International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869 (O) 2454-4698 (P) Volume-8, Issue-5, May 2018

A Combination of Neuron Networks and Genetic Algorithms for Nom Character Recognition

Truong Thi Huong, Nguyen Van Truong

***Abstract*— Nom character, Vietnamese ancient language, is a cultural heritage that has played a particularly important role in creating a brilliant literary culture throughout the centuries in Vietnam. Nom character is very valuable in studying the life of the ancient Vietnamese in many fields literature, thought, philosophy, language, law, morality, etc. Reading and writing the character in Vietnam nowadays are not popular, so the digitization of its letters is an important task. In this paper, we present an approach of Neuron Networks and Genetic Algorithms to provide a good optical recognition method of Nom character. This contribution helps to build a tool that can identify, transform Nom character into Vietnamese language effectively.**

***Index Terms*— Nom character, Recognition, Neuron Networks, Genetic Algorithms.**

I. INTRODUCTION

Nom character is a logographic writing system formerly used to write the Vietnamese language. It used the standard set of classical Chinese characters to represent Vietnamese vocabulary and some native Vietnamese words, while new characters were created on the Chinese model to represent other words. Nowadays, many people are loyal to brush and ink pot in studying Nom Character inherited from their forefathers to study the life of the ancient Vietnamese.

Building Nom optical character recognition software (Nom-OCR) is a necessity as with other languages. Nom-OCR will act as a strong motivator for the study of Nom character, explorer precious resources of the nation for thousands of years in politics, culture and society. Nom recognition systems can refer to the technical models of other OCRs, especially hieroglyphs OCR such as Chinese, Japanese. Based on the study of OCR models, the authors propose the recognition model for Nom as shown in Figure 1.

In the model, the source can be an image or a PDF file. The source of the OCR system can include many types of information, such as images of different language types. Therefore, it is necessary to conduct page analysis, character recognition. After separating the character from the page, we proceed to the necessary pre-processing steps, split into blocks, split the blocks into lines, split the lines into discrete characters. From discrete characters, we extract the character of the character to be used to carry out the identification. The result of the identification step may not be the final step, but will be post-processing, dictionary-based, grammar, etc., to determine the final result [1].

Truong Thi Huong, Faculty of Mathematics, Thai Nguyen University of Education, Thai Nguyen City, Vietnam, Mobile No. +(84)-914061236
Nguyen Van Truong, Faculty of Mathematics, Thai Nguyen University of Education, Thai Nguyen City, Vietnam, Mobile No. +(84)-915016063

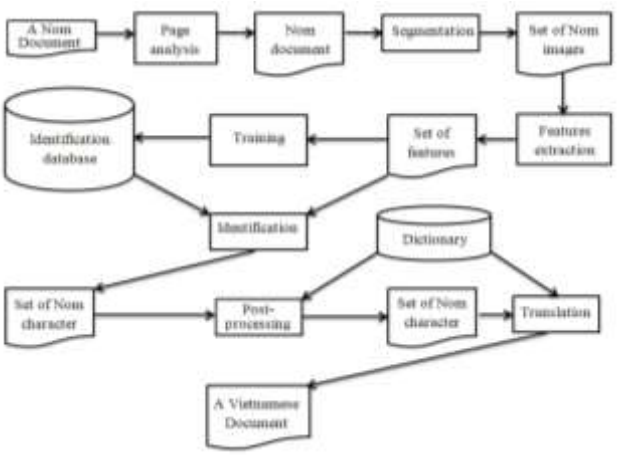

Fig. 1. A Nom character recognition model

II. NEW GANN ALGORITHM FOR NOM CHARACTER RECOGNITION

2.1 Previous works

Artificial Neural Network (ANN) is an information processing model illustrating to the way information is processed by biological neural systems with the expectation that it can handle intelligent tasks in similar way that human brain can do. An artificial neural network is configured for a specific application such as pattern recognition, data classification, etc. by a learning process from a set of training samples. The essence of learning is the adjustment of weights indicate linkage among neurons.

Neural networks have been successfully applied in classification by means of highly adaptive learning, discriminatory and generalization. It has been widely used in the studies of hieroglyphic identification such as Korean [3, 8], Indian [4], Chinese [2, 5, 6, 9], Japanese [7] that can achieve accuracy of 98.82%.

Direct propagation neuron is one of the most common type of networks. Back-propagation algorithms (BPA) are used to calculate the gradient of error functions using the diffusion rule. The errors after the initial calculation on the transmission line will be transmitted back from the output through each layer, called "back-propagation".

