Fuzzy ART for the Document Clustering By Using Evolutionary Computation

Shutan Hsieh 1, Ching-Long Su 2, Jeffrey Liaw 3
1Department of Accounting, National Kaohsiung University of Applied Sciences
415 Chien Kung Road, Kaohsiung 807, Taiwan
shutan@cc.kuas.edu.tw

2Department of Information Management, Chang Jung Christian University
396 Chang Jung Rd., Sec.1,Kway Jen, Tainan 71101, Taiwan
clsu@mail.cjcu.edu.tw

3Corporate 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 \( V = (v_1, v_2, ..., v_i) \), and analyze the new object.

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

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 \( C = (I-\lambda) C + \lambda V \); 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 \((\land)\). In Fuzzy ART, three parameters including choice parameter \((\alpha)\), learning parameter \((\beta)\) and vigilance parameter \((\rho)\), 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 \(\beta\) 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 \(\beta\) 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 \((T_j)\)

\[
T_j = \frac{|I \land W_j|}{\alpha + |W_j|}
\]

Step 4. Select best-matching examplar:

\[T_j = \max_j \{T_j\}\]

Step 5. Resonance test

If similarity = \[
\frac{|I \land W_j|}{|I|} \geq \rho,
\]
go to learning Step 7

Else go to the next step (Step 6).

Step 6. Mismatch reset: Set \(T_J = -1\) and go to Step 4.

Step 7. Update best-matching exemplar (learning law) \[W_j^{\text{new}} = \beta(I \land W_j^{\text{old}}) + (1 - \beta)W_j^{\text{old}}\]

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 \((a, b, \text{and} r)\) 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.
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:


