Fuzzy ART for the Document Clustering By Using Evolutionary Computation

Shutan Hsieh [1], Ching-Long Su [2], Jeffrey Liaw[3]
[1]Department of Accounting, National Kaohsiung University of Applied Sciences
415 Chien Kung Road, Kaohsiung 807, Taiwan
shutan@cc.kuas.edu.tw
[2]Department of Information Management, Chang Jung Christian University
396 Chang Jung Rd., Sec.1,Kway Jen, Tainan 71101, Taiwan
clsu@mail.cjcu.edu.tw
[3]Corporate Planning Office, Uni-President Enterprise Corp., Taiwan
yihjuh@mail2000.com.tw

Abstract: - Many clustering techniques have been widely developed in order to retrieve, filter, and categorize documents available in the database or even on the Web. The issue to appropriately organize and store the information in terms of documents clustering becomes very crucial for the purpose of knowledge discovery and management. In this research, a hybrid intelligent approach has been proposed to automate the clustering process based on the characteristics of each document represented by the fuzzy concept networks. Through the proposed approach, the useful knowledge can be clustered and then utilized effectively and efficiently. In literature, artificial neural network have been widely applied for the document-clustering applications. However, the number of documents is huge so that it is hard to find the most appropriate ANN parameters in order to get the most appropriate clustering results. Traditionally, these parameters are adjusted manually by the way of trial and error so that it is time consuming and doesn't guarantee an optimum result. Therefore, a hybrid approach incorporating an evolutionary computation (EC) approach and a Fuzzy Adaptive Resonance Theory (Fuzzy-ART) neural network has been proposed to adjust the Fuzzy-ART parameters automatically so that the best results of the document clustering can be obtained. The proposed approach is tested by using ninety articles in three different fields. The experimental results show that the proposed hybrid approach could generate the most appropriate parameters of Fuzzy-ART for getting the most desired clusters as expected.

Key-Words: - Documents Clustering, Evolutionary Computation, Fuzzy ART, Knowledge Discovery.

1 Introduction

Nowadays, more and more information is hugely necessary for creating the useful knowledge to maintain the competitive advantage for a successful enterprise. For keeping the competitive advantage, it is necessary to explore the useful technologies in knowledge management. Although information can be obtained much easily via information technology, unfortunately the tremendous amount of information is far more than what people can absorb. However, how to get the "right" information becomes crucial. The methodologies to well organize, store and retrieve the desired knowledge accurately from tons of data become the key issue for assisting the successful knowledge dispersion and utilization in knowledge management.

Group technology (GT) has been proven to be an important tool in many engineering fields [Chung and Kusiak 1994][Kusiak and Chung 1991][Liao and Chen 1993] [Moon and Chi 1992]. Recently, the new advancements of computer technology and artificial intelligence offer good opportunities to apply more advanced clustering techniques to the GT problems. In various research fields, many successful artificial neural networks (ANNs) applications have been reported due to its superiority in terms of robustness to noise, quick response to numerous data population, and self-learning capability compare to traditional serial processing techniques. In literature, many researches demonstrated that ANNs can achieve better results than traditional methods in solving the GT problems [Lee and Chen 2001].

The aim of this paper is to develop the way to appropriately organize and store such information in the form of articles, from which the useful knowledge contained in article(s) can be obtained and utilized. The issue to appropriately organize and store the information in terms of documents clustering becomes very crucial for the purpose of knowledge discovery and management. A neural

network approach, Fuzzy Adaptive Resonance Theory (Fuzzy ART) [Carpenter and Grossberg 1991], has been applied for finding the relationships of interaction among the explicit and implicit information, which can be conducted into knowledge.

In this research, an efficient document-clustering algorithm has been proposed that uses the knowledge representation obtained by using fuzzy concept network [Chen and Horng 1999] for each document instead of using a huge proximity matrix as the clustering input. The methodology is mainly applying the Fuzzy ART neural network to perform an unsupervised clustering using the term frequency vector of each text document.

However, one of the biggest problems encountered in the application is the task of finding an optimal set of parameters for obtaining the most desired Fuzzy-ART's output. Selection of these parameters generally depends on the property of data population and affects nonlinearly to the result of the ANN. Therefore, it is very time consuming and hard to find an optimum set of parameters by means of the inefficient trial and error method. It may get worse to find optimum values using the existing serial processing techniques in the following situations: (1). if the number of ANN's parameters is greater than two, (2). if the optimum results are changing by nonlinear combination among parameters, and (3). if data population is too huge. Especially, it is more critical when the optimum parameter values need to be found for unsupervised ANN because there is too little information about data population.

