Neurocomputing 149 (2015) 1125–1134

Contents lists available at ScienceDirect

Neurocomputing

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

Application of a staged learning-based resource allocation network to automatic text categorization

Wei Song [a,b,*], Peng Chen [a,b], Soon Cheol Park [c]

[a] School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China
[b] Engineering Research Center of Internet of Things Applied Technology, Ministry of Education, China
[c] Department of Electronics and Information Engineering, Chonbuk National University, Jeonju, Jeonbuk 561756, Republic of Korea

ARTICLE INFO

Article history:
Received 11 December 2013
Received in revised form
4 April 2014
Accepted 10 July 2014
Communicated by Y. Chang
Available online 18 July 2014

Keywords:
Resource allocation network
Neural network
Staged learning algorithm
Text categorization
Novelty criteria

ABSTRACT

In this paper, we propose a novel learning classifier which utilizes a staged learning-based resource allocation network (SLRAN) for text categorization. In the light of its learning progress, SLRAN is divided into a preliminary learning phase and a refined learning phase. In the former phase, to reduce the sensitivity corresponding to input data an agglomerate hierarchical k-means method is utilized to create the initial structure of hidden layer. Subsequently, a novelty criterion is put forward to dynamically regulate the hidden layer centers. In the latter phase a least square method is used to enhance the convergence rate of network and further improve its ability for classification. Such staged learning-based approach builds a compact structure which decreases the computational complexity of network and boosts its learning capability. In order to implement SLRAN to text categorization, we utilize a semantic similarity approach which reduces the input scales of neural network and reveals the latent semantics between text features. The benchmark Reuter and 20-newsgroup datasets are tested in our experiments and the extensive experimental results reveal that the dynamic learning process of SLRAN improves its classifying performance in comparison with conventional classifiers, e.g. RAN, BP, RBF neural networks and SVM.

1. Introduction

With the rapid development of Internet technology, a large quantity of online documents and information are growing exponentially. The demand of rapidly and accurately finding out the useful information from such a large dataset has become a challenge for modern information retrieval (IR) technologies. Text categorization (TC) is a crucial and well-proven instrument for organizing large volumes of textual information. As a key technique in IR field, TC has been extensively researched and witnessed in recent decades. Meanwhile, TC has become a hot spot and puts forward a series of related applications, including web classification, query recommendation, spam filtering, topic spotting etc.

In recent years, an increasing number of approaches based on intelligent agent and machine learning, e.g. support vector machine (SVM) [1], decision trees [2,3], K-nearest neighbor (KNN) [4,5], bayes model [6–8], neural network [9,10] etc, have been applied to text categorization. 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 artificial neural network is still one of the most powerful tools utilized in the field of pattern recognition [11], we employ it as a classifier.

As a kind of basic supervised network, back propagation (BP) neural network suffers the fault of slow training rate and high tendency to trap into local minimum. On the contrary, without slow learning rate, the relatively simple mechanism of radial basis function (RBF) neural network [12–14] displays the robust property of global situation approaching. It has been known that the key to build a successful RBF neural network is to insure a proper number of units in its hidden layer [15]. More specifically, the lack of hidden layer nodes always results in a negative influence on its ability to decision-making. Whereas the redundant hidden layer nodes bring about a result of high computing [16–18]. 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 [19,20]. Although more and more learning methods have studied to regulate the hidden nodes to satisfy the demand of the suitable structure for RBF network, the most remarkable approach is resource allocation network (RAN) learning method put forward by Platt [21]. Platt made a significant contribution through the development of the

* Corresponding author at: School of IOT Engineering, Jiangnan University, Lihu Avenue, Wuxi, Jiangsu Province 214122, China.
E-mail address: songwei@jiangnan.edu.cn (W. Song).

http://dx.doi.org/10.1016/j.neucom.2014.07.017
0925-2312/

algorithm which regulates the hidden nodes according to the so called novelty criteria. In other words, 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, and lead to the reduction of the employment effect [22]. Meanwhile, in RAN the least mean-square (LMS) algorithm applied to update its learning parameters usually makes the network suffer from the drawback of lower convergence rate [23,24]. To tackle with these problems, in this paper we propose a staged learning-based resource allocation network (SLRAN) which divides its learning process into a preliminary learning phase and a refined learning phase. In the former phase, to reduce its sensitivity corresponding to the initialized data, an agglomerate hierarchical k-means method is utilized to construct the structure of hidden layer. Subsequently, a novelty criterion is put forward to dynamically add or prune hidden layer centers, and a compact structure is created. That is, the former phase reduces the complexity of the network and builds the initial structure of RAN. Yet in the latter phase a least square method is used to enhance the convergence rate of network and further refine its learning ability. Therefore, SLRAN builds a compact structure which decreases the computational complexity of network and boosts its learning ability for classification.

