Feature Dimensionality Reduction for Recognition of Persian Handwritten Letters Using a Combination of Quantum Genetic Algorithm and Neural Network

Mohammad Javad. Aranian[1], Moein. Sarvaghad Moghaddam[2*], Monireh. Houshmand[3]
1- Department of Electrical and Computer Engineering, Imam Reza International University, Mashhad, Iran.
Email: Mj.aranian@gmail.com
2- Department of Electrical and Computer Engineering, Semnan University, Semnan, Iran.
Email: moeinsarvaghad@yahoo.com (Corresponding author)
3- Department of Electrical and Computer Engineering, Imam Reza International University, Mashhad, Iran.
Email: m.hooshmand@imamreza.ac.ir

Received: January 2016 Revised: January 2017 Accepted: March 2017

ABSTRACT
Curse of dimensionality is one of the biggest challenges in classification problems. High dimensionality of problem increases classification rate and brings about classification error. Selecting an effective subset of features is an important point in analyzing correlation rate in classification issues. The main purpose of this paper is enhancing characters recognition and classification, creating quick and low-cost classes, and eventually recognizing Persian handwritten characters more accurately and faster. In this paper, to reduce feature dimensionality of datasets, a hybrid approach using artificial neural network, genetic algorithm and quantum genetic algorithm is proposed that can be used to distinguish Persian handwritten letters. Implementation results show that proposed algorithms are able to reduce number of features by 19% to 49%. They also show that recognition and classification accuracy of resulted subset of features has risen, by 7/31%, comparing to primitive dataset.

KEYWORDS: Dimensionality Reduction of Features; Recognition of Persian Handwritten Letters; Genetic Algorithm (GA); Quantum Genetic Algorithm (QGA); Neural Networks.

1. INTRODUCTION

Classification is a process in which machine learns to assign new inputs to pre-defined classes [1]. Different methods are used to classify patterns. Artificial neural networks and learning algorithms such as K Nearest Neighbors and Support Vector Machine are some of the existing and practical methods that are used in patterns classification [2].

One practical application of algorithms and classification tools is their use in large datasets to designate one pattern with high number of features to a group of patterns. In this situation to achieve required accuracy and to speed up training and experimenting process, methods for reducing dimensionality of features are needed [6], [7], [8]. A categorization of algorithms used for features' selection and their comparison are also presented in [3].

"Optical Character Recognition" (OCR) [4] is a method in which a computer system can recognize texts available in digital images and convert them to text files. Nowadays, OCR is widely used in many contexts. Handwriting recognition, car plate recognition, extracting keywords from image, and indexing image based on content are some of these applications [4].

One of the challenges of OCR is reduction of classification speed and accuracy as a result of imposing lots of features to classifier algorithm. Hence, to achieve high accuracy and to increase training and testing speed, dimensionality reduction of problem is needed [5].

In [5], a genetic algorithm-based method for selecting subset of features in Persian OCR has been presented. For dimensionality reduction of problem, Bayesian Classification and genetic algorithm have been used in this paper. To recognize printed Persian character a combination of genetic algorithm and simulated annealing has been used in [9]. In [10], Particle Swarm Optimization and genetic algorithm have been used for better recognition of Persian handwritten digits. In [11], a hybrid method including neural networks and ant colony optimization has been presented in which neural network has been used as a classifier function used in ant colony optimization (ACO).

In this paper, firstly, a hybrid approach for reducing features dimensionality of datasets using neural network and genetic algorithm has been presented to recognize Persian handwritten letters. Then, to enhance proposed method performance, Quantum Genetic Algorithm has been used instead of genetic algorithm. Presented hybrid method has been tested using heady dataset that includes 70000 images of scanned handwritten letters. Results show that hybrid genetic algorithm-neural network strategy can reduce features dimensionality by 19% and quantum genetic algorithm can reduce number of features by 49%. It also has affected on accuracy and recognition rate of Persian handwritten letters on related dataset by 10%.

This paper is organized as follows: in Section 5 how the features are extracted has been expressed. Different methods for selecting subset of features have been discussed in section 3. In Section 4, short explanations about quantum genetic algorithm have been presented. The hybrid proposed method have been discussed in Section 5 and datasets and how they are used have been discussed in Section 6. Then, proposed method has been evaluated in Section 7 and finally conclusion has been presented in the last section.

2. FEATURES EXTRACTION

Features extraction is an important phase in OCR systems that may affect recognition phase quality. In recognition phase, a code or feature vector is assigned to each character or word input pattern which is that pattern indicator in features space and distinguishes it from other patterns units.

Due to segmentation and recognition phases, there are two main differences among Persian and Latin OCR systems. Because of major differences between Persian and Latin method of writing, it is not possible to apply Latin OCR's segmentation and recognition methods to Persian texts. Complexities available in Persian writing increase commercial OCR systems complexities. That is why most OCR software packages are not able to support Persian and Arabic languages [12].

Features extraction techniques can be searched for, in methods related to four general groups of pattern recognition.

- Template matching
- Statistical methods
- Structural methods
- Neural networks

In this paper, 63 features have been extracted for each character using statistical methods. Lots of these features have been extracted based on character segmentation to 9 areas. Feature extraction using segmentation method eliminates context's or character's language limitation, remarkably. Therefore, extracted features can be used for many languages.

3. DIFFERENT METHODS FOR SELECTING FEATURES SUBSET

Algorithms for selecting features' subset are divided into two main categories: Filter method and wrapper method [6].

3.1. Filter method

In this technique, no classification function is used. In other words, no feedback from applied learning algorithm will be used. This is a pre-selected method which is independent from applied machine learning algorithm. Features subsets are evaluated using other concepts.

Filter technique works in a way in which, firstly, a weight is calculated for each feature. Then, these weights will be sorted and features with lowest weights are eliminated. A threshold is used for features weights. Then results of a features subset are employed to a classifier system as input. Figure 1 shows how filter technique works.

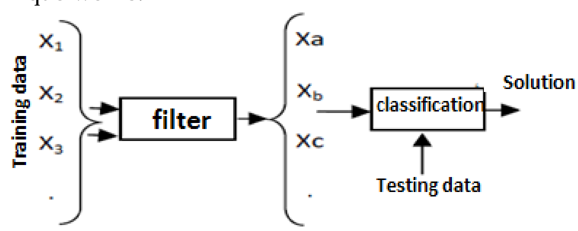

Fig. 1. Filter technique mechanism [7].

3.2. Wrapper method

Wrapper method is known as a black box. In this method a classification function is used for evaluating fitness of feature subsets. This technique uses the feedback that has been applied to learning algorithm. A genetic algorithm has also been used to search for valid features. The main reason for using genetic algorithm is that this algorithm can establish a random search and is not prone to stuck in local minimum [8].

Anyway, crossover operator used in this algorithm works as hybrid solutions, while keeps successful selections of previous feature. In other words, this technique is a feedback method that uses machine learning algorithm in feature selection procedure. Evaluation is done in each feature selection by execution of inductive algorithm during learning and testing phases. Figure 2 shows how this technique works [8].

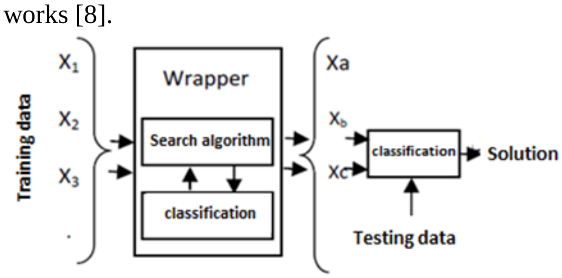

Fig. 2. Wrapper technique mechanism [8].

4. QUANTUM GENETIC ALGORITHM

The smallest unit of information which is saved in a quantum computer is named as q-bit [13]. A q-bit can take values 0 or 1. A q-bit state can be represented as follows:

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \tag{1}$$

While α and β are complex numbers that specify probability amplitude of corresponding states. $|\alpha|^2$ specifies the probability of q-bit to be in state "0" and $|\beta|^2$ specifies the probability of q-bit to be in state "1". Considering the fact that q-bit will take either value "0" or "1", we will have:

$$|\alpha|^2 + |\beta|^2 = 1 \tag{2}$$

Quantum genetic algorithms present replies in a probabilistic format. Each quantum-chromosome in a space is represented with n q-bits:

$$q_j^t = \begin{bmatrix} \alpha_{j1} & \alpha_{j1} & \dots & \alpha_{jn} \\ \beta_{j1} & \beta_{j2} & \dots & \beta_{jn} \end{bmatrix} \tag{3}$$

In which j = (1,2,...,m) shows quantum-chromosome number in solution space, m shows total number of chromosomes in solution space, n shows number of qubits available in quantum-chromosome or optimization problem dimension, and t shows generation number of evolutionary algorithm. In fact,〚〚[α〛^t ji   β^t ji]〛^T shows i-th qubit from j-th chromosome in t-th generation. The main benefit of representing data using qubit is that a chromosome can have different values in solution space. E.g. consider following quantum-chromosome:

$$\begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \frac{1}{2} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & \frac{\sqrt{3}}{2} \end{bmatrix} \tag{4}$$

The states of this quantum-chromosome can be represented as follows:

$\frac{1}{4}|000\rangle + \frac{\sqrt{3}}{4}|001\rangle - \frac{1}{4}|010\rangle - \frac{\sqrt{3}}{4}|011\rangle + \frac{1}{4}|100\rangle + \frac{\sqrt{3}}{4}|101\rangle - \frac{1}{4}|110\rangle - \frac{\sqrt{3}}{4}|111\rangle$

In which the probabilities of states:
$|000\rangle$ ،$|001\rangle$ ،$|010\rangle$ ،$|011\rangle$ ،$|100\rangle$ ،$|101\rangle$ ،$|110\rangle$ ،$|111\rangle$

Are equal to $\frac{1}{16}$ ،$\frac{3}{16}$ ،$\frac{1}{16}$ ،$\frac{3}{16}$ ،$\frac{1}{16}$ ،$\frac{3}{16}$ ،$\frac{1}{16}$ ،$\frac{3}{16}$, respectively. Obviously, mentioned quantum-chromosome can form eight different states. This means each chromosome in quantum genetic algorithm comprises information embedded in many solutions, so only one quantum-chromosome is enough to represent eight different states.

5. PROPOSED HYBRID METHOD

Proposed method of this paper is based on evolutionary theory which says only populations remain in sequential generations that are fitter than others. Operators such as crossover and mutation will also be applied on populations of a generation for diversity purposes. In other words, we are looking for best answer using population evolution and choosing the bests. In this paper, artificial neural network has been used as classifier function for evaluating generated population by genetic algorithm and quantum genetic algorithm. Figure 3 shows hybrid algorithm mechanism that has been used for obtaining smallest features set with maximum efficiency and desired effectiveness on output.

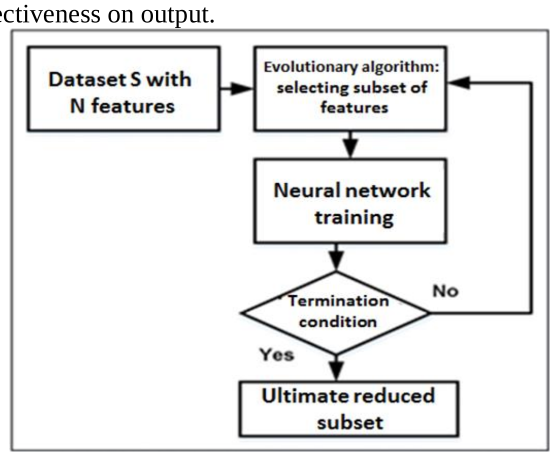

Fig. 3. Hybrid algorithm structure

In figure 3 dataset s and its subsets include n features for each character that most efficient features are specified by applying these features to hybrid algorithm.