Therefore, the purpose of this research is to apply an evolutionary computation approach to find the optimum parameters for the ANN with the unsupervised learning algorithms. If a set of optimum parameters can be automatically generated while the evolutionary computation is incorporated with the ANN, this hybrid approach is expected to be applied in more complicated applications with more accurate solutions since it significantly reduces time for trial and error and improves the reliability of the correct solutions [Chen 1997]. Since the use of evolutionary computation approach is able to automate the parameter-selection process for the ANN, it ultimately realizes that a real time use of ANN implementation is more reliable.

The paper is arranged as follows: following the basic clustering algorithm is briefly described; the general concept of fuzzy ART is described and the method for extracting the document pattern are described in Section 3 and Section 4 respectively; the evolutionary computation is illustrated in Section 5; the hybrid intelligent model and the clustering example with 90 articles are illustrated and discussed in Section 6, followed by conclusions in Section 7.

2 Clustering

A distance metric is needed to find the closest cluster and to determine if it is too far from the object to cluster. Basically, the clustering algorithm can be described as follows:

Step 1. Initially, no cluster prototype vectors have been clustered.

Step 2. Present and transform a new object (V), where $V = (v_1, v_2, \ldots, v_n)$, and analyze the new object.

Step 3. Find the closest cluster $C = (c_1, c_2, \ldots, c_n)$ to minimize the distance (d), if any exist, where $d(C,V) = \sqrt{\sum_i (w_i(c_i - v_i))^2}$.

Step 4. Check if the closest cluster is close enough.

If $d(C,V) > r$, or if there are no cluster prototype vectors yet, then create a new cluster, with prototype vector equal to V; goto step 2

Step 5. Update a matched cluster, let $C = (1-\lambda)\ C + \lambda\ V$; goto step 2

Many different approaches have been proposed in the literature based on knowledge or rules, fuzzy logic and artificial neural networks [Palmero et al. 1996]. However, the learning of cluster process stated as above is similar to the learning of processes of some neural networks. Above all, the learning process in Kohonen networks [Kohonen 1990] in case of fixed number of vectors, and in Adaptive Resonance Theory (ART) networks [Carpenter and Grossberg 1987] for a variable number of vectors. Moreover, the ART based architectures with the real-time property are able to keep learning during performance phase, thus providing a continuous adaptation of the system to the real one, i.e., the capacity for incremental learning. The above property is very significant for the uses of ART based neural networks to achieve the knowledge management purpose so that the Fuzzy ART has been applied as the tool for the clustering purpose.

But one of the difficulties to most of the artificial neural networks including Fuzzy ART and Kohonen networks is hard to decide the number of the clusters. For example, in the Step 4 of the above pseudo algorithm, the *r* value will decide the number of the clusters. But it will be difficult the most appropriate value for *r* so as to decide the most desired outcome of the clustering. For solving this difficulty, a hybrid algorithm combining the evolutionary computation with Fuzzy ART neural networks is proposed to solve the above difficulty.

3 Neural Networks Implementation

Carpenter and Grossberg (1991) introduced the Fuzzy ART which is a unsupervised learning neural network. This algorithm achieves a generalization to learning both analog and binary input patterns by replacing appearances of the intersection operator ($\cap$) in ART 1 by the fuzzy set theory MIN operator ($\wedge$). In Fuzzy ART, three parameters including choice parameter (α), learning parameter (β) and vigilance parameter (ρ), are to be adjusted to form the appropriate number of clusters. The influence of these parameters to Fuzzy ART is noted as follows: 1. when the value of either the learning or vigilance parameter increases, the number of the clusters increases, and 2. when the value of the choice parameter decreases, the number of the clusters increases. The vigilance mechanism helps to ensure that a minimum level of similarity within a cluster is maintained. The learning parameter β defines the degree to which the weight vector W_j learns characteristics of an input vector that is claimed by node *J*. Two choices for this neural network's learning are as follows: 1. fast learning mode, where β is always equal to 1, 2. fast-commit and slow-recode mode, where $\beta = 1$ for a category that is committed for the first time and $0 < \beta < 1$ for other times. As what Carpenter and Grossberg (1991) suggested, the second choice is used in this study for smoothing respondence.

The Fuzzy ART algorithm is illustrated as follows:

Step 1. Initialization

Connection weights: $w_{ji}(0) = 1$,
Choice parameter: $\alpha > 0$,
Learning rate: $\beta \in [0,1]$,
Vigilance parameter: $\rho \in [0,1]$

Step 2. Read new input *I*

Step 3. Compute choice function (Tj)

$$T_j = \frac{|I \wedge W_j|}{\alpha + |W_j|}$$

Step 4. Select best-matching examplar:

$$T_J = \max_j \{T_j\}$$

Step 5. Resonance test

If similarity $= \frac{|I \wedge W_J|}{|I|} \geq \rho$, go to learning Step 7

Else go to the next step (Step 6).

