A Kind of Text Classification Method Based on Fuzzy Vector Space Model and Neural Networks

JunHui PAN
Department of Computer and Information Technology
NorthEast Petroleum University
Daqing, Heilongjiang, China
pjhdqpi@126.com

Hui LI
Department of Computer and Information Technology
NorthEast Petroleum University
Daqing, Heilongjiang, China
lihui_dqpi@163.com

***Abstract*-A kind of text classification method based on fuzzy vector space model and neural networks is proposed in the paper according to the problems that a text can be belongs to many types during the text classification. Fuzzy theory is adopted in the method to look the occurring position of feature items in text on as the important degree (membership) reflecting text subject, and fully considered the position information while the features are extracted, thus the fuzzy feature vectors are constructed, as a result, the text classification is close to the manual classification method. The established networks are constituted of input layer, hidden layer and output layer, the input layer completes the inputs of classification samples, hidden layer extracts the implicit pattern features of input samples, the output layer is used to output the classification results. Finally the effectiveness of this method is proved by some documents of Wan Fang data in experimental section. (Abstract)**

Keywords-Text classification; fuzzy vector space; neural networks; fuzzy feature vector; feature extracted; membership

I. INTRODUCTION

Text classification is a very important task in data mining [1]. The traditional text classification research has abundant research results and extensive application practice, but in reality, a part of text can not accurately classify into a category, a text may belong to two or more given types if it relies on artificial judgment. How to deal with that a text belongs to many types, we can consider the subordinate relations between the text with all types, if can get the text's membership degree to all types, then the problems will be readily solved.

The paper presents a solution based on fuzzy vector space model and BP neural network. Firstly select a training sample set composed with some kinds of mode which is uniformly coverage for domain knowledge, and then construct fuzzy feature vector extracting according to fuzzy feature, submit to the neural network, then train to get fuzzy mapping relation between the generic features of text with class model, at last test text classification according to generated fuzzy mapping relation.

II. FUZZY FEATURE SELECTION

Feature selection chooses the best and most representative feature subsets as the class features according to the characteristic evaluation results; it can reduce the dimension of feature space, so as to reduce the computational complexity and improve the precision of the system and prevent over fitting [2]. The feature extracted methods based on VSM all are the statistics methods, firstly give mark for feature item by using different methods, and then choose the items with higher score composed feature vector space. The common feature extraction method has document frequency, information gain, and mutual information, etc, but these methods generally do not consider features' position in a document, for example, the same feature item appears in title, keywords, abstract and the text will be treated as equally, thus affecting the classification accuracy. The paper is improved by using the fuzzy set theory, look the feature items' appear position in the document on as reflecting the important degree of document theme, namely membership [3], and calculate the frequency of feature item according to the membership degree.

The appearing frequency of a feature in a document can calculate according to the following principle:

(1) If a feature item in the text has been chosen as the keyword should give membership value as 1;

(2) If a feature item appears in the title and abstract, should give higher membership;

(3) If feature item appears in some "key words" of the text, namely those who contain such as "the key lies in……", "aims to.....", "main purpose is.....", should be given greater membership

(4) If a feature item appears in the preface and the conclusion section, should be given a certain membership;

(5) If a feature item appears at the beginning and ending of section, should be given a certain membership;

(6)If a feature item has a higher appearance frequency in text, should increase its membership with frequency increasing.

(7) If a feature item is on the various statuses at the same time, the membership degree should be added in sum.

(8) If the synonyms and homogony or escape words of a feature item appears, should appear in frequency of statistics as one or part of feature item based on the size of semantic relations.

(9) Structure feature vector should also consider the shows degree of feature item. The shows degree can be denoted by the ratio between document total with document numbers containing the feature item. The

Published by Atlantis Press, Paris, France.

feature item with low shows degree will inhibit the accuracy of classification. For feature item with high shows degree should be appropriately increased its document frequency; for feature item with low shows degree should be reduced its document frequency.

