Application of Synergetic Neural Network in Online Writeprint Identification

Sanya Liu, Zhi Liu, Jianwen Sun, Lin Liu
Engineering & Research Center for Information Technology on Education, Central China Normal University
National Engineering Research Center for E-Learning, Hubei Wuhan, China
lsy5918@gmail.com
doi:10.4156/jdcta.vol5.issue3.12

Abstract

Synergetic neural network (SNN) associates synergetics with artificial neural network, it can rigorously deal with the behavior of network in the mathematical theory, and have the advantage of fast learning, short pattern recalling time and so on. In this paper, a pattern recognition method based on the self-adaptive attention parameters presented on the basis of analyzing the key technology of SNN, and the advanced algorithm will be employed in the online writeprint identification, the key point of this algorithm is that it can correct initial mis-identified patterns through measuring similarity between the prototype pattern and the testing pattern in the evolution of order parameters. Experimental results show that the advanced SNN has better performance and robustness than the SNN based on balanced attention parameters. Further, the network's self-learning ability and recognition performance is greatly improved by using advanced SNN.

Keywords: *Online Writeprint, Synergetic Neural Network (SNN), Global Feature, Attention Parameter, Order Parameters*

1. Introduction

With the rapid development of web technologies and growing interest in surfing the Internet, distributing messages in the on line community has become an important means of communication. Unfortunately, on line messages also are usually misused for the transmission of unsolicited or unsuitable information such as spamming and offensive messages. Furthermore, criminals have been using online messages in order to transmit illegal materials, including fraudulent e-commerce activities, pornography materials, and so on. A common characteristic of online messages is an onymity, so the main objective of researching on line writeprint is tracing identities of illegal messages' providers. Writeprint [1] means stylistic features which represent writing-style of authors hidden in online messages, including character n-grams, punctuation, special symbols, digit, high-frequency words, word and sentence-length distributions, function words, content words, and part-of-speech tag n-grams distribution [2] and so on. These abundant features are combined as a global feature，so the problem of writeprint identification is transformed into the matching problem of the different global features. But we find that the similarity in global features which represent different authors' patterns is great, therefore, it become a critical issue of writeprint identification that how to reduce the correlation among different authors' global features and improve self-adaptive ability in the identification using SNN.

In the late 80s, Haken proposed that synergetics could be applied to pattern recognition [3]. An important point of synergetics in pattern recognition's field is that pattern recognition means pattern formation. Pattern's formation of the whole system is that initial state appears as several ordering subsystems, order parameter belonging to a certain subsystem will win in competition with other order parameters. Finally, this ordering subsystem will dominate the whole system and make it into a certain ordering condition. In pattern recognition, this mechanism is used that once the set with a variety of features is given; a certain order parameter will compete with other order parameters. Ultimately, the strongest supporting order parameter will win and force the system to appear original missing features. The process corresponds to the transformation of the testing patterns being mapped to the basic pattern. Application of synergetics in the field of pattern recognition has been a conspicuous achievement,

including character recognition, face recognition, shape recognition, fingerprint recognition, license plate identification, and synthetic aperture radar image target recognition and so on. Yudashkin et al. (1996) [4] transformed competition among different patterns in to competition among several time-independent order parameters based on synergetic self-organizing neural network. Joerg et al. (1995) [5] conducted a series of experiments to apply SNN with constraints to image segmentation. Wagner et al. (1994) [6] proposed an advanced synergetic algorithm and apply it to recognition of tires in industry and classification of coins' echo.

In this study, SNN will be first applied to online writeprint identification. In the step of preprocessing, abundant linguistic features are extracted from 2500 samples, and then we combined different kind of features to form the most representative global feature for each pattern of author. In the step of identification, we present a method of optimizing SNN using the dynamic attention parameter.

The organization of the paper is as follow. After briefly introduced the basic concept of the approach in Section 2, we described the application of SNN in online writeprint identification and improvement of this algorithm in Section 3. In Section 4, experimental results for testing proposed identification method and performance of advanced SNN are given. Concluding remarks and future directions are given in Section 5.