Step 6. Mismatch reset: Set TJ = -1 and go to Step 4.

Step 7. Update best-matching exemplar (learning law) $W_J^{(new)} = \beta(I \wedge W_J^{(old)}) + (1-\beta)W_J^{(old)}$

Step 8. Repeat: go to Step 2.

The flow chart of Fuzzy-ART clustering program has been described in Figure 1. The Fuzzy-ART neural network can perform well for the clustering purpose [Lee and Fischer 1999] [Lee and Chen 2001]. However, selection of the three parameters (a, b, and r) affect nonlinearly and sensitively to the clustering result of the Fuzzy-ART. Thus, it is necessary to have an approach to find the appropriate parameters of fuzzy-ART for obtaining the desired clustering. Evolutionary computation approach has been chosen to achieve this purpose. In this study, the role of the Fuzzy-ART is to cluster the documents, which is similar keyword attribute. It is noted that the keywords of the documents are used as the clustering attribute.

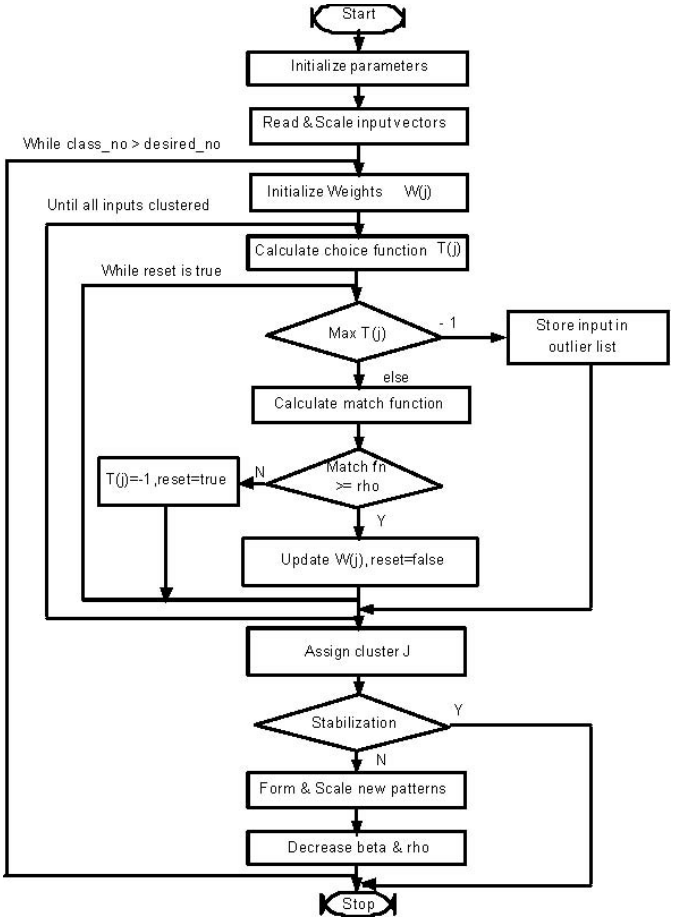

Figure 1. A flow chart of the fuzzy-ART clustering program (source: ref. [Lee and Fischer 1999])

4 Document Pattern Extracting

Since the content of a document is unstructured, the implicit contents are hard to be expressed based on the Boolean logic. One superior approach is called fuzzy concept network [Chen *et al.* 1999] [Hu *et al.* 2000] has been applied to extract the pattern vector from the document as the characteristics of the document. The details of the fuzzy concept network are described as follows.

4.1 Fuzzy Concept network

Notations:

tf_{ij} : the frequency of the j^{th} term in document i.

tf_{ijk}: the co-occruence of the j^{th} term and k^{th} term in the document i.

df_j: the frequency of the j^{th} term in a document.

df_{jk}: the frequency of j^{th} and k^{th} terms appear in a document simultaneously.

N: the number of documents in a cluster.

w_j: a weight for inverse document frequency.

$WF(T_k)$: the specificity of term T_k to a document is determined by the inverse document frequency.