In the process of network training, the initial weight is an importan factor that will affect to acuracy. If the weights are initialized with large values, then the total input signal from the beginning of the absolute value is large and therefore the output of the network is only 0 or 1. This causes the system to be clogged at a local minimum or in a flat area near the starting point. Therefore, these weights are usually initialized by small random numbers in the range $[-1/n, 1/n]$, where n is the number of weights connected to the class l. Wessels and Barnard have investigated and pointed out that the initialization of W_{ij} weights should be within the range of

$\left[-3/\sqrt{k_i}, 3/\sqrt{k_i}\right]$, where k_i is the number of links of the neurons j to neurons i [11].

Studies on GA combined with ANN was initiated by Montana and Davis [13]. In 1989, they presented the successful application of GA in the ANN network and demonstrated that GA found optimal sets of weights better than BP in some cases. Kumar Reddy [12] and Yas Abbas Alsultanny [14] achieved better results and reduced training time when using GA to initialize the initial set of multi-layered linear neural networks. These works provide our motivation for applying GA to optimize the weight for ANN in the optical recognition of Nom characters.

2.2 The convergence of neural networks

Our survey was conducted with a 3-layer network, the input layer consists of 25 neurons, the hidden layer consists of 5 neurons and the output layer consists of 1 neuron. The initial set of weights is randomly taken around the point 0.5, which is the midpoint of Log-Sigmoid activation function (logsig). The results are achieved after 14 training times as shown in Table 1.

Table 1. The survey of the convergence of neural networks

Time	Loop Number	Time	Loop Number
1	failed	8	failed
2	81007	9	failed
3	failed	10	35672
4	85060	11	65742
5	14542	12	failed
6	failed	13	78649
7	42335	14	65903

Results in Table 1 show that with the same algorithms, the structure networks and parameter are selected similarly, then result of the network training is strongly depend on the initialized weights, even there are 6 times is failed in the total of 14 times. This can be explained as follows: nature of learning algorithm back-propagation errors is the method of reducing the deviation gradient so the initialization of the initial value of the weighting of small value will randomly make converged networks on different minimum values. If lucky, the network will converge to the global minimum value, otherwise the network can fall into local extreme and can not get out then leads to network failure.

2.3 GANN algorithm

Because of its extensive, random and selective search mechanism, GA usually finds the global extremes, but it is difficult to reach global extremes. On the one hand, we want GA to maintain population diversity (spreading search space) to avoid premature convergence to local extremes; On the other hand, when "globally localized", we want GA to narrow the search area to "indicate global extremism". The first goal is often achieved by selecting appropriate adaptive function and appropriate population replication. To achieve the second goal, we have to divide the evolution into two phases. In the second phase, we must adjust hybrid operators, mutations, reproducers; selective methods; assessment of adaptation; as well as must revise the parameters of evolution to reach the global extremes. Implementation of such a model would be very complicated. Therefore, it is necessary to combine GA with other local optimization methods.

The learning methods in ANN perform a "local search" in the weighted space (based on information about the derivative of the error) so there are two disadvantages. Firstly, the set of weights obtained is often not globally optimal ones. Secondly, the learning process may not converge or may converge but very slowly. Therefore, it is necessary to combine "localized" methods of ANN with "universal" algorithms such as GA. From these observations, we find that it is possible to combine GA with ANN in an uniform framework to improve the efficiency of ANN. GA will enclose the global minimum region of the error function, then ANN derives from this set of weights to reach the global minimum. The proposed algorithm GANN is depicted in details as in Fig. 2.

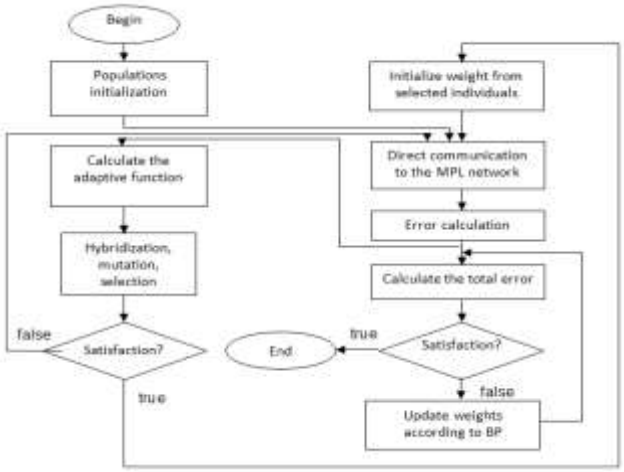

Fig. 2. Diagram of GANN algorithm

The model consists of two stages of network training. The first stage uses GA with a straight forward step to accelerate the whole process of networking. GA performs a global search and optimal search near the initial point, weight vector, for the second stage. Each chromosome is used to encode the neuron's weights. The adaptive function of GA is defined as the sum of the error squares of the corresponding neural network. Thus, the problem becomes unlimited optimization in order to find a set of decision variables that minimize the objective function.

III. EXPERIMENTS

3.1 Experimental procedure

The process is conducted in three main steps: data preparation, network training and recognition.

