The GSA Algorithm at Two Methods of Feature Selection and Weighting Features to Improve the Recognition Rate of Persian Handwrite Digits with Fuzzy Classifier

Najme Ghanbari
Department of Electrical Engineering, Faculty of Engineering, University of Zabol, Zabol, Iran

B. Somayeh Mousavi
Department of Electrical Engineering, Hatef Higher Education Institue, Zahedan, Iran

Mohammad Allahbakhsh
Department of Computer Engineering, Faculty of Engineering, University of Zabol, Zabol, Iran

ABSTRACT

In this paper, using Gravitational search algorithm or GSA recognition rate of Persian handwritten digits can be improved. Two methods have been proposed to improve the recognition rate. in the first method, with using of version binary of Gravitational search algorithm or BGSA, we choose optimal features among the overall features extracted. Finding the best feature sets from entire extracted features, not only the number of features and computational burden will be reduced but also recognition rate will be significantly improved. Also In this paper, real version of GSA or RGSA has been used in different way (secondary methods) to improve recognition rate. In this method, instead of choosing some of the features, one random weight has been assigned to each feature. Indeed, feature vector has been multiplied in weight vector to obtain a new feature vector. This Weight vector is obtained with RGSA. After several iterations, RGSA algorithm determines Weight set of features so that Classification accuracy increases. In this paper, the fuzzy classifier is used for classification. Fitness function in BGSA and RGSA algorithms is the number of fuzzy classifier errors and the aim is to make this value minimum. The obtained results confirmed that these algorithms have proper performance.

Keywords

Increase the recognition rate, feature selection, binary Gravitational search algorithm (BGSA), Real Gravitational search algorithm (RGSA), Weight vector of features, fuzzy classifier and recognition of Persian handwritten digits.

1. INTRODUCTION

Recognizing Persian handwritten digits is a field of pattern recognition and image processing that has been undergone extensive research and is still evolving. Digits recognition is currently used in different applications such as office automation, review and verification of banking cheque, types of banking businesses and banking tickets, financial reporting forms, statistical reporting forms, postal codes, reading mailing address, name, number of national, rial quantities that are written on bills and forms by hand, data importing applications, etc. Many activities have been done regarding handwritten digits recognition. In a research aimed for Persian Handwritten digits recognition, characteristics of places, Bays classification and Markov chain is used [4]. Another research has been done in final Persian Handwritten digits using fuzzy methods [1]. In an another research performed in algorithm, matching shapes and classifying the nearest distance from the representative or representatives of each class for Persian Handwritten digits recognition has been used [5]. In another method aimed of classification by neural networks and combining three features of characteristics of places, gradient is improved and Krish has been used for recognizing digits [8]. In reference [6] of neural networks and combining two-class-ten classification has been used for Persian Handwritten digits recognition. Practically, using just one system of handwritten digits recognition is faced with some challenges; most importantly, high recognition rate is necessary. In the field of Persian language, because of high similarities of the digits and also the difference in how they map, creating a system with an acceptable accuracy for practical purposes is faced with some problems. Hence it is necessary to develop ways to improve their accuracy. One of the most efficient ways to improve accuracy in a handwritten digits recognition system is selecting optimal features among the entire set of extracted features.

By feature selection, it means selecting a small set of features that are collectively ideal to describe specific patterns. This will avoid extracting unnecessary features. Different features can usually be measured from the real world patterns. However, using the overall extracted features is not affordable in the end. Also, is the feature vector becomes very large; recognition system performance will be decreased. feature selection can have a significant effect on appropriate recognition rate of classification algorithm. The main aim of feature selection is to reduce feature vector dimension in classification in a way that a reasonable classification rate is achieved.

Finding a subset of features among a large set of feature is an issue that is faced with in many problems. It is not generally clear that what subset of features make the most distinction for regarded pattern classes, and on the other hand, it is not affordable to consider all the current subsets. Population based search methods are appropriate ones to reduce the number of features.