The concept indicates the keywords of a document that can represent the characteristic of the document. The network is a matrix to illustrate the relations between two terms (term by term). The concept network is described as follows:

$$M = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1n} \\ f_{21} & f_{22} & \cdots & f_{2n} \\ \vdots & & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nn} \end{bmatrix}, \quad C_j \xrightarrow{f_{jk}} C_k,\ f_{jk} \in [0,1],$$

$$1 \leqq j \leqq n,\ 1 \leqq k \leqq n \tag{1}$$

The relevancies of any pairs of two terms can be represented by f_{jk} as follows [Hua et al. 2000]:

$$f_{jk} = relevancy(T_j, T_k) = \frac{\sum_{i=1}^{n} d_{ijk}}{\sum_{i=1}^{n} d_{ij}} \times Weighting\ Factor(T_k) \tag{2}$$

where,

$$d_{ijk} = tf_{ijk} \times \log_{10}\left(\frac{N}{df_{jk}} \times w_j\right) \tag{3}$$

$$d_{ij} = tf_{ij} \times \log_{10}\left(\frac{N}{df_j} \times w_j\right) \tag{4}$$

$$Weighting\ Factor(T_k) = \frac{\log_{10}\frac{N}{df_k}}{\log_{10} N} \tag{5}$$

4.2 Transitive closure matrix

If document i is relevant to document j and document j is relevant to document k, then the relevance of document i and k can be described by f_{ik}, which is described as Equation (6).

$$f_{ik} = \bigcup_{j=1,\ldots,n}(f_{ij} \cap f_{jk}) = Max(\underset{j=1,\ldots,n}{Min}(f_{ij}, f_{jk})), \tag{6}$$

where, $1 \leqq i \leqq n,\ 1 \leqq k \leqq n,\ \bigcup : Max\ ,\ \cap : Min$

The transitive closure matrix is as follows:

$$T = M^2 = M \otimes M$$

$$= \begin{bmatrix} \bigcup\limits_{j=1,\ldots,n}(f_{1j} \cap f_{j1}) & \bigcup\limits_{j=1,\ldots,n}(f_{1j} \cap f_{j2}) & \cdots & \bigcup\limits_{j=1,\ldots,n}(f_{1j} \cap f_{jn}) \\ \bigcup\limits_{j=1,\ldots,n}(f_{2j} \cap f_{j1}) & \bigcup\limits_{j=1,\ldots,n}(f_{2j} \cap f_{j2}) & \cdots & \bigcup\limits_{j=1,\ldots,n}(f_{2j} \cap f_{jn}) \\ \vdots & & \ddots & \vdots \\ \bigcup\limits_{j=1,\ldots,n}(f_{nj} \cap f_{j1}) & \bigcup\limits_{j=1,\ldots,n}(f_{nj} \cap f_{j2}) & \cdots & \bigcup\limits_{j=1,\ldots,n}(f_{nj} \cap f_{jn}) \end{bmatrix} \quad (7)$$

4.3 Document representation

After the three pre-processes including stop lists, stemming and calculating term frequency are achieved, the documents can be represented by an document descriptor matrix D as Equation (8).

$$D = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \vdots & & \ddots & \vdots \\ t_{m1} & t_{m2} & \cdots & t_{mn} \end{bmatrix},$$

$$t_{ij} \in [0,1],\ for\ 1 \leqq i \leqq m,\ 1 \leqq j \leqq n \quad (8)$$

While D is calculated in Equation (8), the matrix D* in Equation (9) is constructed using D and the transitivity closure matrix T of concept matrix M. In the matrix, all the relevancy values are calculated reflecting indirect relationships with other concepts.

$$D^* = D \otimes T$$

$$= \begin{bmatrix} \bigcup\limits_{j=1,\ldots,n}(t_{1j} \cap d_{j1}) & \bigcup\limits_{j=1,\ldots,n}(t_{1j} \cap d_{j2}) & \cdots & \bigcup\limits_{j=1,\ldots,n}(t_{1j} \cap d_{jn}) \\ \bigcup\limits_{j=1,\ldots,n}(t_{2j} \cap d_{j1}) & \bigcup\limits_{j=1,\ldots,n}(t_{2j} \cap d_{j2}) & \cdots & \bigcup\limits_{j=1,\ldots,n}(t_{2j} \cap d_{jn}) \\ \vdots & & \ddots & \vdots \\ \bigcup\limits_{j=1,\ldots,n}(t_{mj} \cap d_{j1}) & \bigcup\limits_{j=1,\ldots,n}(t_{mj} \cap d_{j2}) & \cdots & \bigcup\limits_{j=1,\ldots,n}(t_{mj} \cap d_{jn}) \end{bmatrix} \quad (9)$$

5 Evolutionary Computation

In this research, we applied a hybrid evolutionary optimization approach that combines genetic algorithm [Holland, 1975] and stochastic annealing algorithms (i.e., elitism strategy) to enhance the effectiveness of solution searching.