2. Theory of Pattern Recognition based on SNN

Process of pattern recognition corresponds to evolution process of a kinetic equation [7]. It is assumed that a virtual particle moves on a potential topographic map, when the particle enters the bottom of an attractive valley, the pattern corresponding to it will be identified. Thus, pattern recognition system can be considered as a synergetic identification system, the arrangement of the initial condition manifested as partial ordering subsystems, the order parameter belonging to the strongest supporting subsystem will win in competition and dominate the whole system to make it into a specific ordering condition [8]. According to the basic concept of synergetics, the process of pattern recognition can be considered as the process of competition among several order parameters. Synergetics mainly researches the kinetic equation which is described by condition vector q; it is assumed that a system has M components:

$$q = (q_1, q_2, ..., q_M) \tag{1}$$

so the recognition process can be constructed by the kinetic equation for q as following:

$$\dot{q} = \sum_{k=1}^{M} \lambda_k (v_k^+ q) v_k - B \sum_{k \neq k'} v_k (v_{k'}^+ q)^2 (v_k^+ q) - Cq(q^+ q) + F(t) \tag{2}$$

where B and C are constants and $\lambda = [\lambda_1, \lambda_2, ..., \lambda_M]$ is a vector of constants labeled λ_k, called as attention parameters. In the dynamic process, q is pulled into one of the M prototype patterns v_k $(k = 1,2,...,M)$ via intermediate state $q(t)$, which can be described as

$$q(0) \to q(t) \to v_k \tag{3}$$

The testing pattern q is decomposed in the prototype patterns v_k $(k = 1,2,..., M)$ as

$$q = \sum_{k=1}^{M} \xi_k v_k + w \quad v_k . w = 0 \tag{4}$$

where ξ_k is the order parameter, w is residual vector. The se order parameters governed the

dynamic process in t he trans ition when th e o ld st ate lo ses i ts stabi lity. The int roduction of o rder parameters [9] can greatly simplify the network's behavior as the formula (5)

$$\xi_k = v_k^+ . q \tag{5}$$

wh ere v_k^+ represents the adjoint vector of v_k, in a SNN, v_k^+ s are obtained by the given prototype v_k s and each v_k^+ co rresponds to the v_k. v_k^+ obeys the orthonomality relation:

$$\left(v_k^+, v_j\right) = \delta_{kj}, \quad \delta_{kj} = \begin{cases} 0 & k \neq j \\ 1 & k = j \end{cases} \tag{6}$$

Corresponding dynamic evolution equation of order parameters is

$$\dot{\xi}_k = \xi_k(\lambda_k - D + B\xi_k^2) \tag{7}$$

where $D = (B+C)\sum_{k'} \xi_{k'}^2$, according to equations (1-7), SNN model can be constructed as following:

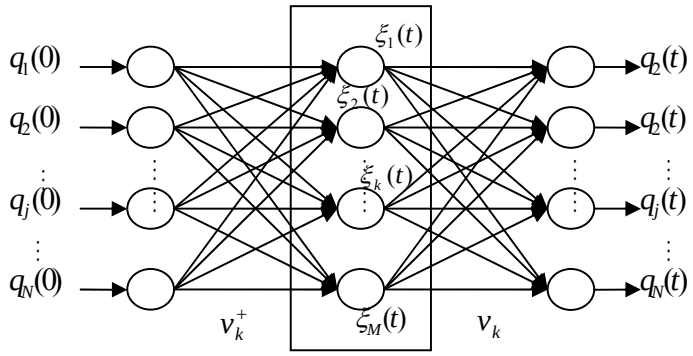

Figure 1. Model of SNN

In the network, N neurons in the input layer correspond to the initial values of N testing patterns; the layer of order parameters has M neurons, whose initial values are obtained from the input layer based on Eq. (5) . The con nection weights betw een input layer and middle layer are deter mined by adjoint patterns v_k^+ s, while the connection weights[10] between middle layer and outp ut layer are determined by prototype patterns v_k s. The relationship among order parameters ξ_k s is established by evolution equation (7); as each un it in the layer corresponds to an internal prototype pattern, each unit can b e used for identifying a s pecific pattern. Due to the pre sence of variable D, all uni ts can interact with each ot her, w hich w ill cau se th e co mpetition a mong the m. Fin ally, only o ne patt ern can win in the competition and force the output layer to present the identification results.