5.1. Appropriate features selection using GA

As mentioned in Section 2, 63 features have been extracted from each handwritten character. In this method using genetic algorithm appropriate features for letters classification are selected among 63 extracted features from characters images. For encoding these 63 features in genetic algorithm, a binary chromosome length in 63 is defined. If a bit in that chromosome is 1, that feature will be used in letters classification and if a bit is 0, that feature will not be used in letters classification (figure 4).

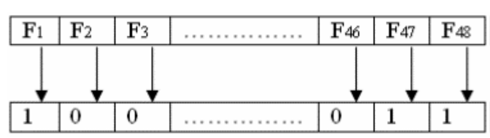

Fig. 4. Presentation of a chromosome and how to select features.

Binary chromosomes are generated randomly to form initial population in genetic algorithm. Then, fitness value is computed for each chromosome using fitness function that artificial neural network method has been used for this purpose in this paper. Appropriate parents will also be selected using Roulette wheel technique.

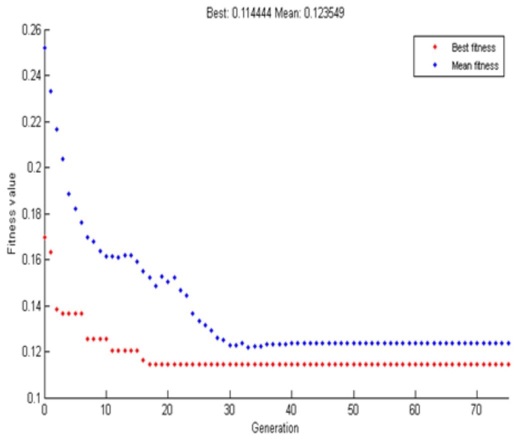

Fig. 5. Genetic algorithm convergence

5.2. Selecting effective features using QGA

Quantum genetic algorithm pseudo code is as Figure 6:

```
Procedure QEA
begin
        t=0
   i.   initialize Q(0)
  ii.   make X(0) by observing Q(0)
 iii.   evaluate X(0)
  iv.   store the best solutions
        among X(0) into b
   v.   while (not termination-
        condition) do
         begin
           t=t+1
  vi.      make X(t) by observing the
            states ofQ(t-1)
 vii.      evaluate X(t)
viii.      update Q(t) using Q-gate
  ix.      store the best solution
            among X(t) into b
         end
end
```

Fig. 6. Quantum genetic algorithm pseudo code [15].

Above algorithm steps are as follows:

i. The first step is initializing *Q(t)*. In this step, initial value $\frac{1}{\sqrt{2}}$ is assigned to all q-bits α_i^0 and β_i^0(i =1,2,…, n) for all quantum-chromosomes q_j^0 (j =1,2,…,m). This means the probability of observing "0" and "1" is equal. Here, n is length of chromosome vector and m is number of chromosomes in solution space.

ii. This step makes a set of binary chromosomes, *P(0)*, from quantum-chromosomes *Q(0)*. In this step, depending on $|\alpha i|^2$ and $|\beta i|^2$ values where i is $(i = 1,2,\dots,n)$, binary chromosomes $p(0) = \{x_1^0, x_2^0, \dots, x_m^0\}$ are made in generation (0). A binary chromosome x_j^0, $(j = 1,2,\dots,n)$, is a binary solution length in n. Making a binary chromosome based on quantum-chromosome is done by observation. To make a xi bit based on a q-bit $\begin{bmatrix}\alpha_i \\ \beta_i\end{bmatrix}$, the following equation is used where U(.,.) is the function that generates uniform random numbers.

$$x_i = \begin{cases} 0 & u(0,1) < \alpha_i^2 \\ 1 & Otherwose \end{cases} \qquad (S)$$

iii. Set of binary chromosomes resulted from second step are evaluated using fitness function.

iv. The best solution among $P(0)$ is stored into b.

v. Algorithm will be executed while termination condition is not satisfied.

vi. In while loop P(t) binary values are made by observing the states of Q(t-1).

vii. P(t) is evaluated.