GAs are efficient search methods based on the principles of natural selection and population genetics in which random operators on a population of candidate solutions are employed to generate new points in the search space [Goldberg 1989]. For any GA, a so-called chromosome representation is needed to describe each individual in the population of interest. Each individual or chromosome contains a sequence of genes from a certain alphabet. Although the alphabet was limited to binary digits in Holland's original design (1975), other very useful problem-specific representations of an individual or chromosome for function optimization have also been proposed. GAs can search the solution space efficiently to find an optimal or near optimal solution by the use of evaluation and genetic operator functions to maintain the useful schemata of chromosome in the population, in which a schema with a higher fitness will have higher probability of survival in each generation and thereby a higher probability of generating offspring. Improved solutions are often found among the new offspring, since they have an inherently good schema, i.e., one that remains in the population over generations. This characteristic has been discussed in detail by Michalewicz (1994). The effectiveness of GAs depends on complementary crossover and mutation operators. The crossover operator determines the rate of convergence, while the mutation operator makes the GAs' search jump out of the local optimum, thus avoiding the premature convergence to a local optimum.

Based on the principles of natural selection and population genetics, GAs are efficient search methods in which random operators on a population of candidate solutions are employed to generate new points in the search space. This efficiency indicates the robustness of the search method that underlies the GA approach and the flexibility of the formulation itself [Goldberg, 1989]. Moreover, the more promising approaches such as evolutionary computations have been widely applied [Bhandarkar et al. 1994] [Chen and You, 2000] [Fogel 1994] [Ma 2000].

Compared with GA, these can show their superior performance in terms of faster convergence and solution quality [Chen and You 2000]. The EC approaches emphasize the behavioral link between parents and offspring, or between reproductive populations, rather than the genetic link [Fogel 1994]. An elitism strategy and adaptation of basic GA operator were found and applied to accelerate convergence successfully [Bhandarkar et al. 1994]. The elitism strategy ensure that the best individual(s) in the current generation always survive into the next generation thereby preventing a potential inadvertent loss of the best individual(s).

Instead of using the roulette wheel method in the selection process, the replacement between each parent and its offspring is decided by the possibility of an annealing function. It has been shown that

microcanonical annealing (MCA) converges to a global minimum with unit probability given a logarithmic annealing schedule [Bhandarkar and Zhang 1994]. It models a physical system whose total energy is always conserved. The basic algorithm of MCA has been illustrated by Bhandarkar et al. (1994). In this paper, the fitness function represented by the mean standard deviation of document-clustering is to be minimized (illustrated in Section 6.2); the deviations of the i^{th} parent and the i^{th} child are denoted by P_i *and* C_i respectively. If $P_i > C_i$ then C_i is accepted as the new solution. If $P_i \geq C_i$ then C_i is accepted as the new solution only if $E_k \geq (P_i - C_i)$. If $P_i \geq C_i$ and $E_k < (P_i - C_i)$, then the current solution P_i is retained. In the event that C_i is accepted as the new solution, the kinetic energy demon is updated $E_k^{n+1} = E_k^n + (C_i - P_i)$ to ensure the conservation of the total energy. The energy E_k is annealed in a manner similar to the temperature parameter T in simulated annealing. The energy provides an extra degree of freedom which helps the solution searching of MCA jump out from the local optima to the global optima.

The implementation of the hybrid EC algorithm in this research is described as follows:

Step 1. Generate an initial population *(G)* randomly and let E_k be a high value.

Step 2. Generate each new population in a mating pool from the current population by using a roulette-wheel selection:

for *i = 1* to *n*, where the *n* is the population size.

Do:

{

(1) Generate a pair of offspring from a pair of parents in the mating pool using the recombination operator and a neighborhood operator. (i.e., crossover and mutation)

(2) ***if*** $C_i \leq P_i$, add C_i to the next generation *G_new*.

else if $(Ci - Pi) \leq Ek$ while $Ci > Pi$,

add *Ci* to the next generation *G_new*

and let $Ek = Ek + Ci - Pi$.

otherwise the current solution *Pi* is retained in the next generation *G_new*.

}

Step 3. Let *G=G_new*; reduce the energy using annealing function *Ek=A(Ek)*

Step 4. Stop while the convergence criterion is met. Otherwise go to *Step 2*.

In our implementation, if *Ci < Pi*, instead of all *Ci* are added to the next generation *G_new,* only the randomly selected C*i* can be added to the next generation *G_new*.

Therefore, in this research, EC based approach is able to provide improved parameter-set relatively quickly to the Fuzzy-ART owing to its inherent genetic reproduction process compare to generic greedy algorithms. More specifically, it is suitable algorithm to find optimum parameter set for the Fuzzy-ART in which the three parameters effect nonlinearly to the result.