3. Application of SNN in Online writeprint identification and its improvement

3.1. The framework of online writeprint identification based on SNN

Based on the SNN algorithm we summarized in section 2, we proposed a framework for online writeprint identification based SNN as shown in figure 2.

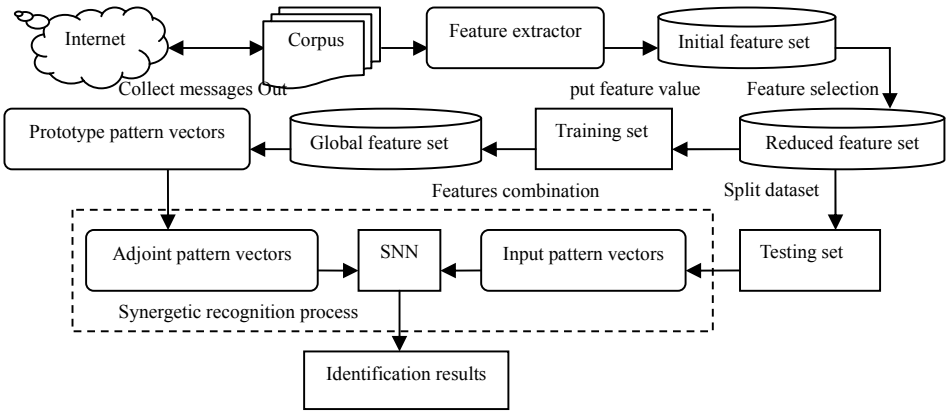

Figure 2. The framework of online writeprint identification based on SNN

In the framework, SNN is applied for identifying authors' identities as the critical classification technique. According to Haken's standard synergetic recognition model, we can construct a 3-layer neural network described in section 2 to realize the learning process of prototype patterns and adjoint patterns, the most important problem in building learning model is obtaining the prototype patterns which represent different authors' writing style. In this study, prototype patterns are obtained by using SCAP [11] algorithm which is rapid for obtaining prototype pattern vectors by averaging the training samples simply. The adjoint patterns are obtained by using pseudo-inverse method, which can make the basic vectors used in representing the different authors' patterns orthogonal. General running process of SNN is divided into the training phase and the recognition phase as following:

1) Training phase:

Step 1: Read the global feature vectors of different authors' training sets and calculate prototype pattern vectors satisfying the condition of normalization and zero-mean.

Step 2: Obtain the adjoint pattern vector of the prototype pattern vector for each author pattern.

2) Recognition phase

Step 1: Read the initial feature vectors $q_k(0)\ (k = 1,2,...,N)$ representing identified patterns satisfying the condition of normalization and zero-mean.

Step 2: Multiply feature vectors $q_k(0)\ (k = 1,2,...,N)$ by network weights v_k^+ according to the equation: $\xi_k(0) = v_k^+ . q(0)$ in order to calculate the initial values of order parameters ξ_k s in the middle layer.

Step 3: Conduct the competition among the order parameters ξ_k s by performing the discrete evolution equation as following:

$$\begin{aligned} \xi_k(n+1) - \xi_k(n) &= \gamma(\lambda_k - D + B\xi_k^2(n))\xi_k(n) \\ D &= (B+C)\sum_{k'} \xi_{k'}^2(n) \end{aligned} \tag{8}$$

where γ is iteration speed, which determines the stability of SNN.

Step 4: Judge whether the values of ξ_k s in the evolution can reach stable condition or not, if the

process of evolution is stable, the pattern which the order parameter with the biggest module value represents will be attributed as the category of identified testing pattern, or continue performing evolution equation in Step 3.

