ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

Web Document Clustering based on a Hierarchical Self-Organizing Model

E.J. Palomo, E. Domínguez, R.M. Luque and J. Muñoz *

E.T.S.I. Informatica, University of Malaga - Department of Computer Science
Campus Teatinos s/n, 29071, Malaga - Spain

Abstract. In this work, a hierarchical self-organizing model based on the GHSOM is presented in order to cluster web contents. The GHSOM is an artificial neural network that has been widely used for data clustering. The hierarchical architecture of the GHSOM is more flexible than a single SOM since it is adapted to input data, mirroring inherent hierarchical relations among them. The adaptation process of the GHSOM architecture is controlled by two parameters. However, these parameters have to be established in advance and this task is not always easy. In this paper, a one parameter hierarchical self-organizing model is proposed. This model has been evaluated by using the 'BankSearch' benchmark dataset. Experimental results show the good performance of this approach.

1 Introduction

Nowadays, main search engines function by matching given keywords with a list of web documents that contain them. However, keyword matching is not enough to satisfy the document research necessities on the World Wide Web. In order to exploit its full potential, the World Wide Web should be labeled and categorized into different subsets accordingly. Unfortunately, the huge amount of web documents added to the World Wide Web every day, over 1.5 million documents according to [1], makes unfeasible manual classification.

One solution is to automatically cluster and categorize web documents through document clustering and content mining methods. There are two main ways to perform this categorization [2]: document classification and document clustering. Document classification uses a set of documents that are previously classified, where new documents are classified into the most similar classes. Typical supervised learning methods are backpropagation, neural networks, naïve Bayesian and support vector machines. Document clustering uses a set of unclassified documents to extract important relations among them and organize them in similar groups. These groups represent similar documents according to a similarity measure, where documents belonging to one group are more similar than documents belonging to different groups. These methods are especially useful when we have no information about the input data.

Document clustering has been faced by many techniques, including neural networks, but above all by self-organizing maps. The self-organizing map (SOM) has been widely used as a tool for knowledge discovery, data mining, detection

*This work is partially supported by the Spanish Ministry of Innovation and Science under contract TIN-07362.

ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

of inherent structures in high-dimensional data and mapping these data into a two-dimensional representation space [3]. This neural network has been successfully applied in multiple areas since the mapping retains the relationship among input data and preserves their topology. On the other hand, SOMs have some difficulties. First, the network architecture has to be established in advance. Second, hierarchical relations among input data are difficult to detect, so that understanding of the data is limited.

The growing hierarchical SOM (GHSOM) [4] was proposed to solve both limitations. The neural architecture is separated into layers, where each layer is composed of different single growing SOMs [5]. After training the growing SOM, each neuron of the map is analyzed to see whether they represent their mapped data at a specific level of granularity. Those neurons that represent too heterogeneous input data are expanded to form a new map at a subsequent layer. Growing and expansion in a GHSOM are controlled by two parameters: $\tau 1$ and $\tau 2$, respectively. These parameters have to be defined prior to training. Although these parameters provide flexibility to choose the size of the neural network, it remains far from trivial to determine and combine the two parameters that provide satisfying results.

In this paper, a new GHSOM model that has just one parameter to control the growing and expansion of the architecture is proposed. This parameter keeps providing the flexibility to choose the size of the network and at the same time makes easy its election. The effectiveness of this approach has been evaluated by clustering web documents from the 'BankSearch' benchmark dataset [6]. The remainder of this paper is organized as follows. The new GHSOM model proposed is described in Section 2. In Section 3, some experimental results about web document mining are presented by using the 'BanckSearch' benchmark dataset. Section 4 concludes this paper.

2 Hierarchical Self-Organizing Model

The starting point for our training process is to compute the quantization error at layer 0 as given in (1), where w_0 is the mean of the all input data I. The qe_0 measures the dissimilarity of all input data and it is used for the expansion process of the neurons.

$$qe_0 = \sum_{x_j \in I} \|w_0 - x_j\| \tag{1}$$

The quality of the adaptation process to input data is measured by the quantization error of a neuron (qe). The qe is a measure of the similarity of data mapped onto each neuron, where the higher is the qe, the higher is the heterogeneity of the data cluster. Let (i, j) be the position of a neuron in a map of NxM neurons. The quantization error of a neuron (i, j) is defined as follows

$$qe_{ij} = \sum_{x_j \in C_{ij}} \|w_{ij} - x_j\| \tag{2}$$

where C_{ij} is the set of patterns mapped onto the neuron (i, j), x_j is the jth input pattern from C_{ij}, and w_{ij} is the weight vector of the neuron (i, j).