According to the above principle, the construction of the fuzzy feature vector can proceed as follows:

step 1: calculate the document frequency of each feature item in feature item sets $\{T_1, T_2, \cdots, T_N\}$ for P document respectively according to (1) - (8);

step 2 : Construct the feature vector $\{ f_T(T_{p1}), f_T(T_{p2}), \cdots, f_T(T_{pN}) \}$; $(p = 1,2,\cdots,P)$.

$$f_T(T_{pk}) = V_k \lg(\frac{N}{N_k} + 0.5) \quad ,$$

$(p = 1,2,\cdots,P; k = 1,2,\cdots,N)$ of P documents according to (9)

Where, V_k denotes the frequency that feature item T_k appears in document p, N denotes the documents numbers of all training text, N_k denotes the document numbers containing feature item T_k.

step 3: Normalize to the above feature vector, can get fuzzy feature vector $\tilde{\boldsymbol{T}}_{\boldsymbol{p}} = \{T_{p1}, T_{p2}, \cdots, T_{pN}\}$ of P documents.

It is pointed out that it must firstly cut text into a sequence of words before extracting the feature of text.

III. NEURAL NETWORK MODEL CLASSIFICATION

A. BP Neural network model

The BP neural network belongs to the feed-forward neural network type [4], the network has three basic layer, namely input layer, hidden layer and output layer. Each layer contains several nodes. The node of input layer is $(x_1, x_2, \cdots, x_n)$ respectively corresponding to the input mode feature vector; the node of output layer is $(y_1, y_2, \cdots, y_m)$ corresponding to the mode output vector; each connection between layers all has a adjusted weight value, it is a coefficient calculated continuously according to the training data, and decides the influence from a input vector to output vector, w_{ij} is the connection weight between the i th input and the neurons j. The mapping relationship between feed-forward neural network input and output can be expressed as formula (1):

$$y = g(\sum_{j=1}^{m} v_j f(\sum_{i=1}^{n} w_{ij} x_i - \theta_j) - \theta) \qquad (1)$$

In the above formula, f is excitation function of hidden layer, θ_j is the j th neurons' excitation threshold of hidden layer; θ is the excitation threshold of output neurons, g is excitation function of output neurons.

By the neural network knowledge, arbitrary function can be approximated by three layer neural network with arbitrary precision, and text classifier essentially is a function mapping process, so we can imagine that construct classifier by adjusting the weights of neural network's each edge, thus approximate text classification mapping function [5]

B. BP Neural network classification algorithm

Put the fuzzy feature vector set of document for $\{X^1, X^2, \cdots, X^P\}$, $\boldsymbol{X^k} = (x_1^k, x_2^k, \cdots, x_n^k)$; The object set is $\{D^1, D^2, \cdots, D^P\}$, $D^k = (d_1^k, d_2^k, \cdots, d_m^k)$, $(k = 1,2,\cdots,P)$; Based on the known classification model, adaptively adjust the neural network parameters with gradient descent learning algorithm, the network error function is:

$$E = \frac{1}{2} \sum_{p=1}^{P} \left\| d^p - y^p \right\|^2 \qquad (2)$$

The purpose of the training is to make $E \le \varepsilon$。In formula (2), y^p is the network output vector of input X^p.

w_{ij} is the stay adjustable parameters，The learning rule of network weights is：

$$\begin{aligned} w_{ij}(t+1) &= w_{ij}(t) - \eta \frac{\partial E}{\partial w_{ij}} + \alpha \Delta w_{ij}(t) \\ \Delta w_{ij}(t) &= w_{ij}(t) - w_{ij}(t-1) \end{aligned} \qquad (3)$$

In formula (3), η is the learning speed, α is the inertial coefficient, t is iterations.