6 Hybrid EC-Fuzzy-Art System and The Numerical Results

In this study, an EC approach is used to provide an optimal set of input parameters for the Fuzzy-ART neural network. The optimal set of parameters is defined as a set of parameters for the Fuzzy-ART that results in the best set of clusters in which the sum of standard deviations among elements is to be minimized. In other words, for each set of parameters of Fuzzy-ART, the fitness function is the sum of deviations among the documents in the same cluster, i.e., the Fuzzy-ART can provide the information of deviations for each set of parameters.

A concept of applying to adjust the values of the Fuzzy-ART parameters (choice parameter α, learning parameter β, and vigilance parameter ρ) is shown in Figure 2. The above three parameters can be represented by a binary string.

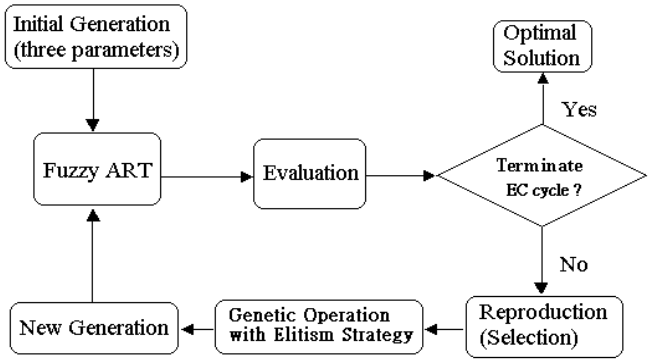

Fig. 2 Hybrid EC-Fuzzy-ART model

In each generation, every string (chromosome) in the current generation has to be evaluated. The

Fuzzy-ART can be looked as an evaluation function for evaluating every string. In the proposed hybrid approach, each set of parameters represented by a chromosome (binary string) in a generation of EC process can be evaluated by the Fuzzy-ART based on the sum of the degree of deviation from the part groups. The detail of fitness function is described above section.

In the reproduction process, an individual with a smaller fitness function value has a higher probability to be selected to propagate a new generation. However, not every offspring can be put in the new generation surely. Both the parent and offspring can be selected to form the new generation based on the elitism strategy that has been illustrated in previous section. The genetic cycle will not stop until the any stop criterion has been reached. Based on our limited experiments in this example, the number of maximum generation is set as 20 which is the only one stop rule in this experiment.

6.1 The representation mechanism

For any genetic algorithm based approach (for example the evolutionary computation approach), a so-called *chromosome representation* is needed to describe each individual in the population of interest. Each individual or chromosome contains a sequence of genes from a certain alphabet. Although the alphabet was limited to binary digits in Holland's original design (1975), other very useful problem-specific representations of an individual or chromosome for function optimization have also been proposed. In our implementation, the three parameters of Fuzzy-ART are represented by strings of binary digits, each string consisting of three substrings to represent the value of each parameter as described in Figure 2. Each binary substring can be decoded into a real number [Michalewicz 1992] to represent the value of each parameter.

6.2 Definition of fitness function

Fitness function is defined as the sum of the standard deviation among elements in each cluster through all clusters. This means that it can be used as a criterion for optimal clustering since it measures similarity levels among elements in a cluster. The smaller fitness values are, the more similar elements are in a cluster. For example, if there are five articles to be classified into two groups based on the similarity of the document-characteristics. First of all, it is supposed that each article should be characterized as a 4-number input vector as shown in Table 1. Based on the characteristics of the vectors, assume parts are classified by Fuzzy-ART in two groups as described in Table 1. The fitness function has been defined as follows:

$$Fitness = \sum_{i=1}^{m} \frac{\sum_{j=1}^{n} Std\_Dev_j}{n},$$

where m is the number of groups, n is the number of elements within a input vector.

Fitness = (.7071 + .7071 + .7071 +.7071)/4 + (1 + 1+ 1+ .5774)/4 = 1.6014

Table 1. Input vectors with the grouping standard deviation.

Group #	Document #	Input vector			
1st	Doc. 1	2	3	5	1
	Doc. 3	3	4	4	2
	Std_Dev.	0.7071	0.7071	0.7071	0.7071
2nd	Doc. 2	4	5	3	7
	Doc. 4	3	6	5	7
	Doc. 5	2	4	4	6
	Std_Dev.	1.0000	1.0000	1.0000	0.5774

