Applied Soft Computing 21 (2014) 210–220

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Taking advantage of improved resource allocating network and latent semantic feature selection approach for automated text categorization

Wei Song*, Jiu Zhen Liang, Xiao Liang He, Peng Chen

School of Internet of Things Engineering, Jiangnan University, Engineering Research Center of Internet of Things Applied Technology, Ministry of Education, Wuxi, Jiangsu 214122, PR China

ARTICLE INFO

Article history:
Received 19 February 2013
Received in revised form 17 January 2014
Accepted 11 March 2014
Available online 1 April 2014

Keywords:
Resource allocating network
Novelty criteria
Root mean square
Least mean square
Extended Kalman filter
Semantic feature selection
Neural network
Text categorization

ABSTRACT

In this study we propose an improved learning algorithm based on resource allocating network (RAN) for text categorization. RAN is a promising neural network of single hidden layer structure based on radial basis function. We firstly use the means clustering-based method to determine the initial centers in the hidden layer. Such method can effectively overcome the limitation of local-optimal of clustering algorithms. Subsequently, in order to improve the novelty criteria of RAN, we propose a root mean square (RMS) sliding window method which can reduce the underlying influence of undesirable noise data. Through the further research on the learning process of RAN, we divide the learning process of RAN into a preliminary study phase and a subsequent study phase. The former phase initializes the preliminary structure of RAN and decreases the complexity of network, while the latter phase refines its learning ability and improves the classification accuracy. Such a compact network structure decreases the computational complexity and maintains the higher convergence rate. Moreover, a latent semantic feature selection method is utilized to organize documents. This method reduces the input scale of network, and reveals the latent semantics between features. Extensive experiments are conducted on two benchmark datasets, and the results demonstrate the superiority of our algorithm in comparison with state of the art text categorization algorithms.

1. Introduction

Digital text is, at present, a leading non-constructed information resource in web. Text classification (TC) is a crucial and well-proven instrument for organizing these large volumes of digital text information which are widespread and increasing continuously. TC is the task of automatically sorting a set of documents into categories from a predefined set [1]. Automated TC is attractive because it frees organizations from the need of manually handling document bases, which can be too expensive, or simply not feasible given the time constraints of the application or the number of documents involved. So far automated TC has become a research hot spot and put forward a series of related applications, e.g. web classification, personalized news, query recommendation, genre identification, spam filtering, and topic spotting etc. A variety of machine learning approaches have been applied to text categorization, including support vector machine [2,3], k-nearest neighbor [4], neural network [5], Bayes model [6], decision tree [7], etc. Although such methods have been extensively researched, yet the present automated text classifiers are still with fault and the effectiveness needs improvement. Thus, text categorization is still a major research field. Since neural network is one of the most powerful tools utilized in the field of pattern recognition, in this study, we employ neural networks as a classifier. Meanwhile, due to the inherent limitations of neural network, we propose a refined algorithm for neural network which will further improve the performance of the classifying system.

Artificial neural network is a self-learning model that can implement different algorithms extensively utilized in the field of pattern recognition [8–12]. Although many neural networks have been proposed, automated text categorization is still a major research spot because the effectiveness of current automated text classifiers is not without fault and still needs improvement. The back propagation neural network (BPNN) is a kind of basic supervised network. BPNN has the problems of slow training speed and the likelihood of becoming trapped in a local minimum, which makes it difficult to use in practical applications, especially when the scale of the network is large [13]. More specifically, in the beginning of its training steps, the learning process proceeds very quickly in each epoch and can make rapid progress. However, it slows down in the later stages

* Corresponding author.
E-mail address: songwei@jiangnan.edu.cn (W. Song).

http://dx.doi.org/10.1016/j.asoc.2014.03.002
1568-4946/© 2014 Elsevier B.V. All rights reserved.

and becomes so called inertia. Meanwhile, it is easy to enter local minima. In comparison with such complicated BPNN, Radial basis function neural network (RBFNN), as a type of feed forward neural network, has aroused scholars' wide concern for its simple structure and robust global situation approaching property [14]. The key to establish a successful RBFNN is to ensure the proper cell number in the hidden layer of the network [15]. It has known that the redundancy or lack of hidden joint of the network will result in a direct influence on the decision-making ability of RBFNN. That is to say, too small architecture of network may cause the problem of under-fitting, while on the other hand, too large architecture of network may lead to over-fitting to data. The frequently utilized learning method usually applies the way of dynamic modifying hidden joint to tackle with these problems and fulfill the requirement of the appropriate network structure. The most notable method is the resource allocating network (RAN) learning method put forward by Platt [16]. It is a promising neural network of single hidden layer based on radial basis function (RBF). RAN can dynamically manipulate the number of the hidden layer units by judging the novelty criteria. However, the novelty criteria are sensitive to the initialized data, which would easily cause the growth of the training time for network. Many extensions and modifications of the RAN have been developed during the past years (see Section 2.2). In general, the basic goals of the modifications have been proposed to raise learning ability and make training faster.