$$\frac{\partial E}{\partial v_{kj}} = (y_j - d_j) z_k \qquad (4)$$

$$\frac{\partial E}{\partial w_{ij}} = \sum_{j'=1}^{p} (y_{j'} - d_{j'}) \sum_{k=1}^{K} b_{kj'} (\frac{\delta(j, s(k,i))}{X} - \frac{Y}{X^2}) \frac{\partial \mu_{A_{ij}}(x_i)}{\partial w_{ij}} \qquad (5)$$

where，

$$X = \sum_{j=1}^{n_i} \mu_{A_{ij}}(x_i) \quad Y = \mu_{A_{i,s(k,i)}}(x_i) \qquad (6)$$

$$\frac{\partial \mu_{A_{ij}}(x_i)}{\partial w_{ij}} = \frac{2(x_i - w_{ij})}{\sigma_{ij}^2}\mu_{A_{ij}}(x_i),$$

$$\delta(j, s(k,i)) = \begin{cases} 1, & j = s(k,i) \\ 0, & j \neq s(k,i) \end{cases}. \quad (7)$$

The description of algorithm is described as follows:

step 1：Construct fuzzy feature vector according to fuzzy feature extraction method;

step 2：Determine the clustering center according to the k-means algorithms;

step 3：Initialize the hidden layer weights and threshold value; Give error precision ε; set accumulated iterations t=0; Maximum iterating times M;

step 4：Compute the error function E according to formula (2), if $E < \varepsilon$ or $t > M$; go to step 6;

step 5：Amend weights and threshold value according to formula (3)-(7); $t + 1 \to t$; go to step 4;

step 6：Output learning results; end.

IV. The Experiment and Result Analysis

The paper use part of the documents based on WanFang database as test sample source, download 800 documents as test sample library, including politics, economy, military, law, education, sports, entertainment, science and technology, life, computers and other 10 topics, each topic include 80 documents. Integrate all documents' feature; extract 82 keywords to compose feature item set. Construct fuzzy feature vector of sample according to the foregoing method. Choose 660 articles as the training set automatically classify and train for the network, and then test the classification situation of rest 140 documents using the trained network. The experimental results are as follows: for the training set, the accuracy of seven classifications can reach above 90% in 10 theme class; the accuracy of the rest three theme class minimum is 86.7%, and the average is 93.6%; For the test set, the accuracy of eight classification can reach more than 80%, the rest two theme class minimum is 78.2%, and the average is 86.8%. This method achieves good results by contrasting with traditional classification methods, and at the same time, this method has stronger generalization promotion ability, is recommended as a practical text classification method.

V. Conclusions

Aim to the traditional VSM model that it has the insufficiency in text characteristic expression, consider the importance when each characteristics reflecting text theme, use the fuzzy set theory knowledge to enhance the network's adaptability for complex classification problem, construct the fuzzy VSM model based on text characteristic, and propose a kind of automatic text classification method in this foundation based on neural network.. The experimental results show that this method achieves a better classification effect. Due to the research of the automatic text categorization problem is a complex problem, therefore, it is very necessary to continue researching many aspects, such as perfecting classification model, improving the learning algorithm and weight evaluation.

Acknowledgments

The authors would like to thank the anonymous reviewers for their critical and constructive comments and suggestions. Many thanks also to go out to National Natural Science Foundation of China (Grant No. 61170132).

References

[1] KeYun Hu, FengZhan Tian, HouKuan Huang. Data Mining Theory and Application. Beijing Jiao tong university press, 2008,200-203

[2]YANGYiming,LiuXin.Are-examinationoftextcategorizationmethods[EBOL].http:citeseer.nj.nec.comyang99reexamination.html,1999.

[3] Peizhuang Wang. The Fuzzy Set Theory and Application. Shanghai Science Press, 1983, 56-59

[4]Robert Hecht-Nielsen.Theory of the Back Propagation Neural Network[J].Proceeding of IJCNN,1989,1(1):593-603.

[5]Gang Liu, SiQuan Hu, ZhiHua Fan. An Application of Neural Network in Text Categorization [J]. Computer engineering and application, 2003, 39(36):73-76