Text Processing Simplified ARTMAP Neural Network

WORAPOJ KREESURADEJ and PUANGPAKA KUNASIT
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Ladkrabang, Bangkok 10520
THAILAND

Abstract: - This paper proposes text processing simplified ARTMAP neural network. The algorithm works directly on textual information without transforming to numerical value. The input layer of the neural network can directly receive a qualitative value without mapping the qualitative value into numerical value. Then, based on simplified fuzzy ARTMAP neural network and the concept of similarity measure for symbolic objects, the proposed neural network can assigns class labels to the objects correctly.

Key-Words: - Text mining, Text classification, Text categorization, Document classification, Document categorization, Simplified ARTMAP neural network and Neural networks.

1 Introduction

Several classification techniques for objects whose feature values are numerical values are well known. Several neural networks such as backpropagation neural networks, ARTMAP, fuzzy ARTMAP and simplified fuzzy ARTMAP neural network are proposed for classification. Recently, the classification problems are extends for document classification. To classify a document by using typical neural networks, a document has to be mapped onto a document representation that has quantitative features. The vector-space model is the most widely-used document representation [1]. Then, the typical neural networks or classifiers can be applied for classifying documents. However, the utilization of the vector-space model may led to a very high dimensional feature space. In addition, this feature space is generally not free from correlation.

Unlike conventional classifiers, the proposed algorithm works directly on textual information without mapping documents onto some numeric document representations. The proposed neural network, text processing simplified ARTMAP neural network, is based on the architecture of the simplified fuzzy ARTMAP neural network and the concepts of symbolic objects similarity measure [2],[3] and [4]. The experimental results show that the text processing simplified ARTMAP neural network can assigns class labels to documents correctly. This paper is introduced as following. In the 2nd section, we present similarity measure for symbolic objects. In the 3rd and 4th section, the text processing simplified ARTMAP neural network is introduced. In the 5th section, we present the concepts of document representation. In the 6th section, experimental results are presented. Finally, the 7th section is conclusions.

2 Similarity Measure

According to K.C. Gowda [3], the definition of similarity between two symbolic objects A and B are written as Cartesian product of A_k and B_k features as:

$$A = A_1 * A_2 * ... * A_d , \quad (1)$$

$$B = B_1 * B_2 * ... * B_d . \quad (2)$$

The similarity between two symbolic objects A and B is written as:

$$S(A,B) = S(A_1,B_1) + S(A_2,B_2) + ... + S(A_d,B_d) . \quad (3)$$

For the k^{th} feature, $S(A_k,B_k)$ is defined using the following three components:

1. the similarity component due to position is $S_p(A_k,B_k)$,
2. the similarity component due to span is $S_s(A_k,B_k)$,
3. the similarity component due to content is $S_c(A_k,B_k)$.

The similarity component due to "position" arises only when the feature type is quantitative. It indicates the relative position of two feature values on real line. Since this paper is only dealing with text data, the similarity component due to position is neglected. The similarity component due to span

indicates the relative sizes of the feature values without referring to common part between them. The similarity component due to content is a measure of common parts between two feature values [3],[4] and [5].
The similarity component due to span is defined as:

$$S_s(A_k,B_k)=\frac{(l_a+l_b)}{2.l_s}, \quad (4)$$

and the similarity component due to content defined as:

$$S_c(A_k,B_k)=\frac{inters}{l_s}, \quad (5)$$

where
l_a = length of A_k or number of elements in A_k,
l_b = length of B_k or number of elements in B_k,
$inters$ = length of the intersection of A_k and B_k,
l_s = span length of A_k and B_k or number of elements in $A_k \cup B_k$.
Therefore, the net similarity between A_k and B_k is

$$S(A_k,B_k)=S_s(A_k,B_k)+S_c(A_k,B_k). \quad (6)$$

3 The Text Processing Simplified ARTMAP Neural Network

The proposed neural network consists of three layers: input layer (F_1), output layer (F_2) and category layer (F_3). Each neural node in the F_1 layer is connected to each neural node in the F_2 layer by two weighted pathways. The F_1 neural node, i.e., X_i, is connected to F_2 neural node by bottom-up weights, b_{ij}. Similarly, the F_2 neural node, i.e., Y_j, is connected to the F_1 neural node by top-down weights, t_{ji}. The F_2 layer is connected to some neural nodes in the F_3 layer if a category is associated with the winning output node. Unlike conventional neural networks, the F_1 layer can receive qualitative values. The bottom-up weight, b_{ij} and top-down weight, t_{ji} contain qualitative values and degrees of association of the qualitative values is defined as [5]:

$$b_{ij}=\{(A_{1ij},e_{1ij}),(A_{2ij},e_{2ij}),(A_{3ij},e_{3ij}),...,(A_{pij},e_{pij})\}. \quad (7)$$