Feature selection with genetic algorithm and feature selection with innovative algorithms such as particle Swarm algorithm, meta-heuristic algorithm such as Ants Algorithms are some instances of these methods. Most of these methods begins the search parallel to an initial population and then each individual’s merits is determined based on a fitness function and the population data are updated using fitness values. This process will be continued until the algorithm is convergent. For example, genetic algorithm has been used in a research to determine a group of features to be applied in a classification and evaluation function clustering of K of the nearest neighbor. This process cause increasing system speed and even its performance [12]. In fingerprint classification with selecting the best features using genetic algorithm, good results have been achieved [11]. Also, in different way, Persian handwritten recognition rate improved. In this method instead of choosing some of the features by GSA, one random weight has been assigned to each feature in order to improve recognition rate. In this method one random weight has been assigned to each feature so that multiplied by the weight vector of features in the feature vector, a new feature vector be achieved and by applying this new vector, the recognition rate and classification accuracy will increase.

Finding the Weight vector by mathematical and statistical computational methods is very difficult. This Weight vector is obtained with real version of GSA (RGSA).

In this paper, HODA database [9] is used which is a very large database of Persian handwritten digits. Also a fuzzy method is used to classification [1]. In the used fuzzy method, the features of database training samples are extracted and Gaussian membership functions are obtained for fuzzy model using them. The applied fuzzy model in this article is a Mamdani model with 10 rules. Each rule is assigned to just one number. In testing phase, this fuzzy model is used to classify the database experimental samples. Some of Population-based algorithms are Binary Gravitational Search Algorithm (BGSA), Binary Genetic Algorithm (BGA) and Binary Particle Swarm Optimization (BPSO). Among them, the BGSA is the newest one. In this paper, two methods were proposed in order to increase the recognition rate of Persian handwritten digits by the fuzzy classifier. One method is finding the optimal features from entire extracted features and other method is weighting to features. To perform two methods, Gravitational Search Algorithm is used. BGSA is used to solve the first method and RGSA, is used to find weight-optimized features in the second method.

The remained samples are used to evaluate the two algorithms.

The obtained results showed that high costs should be spent due to replicate the problem in GSA algorithm in the phase of selection features and finding optimal weights for features through the used methods. However, in the test phase (training and test data) higher recognition rate is achieved. In feature selection method, in addition higher recognition rate the cost is reduced due to reducing the problem dimensions.

2. DATABASE

This recognition method is applied to Farsi hand-written digits in the HODA database [9], which is a standard database. Approximately 11942 binary images have been extracted out of 102353 digits of two types of registration forms filled in by students in undergraduate and graduate. These forms are scanned with a high-speed scanner with an accuracy of 200 pints per inch. These forms contain letter information including name, surname, father's name and some digital information written in separate letters and digits in pre-determined spaces. Two large databases from digits and letters have been prepared out of these forms which digit database is used in this article. The overall digits in this database are 102352 digits in which 60000 digits are considered as training samples, 20000 digits as test samples and 22352 digits as remaining digits. Training samples have been used to train proposed classifier, testing samples to test Classifier and the remaining digits have been used in population-based algorithms.

3. EXTRACTING FEATURES

To recognize Persian handwritten digits, some features such as zoning features, Geometric Moments, Zernike Moments, instant descriptors, Invariant Moments, histogram display, features of characteristics of places, Krish, border profile, gradient, gradient histogram, projection, etc[1-5]. The zoning features are used for recognition in this article. Zoning features has low computational complexity and has often been used in recognizing Persian and Latin handwritten digits [7].

3.1 Image Normalization

By normalization, it means to convert input images of any size to an image with n*m pre-determined dimensions. In binary images of database digits, there are many zeros in four sides (i.e. white spots that are related to the field and contain no data). Initially, the tetragonal containing redundant information is removed and each digit is confined to one tetragonal. The digits are normalized to 32*32 in terms of size. For this reason, 32*32 images are divided into smaller blocks to reduce the number of features and the spots within each block are enumerated and are considered as feature. The block are selected 4*4, therefore the image have 64 blocks and 16 spots. The number of black spots in each 4*4 block is considered as recognition system features. Each digit consists of 64 features in this phase which the digits within these features can be changed from zero to 16.