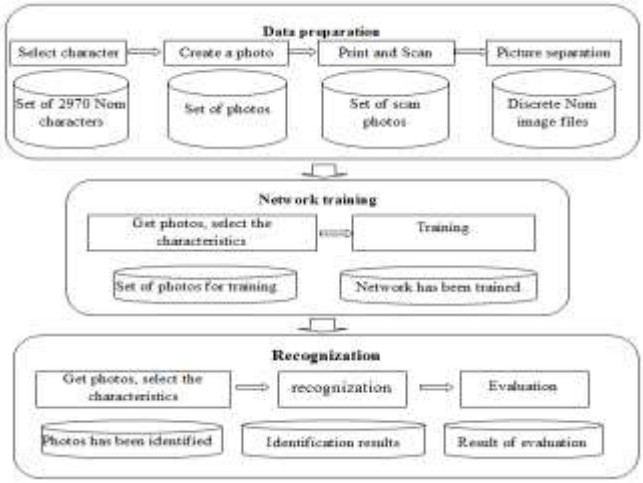

Fig. 3. Experimental procedure

3.2 Experiment data preparation

Data is an important component in the testing and evaluation of identification methods and is also the basis for accurate assessment of the accuracy of the results. Since the database for the subject has not yet been available, the team of embedded systems experiments have conducted a real data set by statistically compiling words that appear over 10 times in the "Kieu story" of great writer named Nguyen Du. Statistical results obtained 495 words. With that 495 data set, we use 3 different types of fonts: Nom A, Nom B, Nom Na Tong, each font is in bold and lower case. So each letter has 6 samples, the total sample of the data set is 2970. Each character pattern is named by the rule: ID_Sample _Font_Type. This set of test data is in good quality, noise is almost zero. Some explanations are:

- ID is the code for the character, each character has a unique ID.
- Sample is a sampling mode, numbered 0, 1, 2 ... with this data set is always 0 because it is also taken to scan a letter printed on white paper A4, black ink.
- Font is the font name of the template.
- Type is 0_0 if lowercase, and 0_1 is bold.

Fig. 4. Some Nom characters in the experimental data set

3.3 Experiment process

Verification method

In this paper, we use 6-fold cross validation for our experiments [10]. That is, with the experimental data set we decide to choose the sample set of 5 sets of 5 different words for training, one belonging to the remaining sample for identification. So the training set and the experiment identification do not intersect in sample and font, the inference of the method of identification is checked. The program will use the knowledge of this type of fonts to identify the image of the other fonts. Therefore, it is possible to measure the practical applicability of the method.

Experiment

The program was tested on the Lenovo X220, with an Intel Core i5-2520M CPU, 2.5GHz 2.5GHz, 2GB RAM and Windows 7.

Figure 5 illustrates the interface of the test program.

Network configuration selection includes the number of classes in the network, the number of neurons in each class, and the selection of learning parameters such as error thresholds, learning coefficients, and transfer functions.

The chromosome configuration allows for the selection of parameters for genetic algorithms including: hybridity, mutation, number of generations, population size. And it allows us to see the optimal set of sequences obtained after the implementation of the genetic algorithm.

Fig. 5. Network training and identification

3.4. Evaluation

To ensure that the combination of genetic algorithms found the best set of weights for the next training, we conduct a GA experiment with different hybrid and mutation operations and found that the BLX 0.5 hybridization and similarity mutations be the most effective one.

For convenience of comparing GANN algorithms with the use of the primitive back propagation algorithm, we use the same set of parameters for both of these tests. The neural network is configured as follows: The number of layers in the network is 5, of which 3 layers are hidden, each has 10 neurons, the input layer is 100 neurons and the output layer is 1 neuron, the learning curve is 0.02 and the error threshold is 0.001. For the GA and GANN algorithms: the BLX 0.5 is used with a hybrid ratio of 0.8 and the mutation was identical with the mutation rate of 0.1, the population size is 100 and the search is conducted through 1000 generations.

Experimental results are shown in Table 2.

The accuracy in the table show that the effect of applying genetic algorithms in initializing the initial set of weights. The GANN algorithm can reach the highest result of 90.01%, and average result of 86.39% while back propagation algorithm achieve average result of 74.49% only. This result imply that initialization of weights will influences the outcome of neural network training. It also shows that the application of genetic algorithms to optimize inputs for network training actually yields better results than neural networks and the original back propagation algorithm.