3.2. Optimization of parameters in SNN

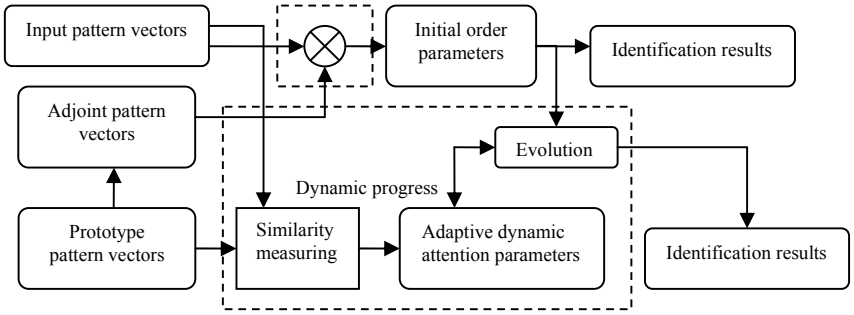

Figure 3. SNN model based on self-adaptive attention parameters

Fig. 3 presents an illustration of the improved SNN model; the key point of the improved algorithm is mainly reflected in the attention parameter. In the Eq. (7), λ_k is called as attention parameter which controls the final results and speed of patterns' evolution. If high-order items and fluctuating force are ignored, the general solution of the differential equation (7) is as following:

$$\xi_k \approx \xi_k(0)e^{\lambda_k t} \tag{9}$$

We can conclude that the attention parameter λ_k controls the speed of patterns' change based on formula (9), if the "attention" needs be paid to a specific pattern, the bigger attention parameter λ_k will be endowed to the pattern. Thus, it is likely to get the final victory even if the order parameter corresponding to the specific pattern has not the biggest in itial module value, which accords with the characteristics of biological sense.

Value of the attention parameter λ_k is divided into balanced and unbalanced. The balanced attention parameter [12] is defined as below:

$$\lambda_k = C > 0, B > 0 \tag{10}$$

Haken has proved that the output pattern of SNN can be obtained by directly comparing the initial module values of order parameters, so the final winner will be the pattern which the order parameter with the biggest initial module value represents. The method using balance parameters needs not go through the evolution of order parameters, but SNN loses self-learning ability in this situation.

The unbalanced attention parameter is more beneficial to selective pattern recognition. Especially when the difference among order parameters is little, the attention parameter plays an important role in the kinetic evolution of order parameters. We propose a self-adaptive method for selecting attention parameters based on the practical application of online writeprint, in this method, the unbalanced attention parameter are selected by measuring similarity between the prototype pattern vector and the identified pattern vector:

$$\lambda_k = \frac{\sum_{i=1}^{N}(v_{ki} - \bar{v}_k)(q_i - \bar{q}_i)}{\sqrt{\sum_{i=1}^{N}(v_{ki} - \bar{v}_k)^2 \sum_{i=1}^{N}(q_i - \bar{q}_i)^2}} \tag{11}$$

The formula (11) reflects the similar degree between the prototype pattern and the testing pattern, the closer the prototype pattern to the testing pattern, the greater corresponding attention parameter will be. It is a method of adaptively determining attention parameters. As for the iteration speed, we also use self-adaptive method to adjust γ, the setting of γ and constants B, C is as following:

$$\gamma = L / D \quad (0 < L < 2) \tag{12}$$

$$B = C = 1 \tag{13}$$

where L represents step length of iteration, which greatly affects the final recognition. Thus, the evolution equation (8) in section 3.1 is transformed as following:

$$\xi_k(n+1) - \xi(n) = \frac{L}{D(n)}(\lambda_k - D(n) + \xi^2(n))\xi(n) \tag{14}$$

where $D(n) = 2\sum_{k'} \xi_{k'}^2(n)$. $D(n)$ will change with the change of the attention parameters, and iteration speed γ reflects variation degree of $D(n)$.