4. FUZZY MODEL

There are 10 rules in the fuzzy system introduced in reference [1] in which each rule is used to recognize each digit. An instance of such rule can be seen in reference [1]. The fuzzy set of $A_{i,j}$ can be obtained from training samples [1]. The algebraic multiplication and minimum are used for "and" operator.

Multiplication algebraic operator results have been reported because the results were better. Two methods are proposed for improving the results obtained in fuzzy method, in which they explain.

5. GRAVITATIONAL SEARCH ALGORITHM

Gravitational search algorithm is presented by Mrs. Rashdi and colleagues in 2009 [10]. The binary version of this algorithm is also provided by the same authors in 1386 [3]. Here, a brief description about Gravitational search algorithm is given. For more details you can refer to the above references. in GSA optimization is done using the laws of gravity and motion, in an artificial system with discrete time. According to the law of gravity, each mass understands the location and status of other objects through the force of gravitational attraction. Therefore, this force can be used as a tool to exchange information. Designed Optimal detector can be used to solve the optimization problem in which each answer to the question if position in space can be defined and its similarities with other answers to the problem is expressed as a distance. The masses are determined according to objective function.

Initially system space is determined. environment including a multidimensional coordinate system in the space that the problem is defined. Every point in space presents a solution. Searcher agents are a set of masses. Each mass has three specifications. A: position b: gravitational mass c: inertial mass. These masses are derived from the concept of active gravitational mass and inertial mass in physics. The position of the mass corresponds to a solution of a problem. In the system referred to there are the laws of gravity and motion. In GSA there are a series of mass m. Each mass that is placed randomly at one point in space is the solution of the problem. In any moment of time, masses are evaluated, then the mass shift is calculated after calculating the relationship that is in [10, 3]. System parameters such as gravitational mass, inertial mass and Newton's gravitational constant are updated at each stage according to the corresponding relations that is in [10, 3]. Stop condition can be determined after the specified period. In figure 1 is given pseudo-code for this algorithm.

```
1- Determine the system environment and primary initialize
2- Initial placement of masses
3- Assessment of masses
4- Update the G, best, worst, Mi, Mg
5- Calculate the force on each mass
6- Calculate the acceleration and velocity of each mass
7- Update the position of masses
8- If the stop condition is not fulfilled go to step 3
9- End
```

Figure1:pseudo-code for used algorithm [10]

5.1 Selecting of Optimal Features for Persian Handwritten Digits using the Algorithm BGSA

Each database digits of 64 features has been extracted from zoning method. The appropriate features are selected for classifying digits here using binary Gravitational search algorithm (BGSA) among this 64 features. The initial population of masses is randomly created. In the problem of feature selection with the BGSA, each mass can be an answer to the problem. Any answer or mass is defined as a string of zeros and ones. String length is equal to the total number of features (64). If each bit of string is one, it means using the feature related to the bit in digits classification and if each bit of string is zero, it means not using the feature related to the bit in digits classification. After creating the initial population of masses, the fitness value is calculated for each mass using the fitness function which is the number of errors here. At each stage, the best mass is selected as for the fitness values. The algorithm runs as the number of iteration. And finally, the good features are obtained by selecting the mass with the lowest fitness. Initial population in our algorithm is about 20 mass. With this method recognition rate is significantly improved. Also number of features dropped from 64 to 30. Results are shown in table (1). Table (2) shows the performance matrix of this way on the test samples.