A_{pij} is the p^{th} qualitative value of the weight and e_{pij} is the degree of association of this qualitative value to the weight. The value of e_{pij} has values between 0 to 1. e_{pij} equals to zero, i.e., e_{pij} = 0, if A_{pij} is not a part of the weight. While e_{pij} equals to one, i.e., e_{pij} = 1, if the qualitative value has strong association with the weight.

$$t_{ji}=\{(B_{1ji},e_{1ji}),(B_{2ji},e_{2ji}),(B_{3ji},e_{3ji}),...,(B_{pji},e_{pji})\}. \quad (8)$$

B_{pji} is the p^{th} qualitative value of the weight and e_{pji} is the degree of association of this qualitative value to the weight. The value of e_{pji} is between 0 to 1. e_{pij} equals to zero, i.e., e_{pji} = 0, if B_{pji} is not a part of the weight. While e_{pij} equals to one, i.e., e_{pji} = 1, if the qualitative value has strong association with the weight. The architecture of the proposed neural networks is shown in figure 1.

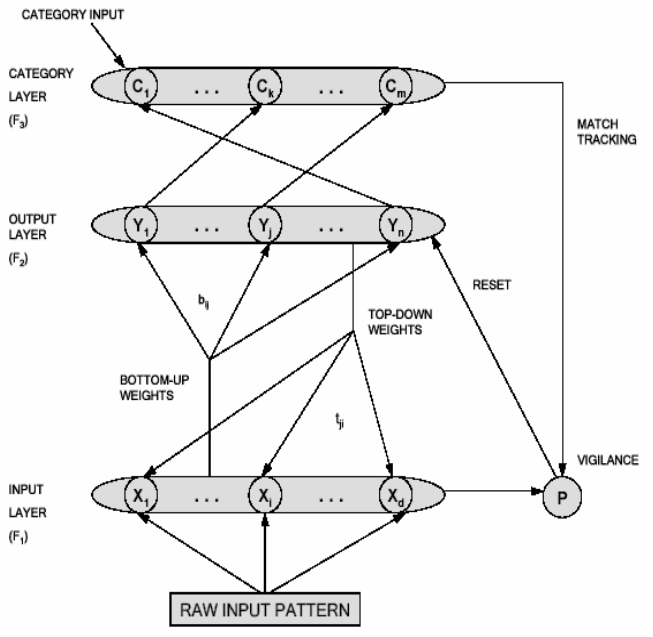

Fig. 1 The architecture of text processing simplified ARTMAP neural network

4 Learning Algorithm

The learning algorithm of The text processing simplified ARTMAP neural network algorithm is based on that of the simplified Fuzzy ARTMAP neural network and the concept of symbolic objects similarity measure as described in the prior section. The algorithm is summarized as below:

0th Step: Set parameters values.

1st Step: While stopping condition is false, do the 2nd -14th step.

2nd Step: For each input vector.

$X = (X_1, X_2, ..., X_d)$, do the 3rd -13th step. (9)

3rd Step: Present the input pattern to the F_1 layer and present the category input to the F_3 layer simultaneously.

4th Step: If output node = 0, then

-Create an output node on the F_2 layer to encode this training pattern.

-Initialize bottom-up weights, b_{ij} and top-down weights, t_{ji} by choosing from the training data set arbitrarily and setting degrees for each element of the weight.

-Set the category of the new output node to that of the input pattern.

-Update the bottom-up weight and the top-down weight, do the 13th step.

5th Step: If output node ≠ 0, do the 6th -13th step.

6th Step: Set the vigilance to the baseline vigilance.

7th Step: Evaluate the activation function between X and bottom-up weight for all output nodes.

$$Y_j = (\sum_{i=1}^{d} S(X_i, b_{ij})) / (\sum_{i=1}^{d} S(b_{ij}, b_{ij})), \quad (10)$$

where d is number of feature values.