4. Experimental works

4.1. Dataset and feature set

In this research, experimental dataset are derived from the customers' reviews in Amazon Commerce Website. Most previous studies conducted the identification experiments for two to ten authors. But in the on line context, reviews to be identified usually have more potential authors, and normally classification algorithms are not adapted to large number of target classes. To examine the robustness of the introduction algorithm, we identified 50 of the most active users (represented by a unique ID and username) who frequently posted reviews in these newsgroups. The number of reviews we collected for each author is 50. The main characteristics of the dataset are listed in Table 1.

Table 1. Description of experimental dataset

Language	No. of authors	Reviews of per author	Avg. length of reviews per author
English 50		50	856 characters

A feature set is composed of writing-style features predefined by previous researchers. As an important component of our framework, the feature set may significantly affect the performance of online writeprint identification. According to the previous stylometric studies and analysis [13], we extracted four types of feature set: lexical, syntactic, content-specfic, and idiosyncratic as shown in Table 2, these four types of feature set will be combined to form the global feature vector for each author' training set.

Table 2. Description of extracted feature set

Type	Category	Description
Lexical	Word Length	frequency of 1–20 letter words
	Sentence Length	sentence length in terms of character & sentence length in terms of word
	Vocabulary richness	richness(e.g., Yule's;K, Simpson's D)
	Function Words	frequency of function words (e.g., of, for)
	Short Words	frequency of 1–3 letter words
	Word Profile	bag-of-words(e.g., "senior", "editor")
	Word Bigrams	word bigrams(e.g., "senior editor")
	Word Trigrams	word trigrams (e.g., "editor in chief")
	Digits	digits (e.g., 1, 2, 20,200…)
	Letter(26 features)	a – z
	Upper-case character	e.g., A,AB,ABC…
	Special Characters	occurrence of special char. (e.g., @#$%^ …)
	Character *n*-grams (fixed-length)	n-gram frequency length of n = 2～5,variable
Syntactic	Punctuation	occurrence of punctuation (e.g., !;:,.?)
	POS Tags	frequency of POS tags (e.g., NP, JJ)
	POS Tag Bigrams	POS tag bigrams (e.g., NP-VB)
	POS Tag Trigrams	POS tag trigrams (e.g., VB-JJ-VBP)
Content	Content words	bag-of-words (e.g.,"senior", "editor")
Idiosyncratic M	isspelled Words	misspellings (e.g., "beleive", "thougth")

4.2. Experimental design

In the preprocessing, the initial feature set will be normalized to reduce the differences among eigenvalues. As the dimensionality of feature space is very large, we use the χ^2 metric [14] as formula 15:

$$\chi^2 = \sum_{i=1}^{k} \frac{(x_i - \mu_i)^2}{\sigma_i} \tag{15}$$

which calculates the expected and observed frequency for every item in every category, to identify features that are able to discriminate between the categories under investigation[15]. The dimension of the initial feature set is more than 15000. The dimension of the initial feature set is reduced to 2000 by using χ^2 feature selection method.

In the module of synergetic recognition, selection methods of the balanced and unbalanced attention parameter are used for analyzing and comparing identification performance. All experiments for the two methods are performed using 10-fold cross-validation which tests the robustness of SNN algorithm adequately. In the method of balanced attention parameters, we set $\lambda_k = B = C = 1$; in the self-adaptive method of unbalanced attention parameters, the iteration step length is set as 1.5, while the time of iteration is set as 500.

We investigate by performing authorship attribution on 5 and 50 authors while gradually increasing the amount for training data. Table 3 will be used to be compared performance of SNN based on balanced and unbalanced attention parameters, and we add identification results of SVM [16] to be compared with improved SNN, the identification rate is computed by formula 16:

$$Accuracy = \frac{Number\ of\ reviews\ whose\ author\ was\ correctly\ identified}{Total\ number\ of\ reviews} \quad (16)$$

Table 3 shows that the comparison results for the three algorithms when number of authors varied from 5 to 50.