5.2 Weighting to Features using RGSA

Another way that we offered for increasing the recognizing rate is, instead of using binary version of GSA and choosing some of features as optimum features, real version of this algorithm and weighting to features is used. In real rate of GSA, first group as a set of real masses. Every mass is as a string with 64 lengths. Every bite of this string has a real amount that these amounts a point the analogous weight with every feature. These amounts a point with fitness function that is the amount of phase classification errors here. After performing RGSA in amount of specific repetition, mass has the lowest fitness function is determined. This mass is the feature weight vector that multiplied in amount of feature vector and will produce a new feature vector. With this way, recognizing rate will improve much. Result is shown in table (1). Table (3) shows the effective matrix of this way in trial samples. Also figure (2) show the convergence of BGSA and RGSA algorithms from two above way for decreasing the phase classification error.

6. CONCLUSION

The Zoning features are used in this article due to low computational complexities it has. This method can be tested with other features. In this paper, the two methods mentioned to increase the recognition rate of Persian handwritten digits. In two methods GSA was used. BGSA was used to select the optimal features and RGSA was used to find the optimal weights of features. Both methods significantly improve the recognition rate. These methods can also be implemented with other algorithms based on population. There is always a tradeoff between training and testing stages. That is, if high speed decision making and low computational costs is achieved in application stage, much more costs should be spent to appropriately select the patterns and features. The more additional costs are incurred for training, the more the costs will be compensated in practice stage. Most methods used until now for recognition of Persian handwriting digits were compared to each other in different articles are different in databases. Hence, their comparison is not very secure, because data quality has been different in different articles. Mr. Khosravi's database is used in this article which is better compared to other databases in terms of both data volume and data diversity [9]. Furthermore, the results of recognition are without any preprocessing and post processing operations; hence the obtained recognition rate is a good one.

Table 1: Recognition rate in different scenarios

Recognition method	Number of features	recognition rate on test samples	The number of error on test samples	recognition rate on train samples	The number of error on train samples
Fuzzy method with Total Number of features	64	80.19	3962	86.21	8276
Feature selection using BGSA with The initial population of 20 masses and 300 iteration	30	84.55	3098	90.01	6003
Feature selection using BGSA with The initial population of 20 masses and 200 iteration	33	84.25	3155	89.39	6330
Weighting to features using RGSA with The initial population of 20 masses and 250 iteration	64	83.33	3345	88.51	6715
Weighting to features using RGSA with The initial population of 20 masses and 200 iteration	64	83.18	3361	88.43	6736

Table 2: Final recognition system performance on test samples while assigned weights to each of the features with BGSA

	0	1	2	3	4	5	6	7	8	9	The number of error	Percent error
0	1821	79	2	2	6	72	6	9	1	2	179	8/950
1	162	1654	84	4	16	2	32	14	6	26	346	17/30
2	0	92	1533	123	64	0	66	52	12	58	467	23/35
3	5	11	177	1613	47	0	120	7	3	17	387	19/35
4	0	29	46	139	1601	38	75	11	19	42	399	19/95
5	50	56	7	1	17	1836	23	1	8	1	164	8/199
6	5	9	156	23	84	3	1405	74	14	227	595	29/75
7	11	24	70	0	17	4	73	1809	0	2	201	9/550
8	7	51	3	0	23	46	17	1	1798	54	202	10/10
9	8	36	0	0	5	7	162	0	76	1706	294	14/70

Table 3: final recognition system performance on test samples while assigned weights to each of the features with RGSA

	0	1	2	3	4	5	6	7	8	9	The number of error	Percent error
0	1835	68	6	2	4	62	10	9	1	3	165	8/2500
1	156	1677	75	1	11	0	37	11	1	31	323	16/1500
2	2	134	1471	119	74	0	40	75	0	85	529	26/4500
3	5	22	156	1686	49	1	62	10	0	11	314	15/7000
4	4	30	86	127	1582	51	54	14	8	44	418	20/9000
5	64	49	3	1	8	1841	22	1	10	1	159	7/9500
6	3	9	107	12	24	5	1411	96	17	316	589	29/4500
7	0	12	118	7	34	1	54	1767	0	7	233	11/6500
8	10	60	2	0	10	38	32	2	1838	8	162	8/0999
9	11	38	0	1	5	0	132	0	36	1777	223	11/1500