8th Step: Select the j^{th} highest activated output node.

$$Y_J = \max\{Y_j : j = 1..n\}. \quad (11)$$

9th Step: Evaluate the match function between X_i and the top-down weight that connects to the winning output node, i.e., J.

$$V = (\sum_{i=1}^{d} S(X_i, t_{Ji})) / (\sum_{i=1}^{d} S(X_i, X_i)), \quad (12)$$

where d is number of feature values.

10th Step: Test mismatch reset

$$V \geq \rho. \quad (13)$$

-If $V < \rho$, suppress activation of the current winning output node, i.e., J, do the 8th step until $V \geq \rho$.

-If $V \geq \rho$, do the 11th step.

11th Step: Test category mismatch between the network category and the category input.

-If the network category and the category input are the same category, do the 13th step.

-If the network category and the category input are not the same category, set vigilance to the match value of the winning output node plus a small value. Then, suppress activation of the current winning output node J and do the 8th -11th step.

12th Step: If all output nodes can not match the category input, then

-Create an output node on the F_2 layer to encode this training pattern and

-Initialize bottom-up weights, b_{ij} and top-down weights, t_{ji}.

13th Step: Update the winner bottom-up weights and top-down weights as following:

$$b_{iJ}^{(new)} = b_{iJ}^{(old)} \cap X$$

$$e_{niJ}^{(new)} = \begin{cases} f(e_{niJ}^{(old)} + \beta) & if \quad A_{nij} \in b_{iJ} \cap X \\ f(e_{niJ}^{(old)} - \beta) & if \quad A_{nij} \notin b_{iJ} \cap X \\ \beta_0 & otherwise \end{cases}, \quad (14)$$

$$t_{Ji}^{(new)} = t_{Ji}^{(old)} \cap X$$

$$e_{nJi}^{(new)} = \begin{cases} f(e_{nJi}^{(old)} + \beta) & if \quad B_{nji} \in t_{Ji} \cap X \\ f(e_{nJi}^{(old)} - \beta) & if \quad B_{nji} \notin t_{Ji} \cap X \\ \beta_0 & otherwise \end{cases}, \quad (15)$$

where $f(.)$ is defined as below:

$$f(x) = \begin{cases} x & if \quad 0 \leq x \leq 1 \\ 0 & if \quad x < 0 \\ 1 & if \quad x > 1 \end{cases}. \quad (16)$$

14th Step: Continue with the 2nd -14th step until the stopping condition is true.

5 Document Representation

A document, i.e., D, can be written as Cartesian product of specific values of its features D_k's as [6]:

$$D = D_1 * D_2 * ... * D_d. \quad (17)$$

Unlike a vector-space model, the features have qualitative values which are a set of words that describe the features. As an example, a document can be written as Cartesian product of Title feature and Keyword feature as:

$$Doc = Title * Keyword. \quad (18)$$

Where the values of the Title feature are a set of words that describe the title of the document and the

values of the Keyword feature are a set of keywords of the document.

6 Experimental Results

In this section, the experiment results of the text processing simplified ARTMAP neural network is presented. We use a training dataset of 200 documents that consist of 3 class. Each document in the training dataset can be represented by title feature and keyword feature according to equation (18).

Each feature of a document is qualitative value. Here, Some English alphabets are used to represent value of each feature. The values of title feature and the value of keyword feature for each class are shown in figure 2 and figure 3 respectively.