Table 3. Identification results (% accuracy) of three algorithms

Algorithm \ No. of authors		50	40	30	20	10	5
SNN	Balanced attention parameter	68.31	70.23	74.13	80.45	87.51	92.87
	Self-adaptive attention parameter	80.49	82.52	84.26	87.28	89.34	95.65
SVM 78.61			80.24	86.15	88.25	92.32	96.19

We observed that the improved SNN kept over 80% accuracy when the number of authors varied from 5 to 50. But SNN based on balanced attention parameters achieved only 70.89% average accuracy when the number of authors varied from 30 to 50, and the method is effective for only identifying 5-20 authors. So it can be seen that improved SNN outperformed SNN based on balanced attention parameters in recognition accuracy and robustness. On the other hand, SVM[17] has shown good performance in previous related research of text categorization; we remarked that SVM outperformed improved SNN in distinguishing within the scope of 30 authors. However, the improved SNN showed better performance than SVM in distinguishing between 40 and 50 authors; the condition may be related with selection method of prototype pattern vectors in SNN. As several authors in the dataset have similar writing-style, they usually use the same keywords, punctuations and other character sets, which may be caused by using quotation from other authors. Pseudo-inverse method [18] is effective in pattern disruption, the similarity among different global feature vectors which represent authors' patterns can be greatly reduced by using the method.

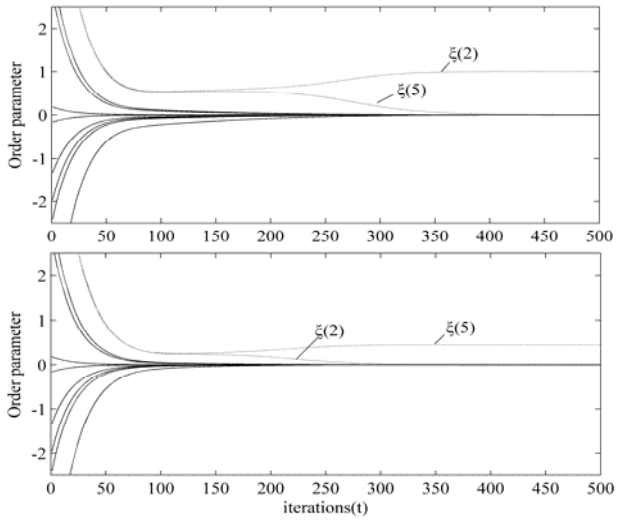

Figure 4. The comparison of identification using balanced attention parameters and self-adaptive attention parameters

Fig. 4 shows the identification process of a testing sample in distinguishing 10 authors based balanced and self-adaptive attention parameters. In the plot top, we can see that the final output pattern of SNN is the pattern which is represented by the order parameter with the biggest initial module value, as $\xi(2)$ has the biggest value in initial condition, after 500 iterations, the testing text is identified as

the second author's work, but the testing samples should be attributed to the fifth author, thus the method of balanced attention parameters causes the incorrect identification in SNN. In the plot bellow, we find that identification process is corrected after using self-adaptive attention parameters, although the initial module value of $\xi(2)$ is bigger than that of $\xi(5)$, because of the auto adjust of attention parameters, $\xi(5)$ gets the final victory in the competition among the different patterns. The process reflects the method of adjusting attention parameters by adaptive selection mechanism is effective for online writeprint identification.

5. Conclusion

In this paper, SNN model was developed to identify online writeprint. The proposed online writeprint identification system based on improved SNN achieved better recognition performance and robustness than the standard SNN. By using the dynamic method, the system obtained fairly strong self-learning ability as it can adaptively change the value of attention parameters by measuring similarity between the prototype pattern and the identified pattern. In our research, we found that the values of λ_k, B and C greatly affected the identification performance, but so far, there is no specific guidance to the values of parameters in SNN. Therefore, in the future, the values of λ_k, B and C will be studied for achieving the better performance the robustness in online writeprint identification.

6. Acknowledgments

This work was supported by National High-tech R&D Program of China (863 Program) (No.2008AA01Z131) and self-determined research funds of CCNU from the colleges' basic research and operation of MOE (No.CCNU09A02006). This study has been carried out in the framework of the 'Key Technology Research of Online Writeprint Identification' project, funded by Central China Normal University; we would like to thank the reviews for their comments which helped improve this paper improve this paper.