In order to consistently combine the growing and expansion criteria, we have to decide in what situations are better to grow a map rather than expand their neurons. The difference between them lies in the use of the data mapped into the map, i.e. after growing the map is trained again with all its input data, whereas after expanding, new maps are just trained with data mapped into their respective parent neurons. Then, growing or expansion are decided depending on the proximity of the error neurons in the map, which are the neurons that no satisfy the condition given in (3). This way, the qe of a neuron (i, j) must be smaller than a fraction τ of the quantization error of its parent neuron u in the upper layer, where $0 < \tau < 1$.

$$qe_{i,j} < \tau \cdot qe_u \tag{3}$$

In the GHSOM, the growing of a map is done by inserting a row or a column of neurons between two neurons, the neuron with the highest quantization error and its most dissimilar neighbor. Here, once the error neurons are computed, if in a row or column there are more error neurons than non-error neurons, a row or a column of neurons is inserted and the map grows. To decide what insert (a row or a column) and where, the row quantization error rqe and the column quantization error cqe must be computed for each of the N rows and the M columns of the map as given in (4). Then, the error row er, i.e. the row with the highest sum of their quantization errors, and the error column ec, i.e. the column with the highest sum of their quantization errors are chosen following the expressions in (5). If the rqe of the er is bigger than the cqe of the ec, a row of neurons is inserted between the er and its neighbor row with the highest rqe. Likewise, if the cqe of the ec is bigger than the rqe of the er, a column of neurons is inserted between the ec and its neighbor column with the highest cqe (see Fig. 1). The weight vectors of the new neurons are initialized as the mean of their respective neighbors.

$$rqe_r = \sum_{j=1}^{M} qe_{rj} \qquad cqe_c = \sum_{i=1}^{N} qe_{ic} \tag{4}$$

$$er = \arg\max_i \{rqe_i\} \qquad ec = \arg\max_j \{cqe_j\} \tag{5}$$

On the other hand, if the error neurons of each row and column are less than the non-error neurons in their respective rows and columns, the error neurons will be expanded in a new map in the next level of the hierarchy. When a new map is created, a coherent initialization of the weight vectors of the neurons of the new map is used as proposed in [7]. This initialization provides a global orientation of the individual maps in the various layers of the hierarchy. Thus, the weight vectors of neurons mirror the orientation of the weight vectors of the neighbor neurons of its parent. The initialization proposed computes the mean of the parent and its neighbors in their respective directions. New maps created from expanded neurons are trained as single SOMs.

ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

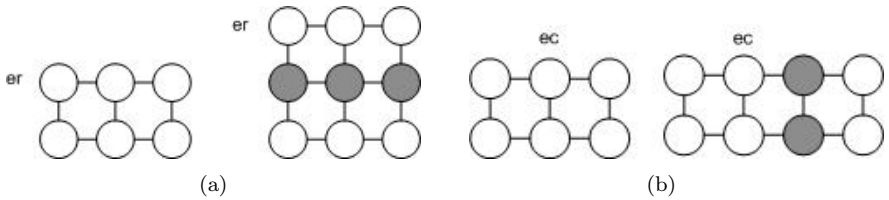

Fig. 1: An example of neurons insertion: (a) A row or (b) a column of neurons (shaded gray) is inserted between the error row *er* or the error column *ec* and its neighbor with the highest quantization error in the row or column.

3 Experimental Results

The proposed hierarchical self-organizing model has been used to perform a web document clustering. One problem in the web mining area is that each research paper uses its own datasets, so comparison between papers is difficult. For this reason, the 'BanckSearch' benchmark dataset has been chosen to test our model, which is a web document dataset proposed in [6] as a standard dataset against which different techniques can be benchmarked and assessed in comparison to each other.

The dataset was selected from the Open Directory Project and Yahoo! Categories, which have already been manually categorized. It consists of $11,000$ documents that correspond with 11 categories ($1,000$ documents per category). Each category has an associated theme, namely "Banking and Finance", "Programming Languages", "Science" and "Sport". The 11 categories are the following: 'Commercial Banks', 'Building Societies', 'Insurance Agencies', 'Java', 'C/C++', 'Visual Basic', 'Astronomy', 'Biology', 'Soccer', 'Motor Sport' and 'Other Sports'.

Each document consists of several features, where each feature represents the frequency of a word in the document that appeared at least once. From these words, stop-words listed in our stop-word list were removed to constitute the features. Our stop-word list is the same as used in [8], which is formed by 319 words. Once all the documents of the dataset were thus processed, a master word list that contains every word in the combined dataset, associated with its overall frequency is created. Then the master list is cut down to contain only the top h% of most frequently occurring words, where h was varied between experiments. Finally, a feature vector v_i was created for each document i, such that the jth element in v_i was w_{ji}/s_i, where w_{ji} is the number of occurrences in document i of the jth most frequent word in the combined dataset, and s_i is the total number of words in document i.

For our experiments, we chose the datasets A&I ('Commercial Banks' and 'Soccer') and B&C ('Building Societies' and 'Insurance Agencies') for the purpose of comparing, where the first dataset has different associated themes whereas the second one has associated the same theme. The training was done by using these datasets with four h values: 0.5, 1, 1.5 and 2, and the τ parameter was set

to 0.1 and 0.2. Results achieved by our model are given in Table 1. The accuracy is the percentage of documents that were correctly categorized and the category of a cluster is the category of the larger number of documents in that cluster. The F-measure is the weighted harmonic mean of precision and recall, where the higher the F-measure, the better the clustering due to the higher accuracy of the resulting clusters mapping to the original categories. It is shown that the results are better for $\tau = 0.1$, which generates bigger neural architectures in terms of neurons than those obtained after training for $\tau = 0.2$.