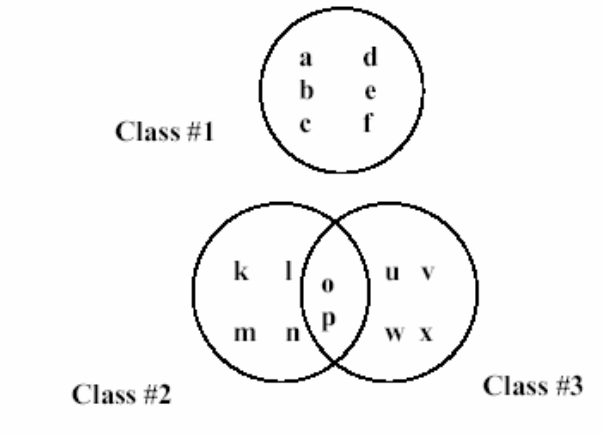

Fig. 2 Venn Diagram of Title feature

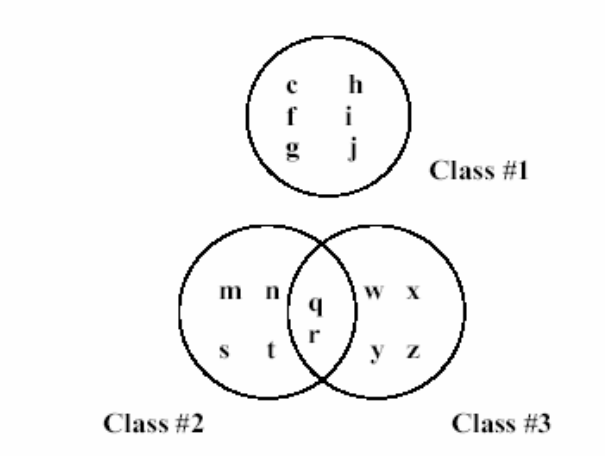

Fig. 3 Venn Diagram of Keyword feature

Some documents of the training dataset are shown in table 1.

Table 1 Some documents of the training dataset

No.	Data Values		Category Input
	Titles	Keywords	
1	l,n,o,p	m,n,q,r,t	2
2	a,d,e,f	c,f,g,h	1
3	k,m,n	m,r,s,t	2
4	o,p,u,w	q,y,z	3
5	v,w,x	r,w,x,y,z	3
6	b,c,d	g,h,i,j	1

The training data consist of 67 documents from the class number 1, 66 documents from the class number 2 and 67 documents from the class number 3.

Then, we use 5 testing datasets of 1000 documents, as shown in table 2, to evaluate the performance of the proposed neural network.

Table 2 The testing dataset

Dataset No.	Members		
	Class 1	Class 2	Class 3
1	330	336	334
2	328	338	334
3	332	333	335
4	332	336	332
5	330	332	338

The evaluate measurement of the proposed neural network is the accuracy rate (r) defined as:

$$r = [(\sum_{i=1}^{c} doc_i) / n] * 100, \qquad (19)$$

where c is a number of data that are the same class,
n is a number of all data.

The results from table 3 show 100% accuracy rate for all data set. The experimental results show good performance of the proposed neural network.

Table 3 The experimental results

Data Set No.	Members			Experimental Result			Accuracy Rate
	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3	
1	330	336	334	330	336	334	100%
2	328	338	334	328	338	334	100%
3	332	333	335	332	333	335	100%
4	332	336	332	332	336	332	100%
5	330	332	338	330	332	338	100%

7 Conclusion

In this paper, we applied the proposed neural network for text classification. The text processing simplified ARTMAP neural network can receive qualitative value directly without transformation. According to the experimental results, the proposed neural network has good performance in classifying documents. In the future, we will conduct some experimentations on the Reuter-21578 news article [7].

References:

[1] G. Salton, *Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer*, Addison Wesley Publishing Company, New York, 1989

[2] T. Kasuba, Simplified Fuzzy ARTMAP, *AI Expert*, Vol.8, 1993, pp. 18-25.

[3] K.C. Gowda and E. Diday, Symbolic Clustering Using a New Similarity Measure, *IEEE Trans. on Systems, Man, and Cybernetics*, Vol.22, No.2, 1992, pp. 368-378.

[4] T.V. Ravi and K.C. Gowda, Clustering of Symbolic Objects Using Gravitational Approach, *IEEE Trans. on Systems, Man, and Cybernetics*, Vol.29, No.6, 1999, pp. 888-894.

[5] W. Kreesuradej, N. Chantasut and W. Kruaklai, Clustering Text Data Using Text ART Neural Network, *The WSEAS Trans. on Systems*, Vol.3, I.1, 2004, pp. 200-205.

[6] Y. El-Sonbaty and M.A. Ismail, Fuzzy Clustering for Symbolic Data, *IEEE Trans. on Fuzzy Systems*, Vol.6, No.2, 1998, pp. 195-204.

[7] D.D. Lewis, Reuters-21578 Text Categorization Test Collection Distribution 1.0, available on http://www.daviddlewis.com/resources/testcollections/reuters21578