viii. Qubits are updated using quantum-gates. Quantum-gate is an operator that is applied on qubit where $|\alpha'|^2 + |\beta'|^2 = 1$. Here α' and β' are qubit's updated values. For updating quantum-gate [α β], following equation is used:

$$[\,\alpha'_i \quad \beta'_i\,]^T = \begin{cases} U(\Delta\theta_i)[\,\alpha_i \quad \beta_i\,]^T & if\ \alpha\beta > 0 \\ U(-\Delta\theta_i)[\,\alpha_i \quad \beta_i\,]^T & Otherwise \end{cases} \qquad (6)$$

In which Δϴ is extracted from table 1. In this paper, the following equation has been used for quantum-gate operator:

$$U(\Delta\theta_i) = \begin{bmatrix} cos(\Delta\theta_i) & -sin(\Delta\theta_i) \\ sin(\Delta\theta_i) & cos(\Delta\theta_i) \end{bmatrix} \qquad (7)$$

ix. The best found solution is stored into b.

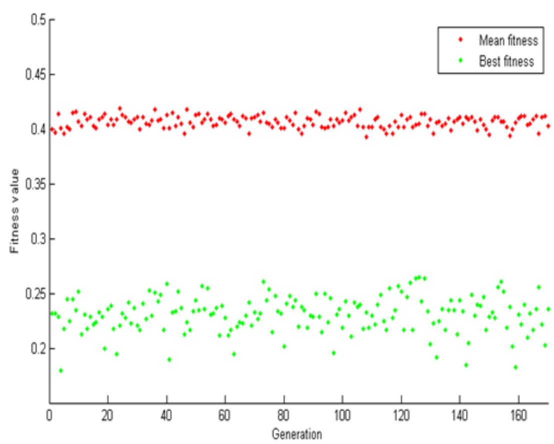

Fig. 7. Quantum genetic algorithm convergence.

Table 1. The Δɵ values for updating q-bit [14].

Xi bi	F(x)>f(b)	Δɵi
0 0	0	0.001π
0 0	1	0.001π
0 1	0	0.08π
0 1	1	0.001π
1 0	0	0.08π
1 0	1	0.001π
1 0	0	0.001π
1 0	1	0.001π

5.3. Calculating features' set fitness using neural network

As mentioned in previous section, in this paper neural network has been used as classifier function for determining fitness value of features' subset. In this phase, features subset has been given to neural network with constant number of neurons in hidden layer for training purposes, which number of neurons in input layer depends on number of features in related subset. Each subset of data has been divided into two sections training and testing on a 70/30 ratio.

After training, neural network will be tested using some new data. Wrong number of classified samples is considered as related subset's error. Error inverse is considered as a criterion for that subset fitness. Hence, each subset of features has an estimation error that helps to determine best subset. Therefore, neural network helps individuals in the population of genetic algorithm and quantum genetic algorithm to find best solution.

6. DATA SETS

One of the problems with Persian OCR is lack of standard datasets for comparing different applied methods. In this paper the heady dataset [22] including 70 thousand and 17 thousand training and testing data from Persian handwritten character images with a resolution of 200 dpi (dot per inch) has been used. At first, all characters' images were resized to a constant size using MATLAB software and then data pre-processing functions like noise removal were applied on them. In the next phase, by combining statistical and structural techniques, 63 features were extracted from all characters. Therefore, a resulted dataset including 63 features was created for 17000 Persian letters to be used in proposed method.

7. EVALUATION OF PROPOSED METHOD

In this research firstly all extracted features including 63 features were analyzed using SMO algorithm in RapidMiner 5 data mining software. Achieving an acceptable accuracy in recognition of Persian handwritten letters was the result of this work. Then, firstly using proposed hybrid genetic algorithm-neural network method, 63 selected features were applied to proposed algorithm as input. Proposed hybrid technique could reduce number of features by 19.04%. Once more features resulted from this algorithm were tested using SMO technique in RapidMiner software. The result showed recognition precision of about 90.49% for Persian handwritten letters. In the next step, 63 features were applied on quantum genetic algorithm. Quantum genetic algorithm could reduce feature dimensionality by 49/20%. The result of applying these features to SMO algorithm in RapidMiner software was achieving 91.83% accuracy for recognition of handwritten letters. Here, an important issue is use of features that have been extracted by authors of this research to achieve high accuracy in recognition of handwritten character. That is the reason why results did not change considerably after applying proposed algorithm. It is predicated that by applying this algorithm on a dataset with more variety of features and greater number of inappropriate features better results can be achieved. It is worth mentioning that one of the objectives of this paper is comparing functionality of genetic algorithm with that of quantum genetic algorithm regarding convergence rate and finding global optimum of problem. Obtained results can be found in Table 2. Considering the fact that evolutionary algorithms are heuristic and time-consuming techniques and the fact that proposed algorithm needs chromosome fitness to be specified by neural network in each phase, the algorithm mechanism is time-consuming comparing to other methods. However, the results obtained from features reduction may cause an increase in speed and accuracy of character recognition during test phase.