7. References

[1] Jiexun Li, Rong Zheng, Hsinchun Chen, "From fingerprint to writeprint", Communication of the ACM, ACM press, vol. 49, no. 4, pp.76-82, 2006.
[2] Burrows, J. F., "Word patterns and story shapes: The statistical analysis of narrative style", Literary and Linguistic Computing, vol. 2, pp.6-67, 1992.
[3] H. Haken, Synergetics, Springer-Verlag, 1977.
[4] Yudashkin and Alexander, " Topological Approach to the Pattern Classification in Neural Networks", IEEE Int'l Conf. on Neural Networks, pp.1484-1487, 1996.
[5] Peter Joerg, Freyer Richard, " Image Segmentation with Constraint Satisfaction Synergetic Potential Network", In Proceedings of SPIE -Int'l. Society for Optical Engineering, pp.384-393, 1995.
[6] T. Wagner, F. G. Boebel, " Testing Synergetic Algorithms with Industrial Classification Problems", Neural Networks, vol. 7, no. 8, pp.1313-1321, 1994.
[7] Liu Min-hang, Fu Yu-kun, "On Synergetic Theory and Application in Sports Curriculum Reform", In Proceedings of the 2009 First International Workshop on Education Technology and Computer Science, pp.557-562, 2009.
[8] Tong Zhao, Lilian H. Tang, Horace H. S. Ip, Feihu Qi, "On relevance feedback and similarity measure for image retrieval with synergetic neural nets", Neouralcomputing, Elsevier Science, vol. 51, pp. 105-124, 2003.
[9] Dongliang Hu, Feihu Qi, Jianfeng Liu, " Recognition of objects with skew distortion based on synergetics", Pattern Recognition Letters, vol. 20, no. 3, pp.255-265, 1999.

[10] Shouju Li, Li Wu, Fuzheng Qu, Wei Sun, "Parameter Estimation of Particle Flow Model for Soils Using Neural Networks", Journal of Convergence Information Technology, AICIT, vol. 5, no. 8, pp.29-35, 2010.
[11] Wagner, T., Boebel, F. G., "Testing synergetic algorithms with industrial classification problems". Neural Networks. vol. 7, no. 8, pp.1313-1321. 1994.
[12] H. Haken, Synergetic Computers and Cognition-Top-down Approach to Neural Nets, Springer-Verlag, 1991.
[13] Rong Zheng, Jiexun Li, Hsinchun Chen, Zan Huang, "A Framework for Authorship Identification of On line Messages: Writing-Style Features and Classification Techniques", Journal of the American Society for Information Science and Technology, Interscience, vol. 57, no. 3, pp.378-393, 2006.
[14] E. Bura, J. Yang, "Dimension estimation in sufficient dimension reduction: A unifying approach", Journal of Multivariate Analysis, ACM press, vol. 102, no. 1, pp.130-142, 2011.
[15] Shifei Ding, Yu Zhang, Li Xu, Jun Qian, "A Feature Selection Algorithm Based on Tolerant Granule", Journal of Convergence Information Technology, AICIT, vol. 6, no. 1, pp.191-195, 2011.
[16] Jair Cervantes, Xiaoou Li, Wen Yu, Kang Li, "Support vector machine classification for large data sets via minimum enclosing ball clustering", Neurocomputing, Elsevier Science, vol. 71, no. 4, pp.611-619, 2008.
[17] Siwar Zribi Boujelbene, Dorra Ben Ayed Mezghani, Noureddine Ellouze, "Improving SVM by Modifying Kernel Functions for Speaker Identification Task", International Journal of Digital Content Technology and its Applications, AICIT, vol. 4, no. 6, pp.100-105, 2010.
[18] Rong-Hua Li, Zhiping Luo, Guoqiang Han, "Pseudo-inverse Locality Preserving Projections", In Proceedings of the 2009 International Conference on Computational Intelligence and Security, pp.363-367, 2009.