ewe, "GACO- a hybrid ant colony optimization meta heuristic for the dynamic load-balanced clustering problem in ad hoc networks", *Applied Artificial Intelligence*, Vol.23, pp. 570-598, 2009.
[21] R. K. Sivagaminathan and S. Ramakrishnan, "A hybrid approach for feature subset selection using neural networks and ant colony optimization", *Expert Systems with Applications*, Vol.33, pp. 40-60, 2007.
[22] C. L. Huang, "ACO-based hybrid classification system with feature subset selection and model parameters optimization", *Neurocomputing*, Vol.73, pp. 438-448, 2009.
[23] H. R. Kanan, K. Faez, and M. Hosseinzadeh, "Face recognition system using ant colony optimization-based selected features", in *IEEE Symposium on Computational Intelligence in Security and Defense Applications*, pp. 57-62, 2007.
[24] G. V. Wouwer, S. Livens, P. Scheunders, D. V. Dyck, "Color Texture Classification by Wavelet Energy Correlation Signatures," *Proceedings of the 9th International Conference on Image Analysis and Processing*, vol. 1, pp.327 - 334, 1997.

M. B. Imani was born in Tehran, Iran on Sep 1985. She received her Bachelor of Science degree in Computer Science from Shahid Beheshti University, Tehran, Iran and her Master of Science degree in Artificial Intelligence Engineering from Alzahra University, Tehran, Iran. She is currently PhD student at Computer and Information Technology Department, Amirkabir University of Technology, Tehran, Iran.
Her research interests are artificial intelligence, machine learning, data and text mining, educational data mining and swarm intelligence.

T. Pourhabibi was born in Tehran, Iran on July 1983. She received her Bachelor of Science degree in Software Engineering from Shahid Beheshti University, Tehran, Iran on February 2007 and her Master of Science degree in Artificial Intelligence Engineering from Alzahra University, Tehran, Iran on March 2011. Her interest fields of study are: artificial immune system, intrusion detection, machine learning, parallel processing, cloud computing, artificial intelligence and data mining.

Dr. M. R. Keyvanpour is an Associate Professor at Alzahra University, Tehran, Iran. He received his BS in software engineering from University of science and Technology, Tehran, Iran. He received his M.S and PhD in software engineering from Tarbiat Modares University, Tehran, Iran. His research interests are including image retrieval and data mining.

Dr. R. Azmi was born in Iran on May 1968. He received his BS degree in Electrical Engineering from Amirkabir University of Technology, Tehran and his MS and PhD degrees in Electrical Engineering from Tarbiat Modares University, Tehran, Iran in and 1999 respectively. Since 2001, he has joined Alzahra University, Tehran, Iran. He was an expert member of Image Processing and Multi-Media working groups in ITRC (From 2003 to 2004), Optical Character Recognition working group in supreme council of information and communication technology (From 2006 to 2007) and Security Information Technology and Systems working groups in ITRC (From 2006 to 2008). He was Project Manager and technical member of many industrial projects. Dr Azmi is founder of Operating System Security Lab (OSSL), Medical Image Processing Lab (MIPL), Face and Facial Expression Recognition Lab (FFERL), Web-based Anomaly Detection Lab (WADL) and Optical Character Recognition Lab (OCRL) in Alzahra University. He is currently an Assistant Professor of Computer Engineering at Alzahra University.