Table 2. The results obtained prior to the execution of algorithm and after it.

DATA SET	***TRAIN-ING DATA***	***NO. OF TESTED DATA***	***NO. OF FEATURES***	***INITIAL ACCURACY WITH SMO ALGORITHM***	***NO. OF FEATURES WITH GA-ANN ALGORITHM***	***DIMENSION REDUCTION PERCENTAGE WITH GA-ANN***	***RECOGNITION ACCURACY WITH GA-ANN***	***NO. OF FEATURES WITH QGA-ANN ALGORITHM***	***DIMENSION REDUCTION PERCENTAGE WITH QGA-ANN***	***RECOGNITION ACCURACY WITH QGA-ANN***
HODA	**70.645**	**17.706**	**63**	**85.11**	**51**	**19.04**	**90.49**	**32**	**49.20**	**91.83**

8. CONCLUSION AND FUTURE WORKS

All OCR systems in all existing languages are in need of high accuracy and speed for characters recognition. High dimensionality of problem and great number of features are main reasons for speed and accuracy reduction in these systems. Numerous methods have been proposed for features reduction. In this paper a hybrid approach combining neural network with genetic algorithm and quantum genetic algorithm is used to select an effective subset of features.

Since genetic and neural network learning algorithms are heuristic algorithms, runtime prolonged in training phase. However, the purpose is reducing the number of features in order to boost speed and accuracy in recognition of Persian handwritten characters in test phase. Positive results obtained from dimensionality reduction compensate training phase costs in test phase. In this paper, by combining genetic algorithm with neural network classifier, number of features fell from 63 to 51. Additionally, combining quantum genetic algorithm network with neural network decreased number of features by 49%. This reduction in number of features means selecting the features that are more influential in characters classification and leads to the result to be obtained more accurately and faster. As mentioned in section 7, features used in this paper are extracted by its own authors. Since extreme efforts have been made for appropriate feature to be extracted, selecting a useful subset among good features seems delicate and difficult.

As mentioned before, comparing effectiveness of genetic algorithm and quantum genetic algorithm is another purpose of this research. Results show that quantum genetic algorithm is more capable to find optimized answer comparing with genetic algorithm. This point encouraged this paper's authors to further research on solving other optimization problems and features dimensionality reduction using quantum genetic algorithm.

REFERENCES

[1] Kulkarni, R. S.; Vidyasagar, M., **"Learning decision rules for pattern classification under a family of probability measures"**, *IEEE Transactions on Information Theory*, Vol. 43(1), 154-166, 1997.

[2] Kulkarni, R. S.; Lugosi, G.; Santosh, V. S.**, "Learning pattern classification – A survey"**, *IEEE Transaction on Information Theory*, Vol. 44(6), 1998.

[3] A. Ahmad; L. Dey **,"A feature selection technique for classificatory analysis "** , *Pattern Recognition Letters* 26, 2005.

[4] Liana, M. Lorigo and Venu Govindaraju, **"Offline Arabic Handwriting Recognition: A Survey"** *IEEE Transactions on pattern analysis and machine intelligence*, Vol. 28, No. 5, May 2006.

[5] Ahmadreza Kheirkhah, Esmaeil Rahmanian, **"optimization of recognition of Farsi handwriting characters based on effective feature selection by GA",** *8th Conference on intelligent systems*, Ferdowsi University of Mashhad, 2007(in Farsi).