Table 1. Performance comparison of GANN with ANN

Steps	Number of training characters	Recognition set	Number of recognition characters	Accuracy (%)	
				GANN	ANN
1	2475	HanNomA_0_0	495	90,01	89.89
2	2475	HanNomA_0_1	495	84,48	60.00
3	2475	HanNomB_0_0	495	87,27	75.55
4	2475	HanNomB_0_1	495	88,28	74.54
5	2475	NomNaTong_0_0	495	82,82	86.26
6	2475	NomNaTong_0_1	495	85,45	60.67
Average				86,39	74,49

IV. CONCLUSIONS

This article presents results of neural network convergence, discusses the advantages and disadvantages of using neural networks and genetic algorithms in optimal character recognition. Based on this survey, we proposed algorithm that combines neural network and genetic algorithm for effective recognition of Nom characters. Our contribution is expressed in two folds. Firstly, we propose the first combined model for

the problem of Nom character recognition. Secondly, some parameter optimization technique are deployed to improve the recognition performance. Experimental results shows that the proposed algorithm produces better performance in comparison with that of recently published papers. In the future, we will build sets of data with varying levels of noise to improve the tool's ability to recognize different types of interruptions. We also would like to apply the method for other ORC problems with larger datasets.

REFERENCES

[1] P.V.Huong, T.M.Tuan, N.T.N.Huong, B.T.H.Hanh, L.H.Trang, V.T.Nhan, T.A.Hoang, V.Q.Dung, N.N.Binh, "Some methods of Nom recognization", ICT.rda, 2008.

[2] Mingrui Wu, Bo Zhang, Ling Zhang, "A Neural Network Based Classifier for Handwritten Chinese Character Recognition", *ICPR'00* – Vol.2, 2000.

[3] Il-SeokOh, Ching Y. Suen, "A class-modular feedforward neural network for handwriting recognition", *Pattern Recognition* 35, pp. 229-244, 2002.

[4] Srinivasa Kumar Devireddy, Settipalliappa Rao, "Hand written character recognition using back propagation network", *Journal of Theoretical and Applied Information Technology*, 2009.

[5] Richard Romero, Robert Berger, Robert Thibadeau, and Dave Touretsky, "Neural Network Classifiers for Optical Chinese Character Recognition". *Proceedings of the Fourth*, 1995

[6] Richard Romero, David Touretzky, and Robert Thibadeau, "Optical Chinese Character Recognition using Probabilistic Neural Networks", *Article in Pattern Recognition* vol. 30, no 8, pp.1279-1292, August 1997

[7] Tadashi Horiuchi, Satoru Kato, "a study on japanese historical character recognition using modular neural networks", *International Journal of InnovativeComputing, Information and Control*, vol. 7, No. 8, 2011.

[8] H.-H. Song, S.-W.Lee, "A self-organizing neural tree forlarge-set pattern classication", *IEEE Trans. Neural Net-works 9*, vol. 3, pp. 369-380, 1998.

[9] H.-M. Lee, C.-C.Lin, J.-M.Chen, "A preclassication method for handwritten Chinese character recognition viafuzzy rules and SEART neural net", *Int. J. Pattern Recogni-tion Artif.Intell*. vol. 12, no. 6, pp. 743-761, 1998.

[10] Juan Diego Rodrıguez, Aritz Perez, Jose Antonio Lozano, "Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation", *IEEE Transactions on pattern analysis and machine intelligence*, vol. 32, no. 3, March 2010.

[11] L.F.A. Wessels, E. Barnard, "Avoiding false local minima by proper initialization of connections", *IEEE Trans. Neural Networks* vol. 3, pp. 899-905, 1992.

[12] R.Ashok Kumar Reddy, G. Venkata Narasimhulu, Dr. S. A. K. Jilani, Dr D.Seshappa, "Genetic Algorithm based Gait Recognition", *International Journal of Electronics and Computer Science Engineering ISSN*- 2277-1956.

[13] David J. Montana, Lawrence Davis "Training feedforward neural networks using genetic algorithms" *IJCAI'89 Proceedings of the 11th international joint conference on Artificial intelligence, vol 1, 1989.*

[14] Yas Abbas Alsultanny, Musbah M. Aqel, "Pattern recognition using multilayer neural-genetic algorithm", *Neurocomputing,* no.51, pp. 237 – 247, 2003.

BIOGRAPHY

Truong Thi Huong is a a lecturer in the Faculty of Mathematics at Thai Nguyen University of Education, from where she received a Bachelor of Informatics in 2010. She finished her master course on Software Technology at at Vietnamese National University in 2014. She is interested in artificial intelligence, natural language processing with some published works in international conference such as ICIICA2012, and KSE2017.

Nguyen Van Truong is a lecturer in the Faculty of Mathematics at Thai Nguyen University of Education, from where he received a Bachelor of Mathematics and Informatics in 2000. He finished his master course on Computer science at Vietnamese National University in 2003. He is currently a PhD student at Institute of Information Technology (IOIT), Vietnamese Academy of Science and Technology (VAST). He has taught a wide variety of courses for UG students and guided several projects. He has published several papers in National Journals & International Conferences. His research interests are embedded systems and artificial immune systems..