** Std_Dev: Standard Deviation*

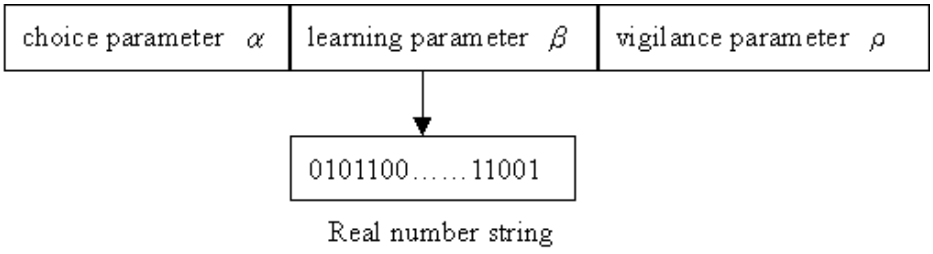

Fig. 2 The chromosome representation

6.3 Numerical Results

In order to illustrate the proposed model of our study and evaluating the performance of the hybrid intelligent approach, ninety articles are tested. These articles are obtained from three fields including Accounting, Information Management and Economics. The first group with thirty articles in the Accounting filed are denoted from no.1 to no.30, the second group with thirty articles in the Information Management field are denoted from no. 31 to no.60, and the other thirty articles in third group in Economics field are denoted from no.61 to no.90 respectively. Totally, these ninety articles are to be clustered by the proposed methodology without any field-predefined in advanced. Ideally, they are to be clustered into three groups.

All articles are represented by ninety vectors respectively as an input data are to be clustered. The characteristic vector of each article is extracted by using fuzzy concept network which has been described in Section 3.

The proposed hybrid evolutionary approach have been coded in MATLAB® on Pentium II 350 MHZ Personal Computer. In our experiments, the following values of evolutionary computations are given as 20 for the maximum number of generation, 30 for the size of population, 0.95 for the crossover rate, and 0.03 for the mutation rate. The determination of the parameters in evolutionary algorithm is a significant problem for the implementation. However, no formal methodology can be used to solve the problem because various value-combinations of the parameters result to different characteristics as well as different performance. Therefore, one should note that the best values for the parameters in the algorithm are case-dependent and based upon the experience from preliminary runs. Also, in this study, the parameters of the Fuzzy-ART, i.e., choice parameter α, learning parameter β, and vigilance parameter ρ, are limited to a 10-digit binary string. The individual lower and upper limits are α=[0.01, 0.999], β=[0.001, 0.999], and ρ=[0.01, 0.99].

By using the proposed hybrid intelligent approach, the most appropriate set of parameters that resulted in the most desired clustering could be found consistently and quickly as shown on Table 2. It is show that no any article is clustered in the wrong cluster, i.e., the clustering accurate rate is 100% in this experiment. Based on our limited experiments in this research, the number of generations of the proposed approach is always not more than ten generations to be convergent for finding the best set of parameters and generating the best clustering. It indicates that the proposed hybrid approach could generate the appropriate parameters for Fuzzy-ART consistently and efficiently to form the most appropriate part families.

Table 2. Experiment results.

Cluster #	No. of Article	Articles in the correct cluster	Accurate rate (%)
# 1	1 2 3 7 8 10 11 12 13 14 15 18 19 20 22 23 25 4 5 6 9 16 17 21 24 26 27 28 29 30	30 articles	100%(=30/30)
# 2	31 32 33 34 35 37 38 39 40 42 43 44 45 46 47 48 49 56 60 36 41 50 51 52 53 54 55 57 58 59	30 articles	100%(=30/30)
# 3	61 62 63 64 65 66 67 68 69 71 72 73 74 75 76 70 77 78 79 80 81 83 84 85 88 82 86 87 89 90	30articles	100%(=30/30)
The optimal parameters of Fuzzy ART：α=0.0387, β=0.4385, ρ=0.9593			

7 Conclusions

In this paper, a hybrid intelligent approach combining genetic algorithm, stochastic annealing algorithms (elitism strategy) and fuzzy ART has been proposed to achieve the automatic document clustering purpose based on the document characteristics extracted by using the fuzzy concept networks methodology. In the research, the evolutionary computation is to provide the most suitable parameters (i.e., choice parameter α, learning parameter β, and vigilance parameter ρ) to make the Fuzzy-ART generate the optimal part families in which the sum of standard deviation to be minimized. In this study, the 90 documents can

be properly clustered in the similar fields including Accounting, Information Management and Economics automatically instead of using the supervised learning approach, the proposed approach is able to cluster the documents with same field in the same group correctly. By using the propose approach, when new documents are added, the previous documents are not needed to be clustered again. This is the advantage more than other approaches in the literature. So, the document can be added and clustered in the one of original groups or even to generate the new group.

Although the fuzzy ART may be successfully implemented into the proposed system, the system is neither useful nor efficient while much time is needed to find the proper input parameters, improper parameters are selected, and there is no way to determine whether the current solution is the optimal or not. Furthermore, it becomes a more significant matter if a novice user who lacks experience for using the systems. To deal with the above difficulties, the hybrid intelligent approach has been proposed in this study. Based on our limited experience, it suggests that the proposed hybrid intelligent approach significantly improves the usability and the proper clustering reliability which are comparable to that of using trial and error for finding the most appropriate input parameters.