[6] Sakineh Zanganeh; Reza Javanmard; Mohamad mahdi Ebadzadeh, "**A Hybrid Approach for Features Dimension Reduction of Datasets using Hybrid Algorithm Artificial Neural Network and Genetic Algorithm-in Medical Diagnosis**", *3rd data mining conference*, 2009(in Farsi).

[7] J. Jarmulak; S. Craw, **"Genetic Algorithms for Feature Selection and Weighting"**, Appears in *Proceedings of the IJCAI'99 workshop on Automating the Construction of Case Based Reasoners*, 1999.

[8] S. Mika; G. Ratsch; J. Weston; B. Scholkopf; A. J. Smola; K. R. Muller, **"Invariant feature extraction and classification in kernel spaces",** *Advances in Neural Information Processing Systems*, Massachusetts, USA: MIT Press, vol. 12, 2000.

[9] M. Saberi; D. Safaai, **"Feature Selection Method Using Genetic Algorithm For The Classification Of Small and High Dimension Data"**, *IEEE Transaction On Pattern Analysis And Machine Intelligence*, Vol. 23, No. 11, 2005.

[10] Reza Azmi , Boshra Pishgoo , Narges Norozi , Maryam koohzadi , Fahimeh baesi **" A hybrid GA and SA algorithms for feature selection in recognition of hand-printed Farsi characters "** 978-1-4244-6585-9/10/2010 IEEE .

[11] Najme ghanbari, Seyyed Mohammad razavi, Sedighe ghanbari **"Optimizing Recognition System of Persian handwritten digits"** *Majlesi Journal of Multimedia Processing* Vol. 1, No. 2, June 2012.

[12] Rahul Karthik Sivagaminathan, Sreeram Ramakrishnan **"A hybrid approach for feature subset selection using neural networks and ant colony optimization"**, *Expert Systems with Applications*, Vol. 33, 2007, pp. 49–60.

[13] Kuk-Hyun Han, Jong-Hwan Kim **"Genetic Quantum Algorithm and its Application to Combinatorial Optimization Problem",** 2002.

[14] Zakaria Laboudi, Salim Chikhi, **"Comparison of Genetic Algorithm and Quantum Genetic Algorithm",** *SCAL Group of the MISC Laboratory*, University Mentouri, Algeria. 2012

[15] Kulkarni, R. S.; Lugosi, G.; Santosh, V. S., **"Learning pattern classification–A survey"**, *IEEE Transaction on Information Theory*, Vol. 44(6), 1998.

[16] S. Mika; G. Ratsch; J. Weston; B. Scholkopf; A. J. Smola; K. R. Muller, **"Invariant feature extraction and classification in kernel spaces",** Advances in Neural Information Processing Systems, Massachusetts, USA: MIT Press, Vol. 12, 2000.

[17] M.Soryani and N.Rafat, **"Application of Genetic Algorithms to Feature Subset Selection in a Farsi OCR"**, *World Academy of Science, Engineering and Technology* 18, 2008.

[18] L. S. Oliveira, N. Benhmed, R. Sabourin, F. Bortolozzi, C.Y., Suen, **"Feature Subset Selection Using Genetic Algorithms for Handwritten Digit Recognition"** *Proc. XIV Brazilian Symposium on Computer Graphics and Image Processing*

(SIBGRAPI'01), pp.362, 2001.

[19] Tahereh Pourhabibi, Maryam Bahojb Imani, Saman Haratizadeh, **"Feature selection on Persian Fonts: A Comparative Analysis on GAA, GESA and GA"**, *Procedia Computer Science 3*, 2011.

[20] S. K. Singhi; H. Liu, **"Feature Subset Selection Bias for Classification Learning"**, *Appearing in Proceedings of the 23rd International Conference on Machine learning*, Pittsburgh, PA, 2006.

[21] I. S. Oh; J. S. Lee; B. R. Moon, **"Hybrid Genetic Algorithms for Feature Selection"**, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 11, November 2004.

[22] http://farsiocr.ir