The rest of this paper is organized as follows. Section 2 introduces the basic concepts of the RAN. Section 3 proposes the algorithm of SLRAN as an efficient text classifier and describes its details. The steps to generate the latent semantic feature of text documents, which helps enhance the text categorization performance, are depicted in Section 4. Experimental results and analysis are illustrated in Section 5. Conclusions are discussed in Section 6.

2. Resource allocation network (RAN)

RAN is a promising and sequential learning algorithm based on RBF neural network. The architecture of RAN includes three layers, i.e. an input layer, an output layer and a single hidden layer. The topology of the RAN is shown in Fig. 1. During the training process of RAN, a sample of n dimensional input vector is given to the input layer, and based on the assigned input pattern, RAN will compute the output of m dimensional vector in the output layer. That is to say, the aim of the RAN network is to define an approach to map from the input space of n dimensions to output space of m dimensions. Eventually, the network calculates the output vectors that match the desired output vectors.

In the structure of RAN, the input layer, the hidden layer and the output layer are $x=(x_1,x_2,...,x_n)$, $c=(c_1,c_2,...,c_h)$ and $y=(y_1, y_2,...,y_m)$ respectively, $b=(b_1,b_2,...,b_m)$ is the offset item of the output layer, where n, h and m are the respective number of units in these three layers. The units of the hidden layer take advantage of Gaussian function as its activation function which implements a locally tuned unit, and the Gaussian function of hidden layer is defined as

$$\Phi_i(x)=\exp\left(-\frac{\|x-c_i\|^2}{\sigma_i^2}\right) \tag{1}$$

Where c_i and σ_i are the *ith* node center and the width of this center respectively. While the output of hidden layer node is linearly weighted for the output layer, the function for output layer is given by

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

Where m and h are the respective number of the nodes in output layer and hidden layer, x is a input sample, w_{ij} is the connecting weight between the ith hidden layer node and the jth output layer node.

At the beginning of the training stage, there is no hidden neuron in the network. RAN initializes the parameters of neural network in terms of the first couple of input training sample. Subsequently, if the training sample satisfies its novelty criteria, it will be added into the network as a new neuron center in hidden layer, otherwise LMS is used to update the parameters of current network including hidden layer centers and the connection weights between hidden layer and output layer. However, the novelty criteria of RAN are sensitive to the initialized data, which causes the growth of the training time. Meanwhile, LMS usually makes the network suffer from the drawback of lower convergence rate.

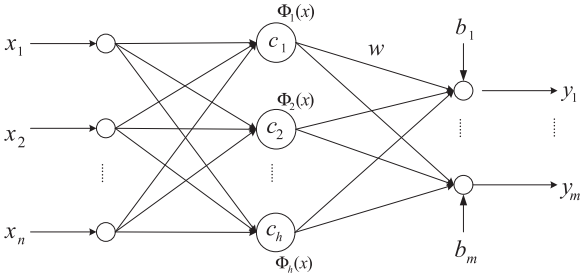

Fig. 1. The three layers topology of RAN neural network.

3. Staged learning-based resource allocation network (SLRAN)

To handle the above-mentioned problems of RAN, in this paper we propose a staged learning-based resource allocation network (SLRAN) which divides the learning process into two phases. In order to reduce its sensitivity corresponding to the initialized data, in the preliminary learning phase of SLRAN an agglomerate hierarchical k-means method is utilized to construct the structure of hidden layer. Subsequently, a novelty criterion is put forward to dynamically add or prune hidden layer centers. That is, this phase reduces the complexity of the network and builds a compact structure of RAN.

3.1. Determination of initial hidden layers centers

We apply an agglomerate hierarchical k-means algorithm to initialize the structure of hidden layer, i.e. the centers of hidden layer and the widths of the clusters for each center. For a given initial documental dataset D, after the clustering process, the generated cluster centers are defined as $C=(c_1,c_2,c_3,...,c_k)$, where k is the number of the clusters. The k-means algorithm helps obtain the hidden layer center c_i and the cluster width σ_i. The algorithm process is shown as follows:

Step 1: Through random sampling of m times, which ensure that the data would not be distorted after random sampling dataset and can reflect the feature of natural distribution for data. So the primitive dataset is divided into m parts, and the size of each part is n/m, n is the total number of the texts. That is, we get m subsets that can be expressed as $S=(s_1,s_2,...,s_m)$. Clustering analysis is executed for every subset s_i using k-means algorithm. In this way, we obtain a group of $k'(k'>k)$ clustering centers where k is the predefined number of categories. That is, we take such a step possible to guarantee that the generated centers can represent all clusters and avoid initial object around the isolated point. In our method k' is empirically defined as $2\times k$. Thus, the appropriate k' can ensure the uniform distribution of centers in each sample as possible and lead to a good sampling performance. In comparison with the effect of k', m is a secondary coefficient. We set it several