We do hope that this paper will interest other researcher to extend the proposed idea as well as to use ANN related approach more efficiently and effectively in the automatic document clustering purpose and more research fields to improve the necessary technologies in knowledge management.

References:

[1] Bhandarkar, S.M., Zhang, Y., and Potter, W.E. (1994), "An edge detection technique using genetic algorithm-based optimization," *Pattern Recognition*, 27(9), 1159-1180.

[2] Carpenter, G., and Grossberg, S., (1987), "A massively parallel architecture for a self-organizing neural patter recognition machine," *Computer Vision, Graphics, and Image Processing*, 37, pp. 54-115.

[3] Carpenter, G., and Grossberg, S. (1991), "Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system," *Neural Networks*, 4, pp. 759-771.

[4] Chen, Shyi-Ming and Horng, Yih-Jen (1999), "Fuzzy Query Processing for Document Retrieval Based on Extended Fuzzy Concept Networks," *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, 29(1), pp. 96-104.

[5] Chen, T.C. (1997), "A Hybrid Intelligent System for Process Modeling and Control Using a Neural Network and a Genetic Algorithm," *Ph.D. Thesis in the University of Iowa*, Iowa, USA.

[6] Chen, T. C. and You, P. S. (2000), "An efficient evolutionary computation approach for the vending machine inventory control problem," *Journal of the Chinese Institute of Industrial Engineers*, 17(4), pp. 451-457.

[7] Chung, Y. and A. Kusiak (1994), "Grouping parts with a neural network," *Journal of Manufacturing Systems*, **13(4)**, pp. 262-275.

[8] Fogel, D. D. (1994), "An Introduction to simulated evolutionary optimization," *IEEE Transactions on Neural Networks*, 5(1), pp. 3-14.

[9] Goldberg, D. D. (1989), *Genetic Algorithm Search, Optimizing and Machine Learning*, Addion-wesly, New York, Massachusetts.

[10] Holland, J.H. (1975), Adaptation in natural and artificial systems, *University of Michigan Press*, Ann Arbor

[11] Huh, Wonchang, Lee, Sangjin, Kim, Yeongho, and Kang, Suk-Ho (2000), "Automatic Classification of WWW Documents Using a Neural Network," *International Conference On Production Research – 2000, Bangkok, Thailand.*

[12] Kohonen, T. (1990), "The self-organizing map," *Proc. IEEE, 78*, pp. 1464-1480.

[13] Kusiak, A. and Chung, Y. (1991), "GT/ART: using neural networks to form machine cells," *Manufacturing Review*, **4(4)**, pp. 293-301.

[14] Lee, S.Y. and Fischer, G.W. (1999), "Group parts based on geometrical shapes and manufacturing attributes using a neural network," *Journal of Intelligent Manufacturing*, 10(2), pp. 199-209.

[15] Lee, S.Y. and Chen, T.C. (2001), "Using evolutionary computation approach to improve the performance of the fuzzy-art for grouping parts," *Journal of the Chinese Institute of Industrial Engineers*, 18(5), pp.55-62.

[16] Liao, T. W. and Chen, L. J. (1993), "An

evaluation of ART1 neural models for GT part family and machine cell forming," *Journal of Manufacturing Systems*, **12(4)**, pp. 282-290.

[17] Ma, J., Tian, P. and Zhan, D. M. (2000)g, "Global optimization by Darwin and Boltzmann mixed strategy," *Computers and Operations Research*, **27**, pp. 143-159.

[18] Merkl, Dieter (1998), "Text classification with self-organizing maps: some lessons learned," *Neurocomputing*, 21, pp. 61-77.

[19] Michalewicz, Zbigniew (1992), Genetic *Algoritms + Data Structures = Evolution Programs*, Springer-Verlag, New York.

[20] Moon, Y. B. and Chi, S. C. (1992), "Generalized part family formation using neural network techniques," *Journal of Manufacturing Systems*, **11(3)**, pp. 149-159

[21] Palmero, G.S., Izquierdo, J.C., Dimitriadis, Y., and Coronado, J.L. (1996), "Document understanding based on a neuro-fuzzy approach," *ProceedingsANN96*, London, UK, pp. 17-19.

[22] Tauritz, D.R., Kik, J.N.m and Sprinkhuizen-Kuyper, I.G. (2000), "Adaptive information filtering using evolutionary computation," *Information Sciences*, 122, pp. 121-140.