		GHSOM 1 parameter	
		$\tau = 0.1$	$\tau = 0.2$
Set	h(%)	Accuracy(%)/F-measure	Accuracy(%)/F-measure
A&I	0.5	94.3/0.9440	85.95/0.8735
	1	94.35/0.9447	92.3/0.9265
	1.5	94/0.9414	88.85/0.8982
	2	94.55/0.9470	89.85/0.9068
B&C	0.5	93.55/0.9373	88.15/0.8913
	1	94.8/0.9485	88.7/0.8946
	1.5	93.7/0.9386	89.95/0.9052
	2	94.5/0.9458	89/0.8954

Table 1: Results for the proposed model with $\tau = 0.1$ and $\tau = 0.2$.

		GHSOM		K-means
		$\tau 1 = 0.1$ $\tau 2 = 0.3$	$\tau 1 = 0.2$ $\tau 2 = 0.3$	
Set	h(%)	Ac.(%)/F-m.	Ac.(%)/F-m.	Accuracy(%)
A&I	0.5	72.43/0.7459	75.93/0.7616	92.8
	1	69.75/0.8204	79.33/0.7366	94.2
	1.5	62.05/0.456	65.85/0.4135	94.3
	2	71.66/0.7244	81.0345/0.7571	94.4
B&C	0.5	87.88/0.8885	68.3081/0.6903	90.8
	1	89.07/0.8889	68.70/0.6918	90.5
	1.5	91.47/0.9105	68.15/0.6826	90.1
	2	92.7/0.9210	67.69/0.6771	90.6

Table 2: Results for the GHSOM and the k-means algorithm.

Moreover, our results have been compared with those achieved by the GHSOM and the k-means algorithm for the same datasets used in this work. These results are shown in Table 2. The achieved results by using the k-means algorithm with $k = 2$ were extracted from [6], where only accuracy was provided. Note that the accuracies achieved by our model for both datasets when $\tau = 0.1$ are better than those achieved by the k-means clustering. For the training of the GHSOM, we also used 0.1 and 0.2 as values for the parameter $\tau 1$, which control

the depth/shallowness of the resulting hierarchical GHSOM. For $\tau 2$, we chose 0.3 to control the expansion of the architecture since with a smaller value the GHSOM returns an extremely deep hierarchy. GHSOM results show that both accuracy and F-measure are worse than the achieved by our model.

4 Conclusions

A new hierarchical self-organizing model for web document clustering is proposed in this paper. Our model is based on the GHSOM, however the growth and expansion of the GHSOM is controlled by two parameters, which have to be defined in advanced. The proposed model uses just one parameter to control both growing and expansion, making easy the configuration prior to training and obtaining a neural network that reflects the inherent hierarchical relations according to input data.

Furthermore, web document clustering has been addressed in this work. For this purpose, our model has been trained with the 'BankResearch' benchmark dataset, which was proposed as a standard dataset with a variety of properties suitable for a wide range of clustering and related experiments. By applying our hierarchical model, we overcame the results achieved in [6] with the k-means algorithm and also the achieved with the GHSOM, using the same datasets in both cases. Moreover, the obtained neural architectures were automatically determined during the unsupervised learning process, mirroring the inherent hierarchical relations among the input documents and providing understanding of the data structure.

References

[1] J.M. Pierre. Practical issues for automated categorization of web sites. In anonymous, editor, *Electronic Proc. ECDL 2000 Workshop on the Semantic Web*, 2000.

[2] R.T. Freeman and H.J. Yin. Web content management by self-organization. *IEEE Transactions on Neural Networks*, 16(5):1256–1268, 2005.

[3] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

[4] A. Rauber, D. Merkl, and M. Dittenbach. The growing hierarchical self-organizing map: Exploratory analysis of high-dimensional data. *IEEE Transactions on Neural Networks*, 13(6):1331–1341, 2002.

[5] D. Alahakoon, S.K. Halgamuge, and B. Srinivasan. Dynamic self-organizing maps with controlled growth for knowledge discovery. *IEEE Transactions on Neural Networks*, 11:601–614, 2000.

[6] M.P. Sinka and D.W. Corne. The banksearch web document dataset: investigating unsupervised clustering and category similarity. *Journal of Network and Computer Applications*, 28(2):129– 146, 2005. Computational Intelligence on the Internet.

[7] M. Dittenbach, A. Rauber, and D. Merkl. Recent advances with the growing hierarchical self-organizing map. In *3rd Workshop on Self-Organising Maps (WSOM)*, pages 140–145, 2001.

[8] C.J. Van Rijsbergen. *Information Retrieval*. Butterworths, 2nd ed. edition, 1979.