In this study we propose an improved RAN learning algorithm as a document classifier. We firstly use the means clustering-based method to determine the initial centers in the hidden layer. Subsequently, to reduce the underlying influence of initial noise data, we propose a root mean square (RMS) sliding window method to improve the novelty criteria of RAN. Moreover, we divide its learning process into a preliminary learning phase and a subsequent study phase. The former phase decreases the complexity of network and initializes the preliminary learning structure of RAN, while the latter phase refines its learning ability and improves the classification accuracy of the network. Such a compact structure decreases the computational complexity of network and maintains a higher convergence rate, which effectively improve the classification performance of RAN.

In the research literature of information retrieval (IR) and data mining (DM), vector space model (VSM) is a classical method frequently implemented to express document. Its premise hypothesis is based on the principle of independence between text features. However, it has the challenges of high dimensionality, sparse space and semantic concern. In order to perform text categorization, in this paper we make use of a method of latent semantic feature selection to reduce the large number of dimensions and create a latent semantic space to further enhance the effect of text classification. The derived indexing features in latent semantics space, rather than independent words, greatly reduce the dimensionality of input space and reveal the actual semantics between terms. Therefore, the main contributions of this paper are three folds. (1) Proposing an improved RAN algorithm as a document classifier. (2) Utilizing the latent semantic feature selection method for document representation. (3) Taking advantage of the refined RAN and the latent semantic feature selection method as a document classifying system, and successfully applying it to text categorization.

The rest of this paper is organized as follows. Section 2 introduces the related work of this study. In this section, the basic concepts of RAN are firstly described, and then the main issues of RAN encountered are extensively discussed. Based on the analysis of Section 2, we propose the improved RAN algorithm as an efficient text document classifier in Section 3. Section 4 describes how to generate the semantic features for our system which can improve the categorization performance. The experimental results and analysis are given in Section 5. The conclusions are given in Section 6.

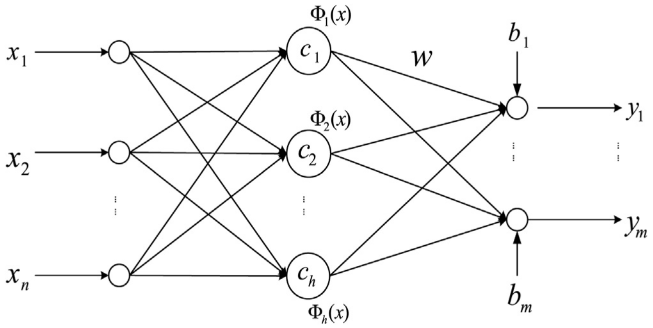

Fig. 1. The three-layer topology of RAN neural network.

2. Related work

2.1. Resource allocating network (RAN) learning algorithm

RAN is a supervised learning technique infused in sequential learning strategy. In RAN, there is an input layer, an output layer, and a single hidden layer based on radial basis function. RAN dynamically adjusts the number of hidden layer nodes or the existing network parameters through judging whether the training samples satisfy its novelty criteria. The topology of RAN is shown as Fig. 1. During the step of training, an input pattern of n dimensions vector is assigned to the input layer of the network. Based on the given input pattern, the network will compute the output of m dimensions vector in the output layer. That is, the whole network attempts to a mapping from n dimensional input space to m dimensional output space. In this three-layer structure network, the input layer, the hidden layer and the output layer are $X=(x_1, x_2, \ldots, x_n)$, $C=(c_1, c_2, \ldots, c_h)$, and $Y=(y_1, y_2, \ldots, y_m)$, respectively. $b=(b_1, b_2, \ldots, b_m)$ is the polarization item of output layer. The neuron of the hidden layer uses Gaussian function, and the output of hidden layer neuron is linearly weighted for the output layer which is shown as

$$f_j(x) = \sum_{i=0}^{h} w_{ij}\Phi_i(x) + b_j \quad (j = 1, 2, \ldots, m) \tag{1}$$

where h is the number of hidden layer neuron, and m is the number of output layer neuron. x is the sample of input, w_{ij} is the connecting weight between the i_{th} neuron of the hidden layer to the jth neuron of the output layer. $\Phi_i(x)$ is the Gaussian function of hidden layer which is shown as

$$\Phi_i(x) = \exp\left(-\frac{\left\|x - c_i\right\|}{\sigma_i}\right) \tag{2}$$

where c_i and σ_i are the neuron center and the center width of the ith neuron, respectively.

Once the RAN learning algorithm is well loaded, it faces a RBF network without hidden layer neuron. RAN initializes the network parameters through the first couple of input sample (x_0,y_0), and then it judges the novelty of each couple of training data. The standard RAN adjusts the number of hidden layer nodes in terms of judging its novelty criteria. The number of hidden nodes will increase if it satisfies novelty criteria, otherwise it will adjust the parameters of the current network by means of LMS algorithm (including the neuron centers of hidden layer and the network weights).