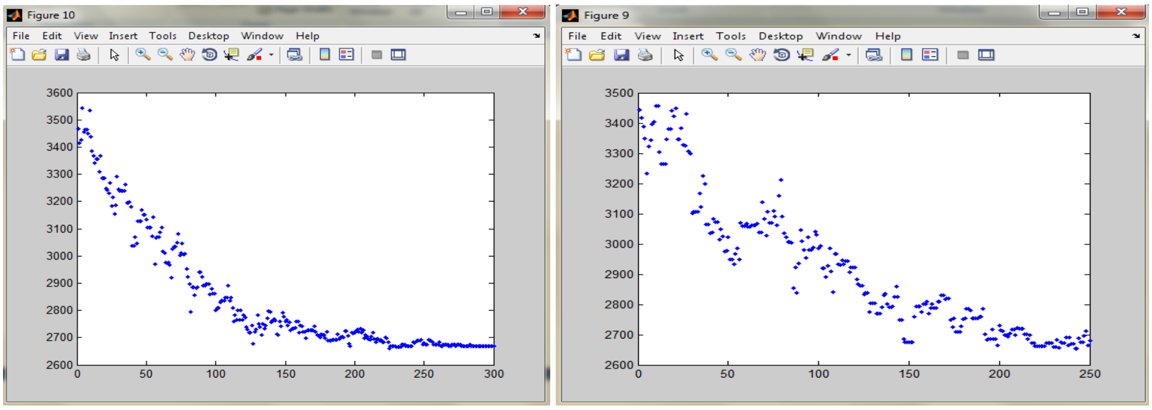

Figure 2: from left to right Convergence of BGSA and RGSA at Methods of feature selection and features weighting shows in order to Reduction of Classification Error

7. REFERENCES

[1] Johari majd. V, razavi. S. m, fuzzy recognition of Persian handwritten digits, First Iranian Conference on Machine Vision and Image Processing, pp. 144 -151, 2000.

[2] Razavi. S. m, sadoghi yazdi. H, kabir. E, feature selection for Persian handwritten digits recognition using of Genetics algorithm, Seventh Annual Conference of Computer Society of Iran , pp. 285-292, 2001.

[3] Rashedi,esmat, Nezamabadi Pour,H, Saryazdi,S. binary Gravitational search algorithm, the first member of congress on fuzzy and intelligent systems,mashhad, 2007.

[4] Nafisi. H. r, kabir. E, Persian handwritten digits recognition, Second Iranian Conference on Electrical Engineering, pp. 295- 304,1994.

[5] Darwish,A., Kabir,E., Khosravi,H., application of shape match in Persian handwritten digits recognition, Journal of Jihad, No. 22, 2005.

[6] Nahvi,M., Rafiei,M.,Ebrahim Pour,R.,Kabir,E.,The combination of the two-class classifiers for recognition of Persian handwritten figures, Sixteenth Conference of Electrical Engineering, Iran, 2008.

[7] Tohidi,H., Nezamabadi Pour,H,Saryazdi,S., feature selection algorithm using binary population of ants", First International Conference on Fuzzy and Intelligent Systems, Ferdowsi University of Mashhad, 2005.

[8] Khosravi,H., Kabir,E., The letters and numbers recognition of Persian handwritten in the national exam registration forms , Fifteen of the first page master's thesis,2006.

[9] H. Khosravi, E. Kabir," Introducing a very large dataset of handwritten Farsi digits and a study on their varienties", Pattern recognition letters 28, 1133-1141, 2007.

[10] E. Rashedi, H. Nezamabadi-pour, S. Saryazdi, "GSA: A Gravitational Search Algorithm", Information Sciences 179, 22322248, 2009.

[11] Y.Qi, J.Tian, R.W.Dai, "Fingerprint classification system with feedback mechanism based on genetic algorithm", IEEE, 1998.

[12] J.G.Smith, T.C.Fogarty and I.R.Johnson, "Genetic selection of feature for clustering and classification" , 1994.