Supervised Learning for Automatic Classification of Documents using Self-Organizing Maps

Dina Goren-Bar, Tsvi Kuflik, Dror Lev
Information Systems Engineering Department
Ben Gurion University of the Negev
Beer-Sheva
Israel
(email: dinag, tsvikak, dlavie@bgumail.bgu.ac.il)

Abstract

Automatic Document Classification that corresponds with user-predefined classes is a challenging and widely researched area. Self-Organizing Maps (SOM) are unsupervised Artificial Neural Networks (ANN) which are mathematically characterized by transforming high-dimensional data into two-dimension representation, enabling automatic clustering of the input, while preserving higher order topology. A closely related algorithm is the Learning Vector Quantization (LVQ), which uses supervised learning to maximize correct data classification. This study presents the application of SOM and LVQ to automatic document classification, based on predefined set of clusters. A set of documents, manually clustered by domain expert was used. Experimental results show considerable success of automatic document clustering that matches manual clustering, with a slight preference for the LVQ.

Introduction

Motivation

Information users define subjective topic trees or categories, based on personal preferences and assign documents they write or receive to categories according to this subjective definition. Today, when the amount of publicly available information on the web is increasing rapidly by roughly a million pages every day [Chacrabarti, 1999], automation of categorization and filtering of documents is a must. For years, manual categorization methods were defined and employed in libraries and other document repositories, where the actual categorization was based on human judgment. Organizations and users search for information and save it in categories meaningful to themselves. The categories used by information users may be idiosyncratic to the specific organization or user, but not to an automatic classification system. Various methods and applications were developed and used for the purpose of automatic document classification. A major question in generating an automatic, adaptable system, for a specific user, is how close can the automatic system reflect the subjective point of view of the user, regarding his/her domain(s) of interest. Users may be inconsistent, changing their document classification over time. Moreover, users may find a document relevant to more then one category, choosing usually just one, arbitrarily, to host the document.

On the other hand, taking a well-defined set of categorized documents, agreed by a set of experts does not solve this problem.

The present study implements and evaluates, two artificial neural net methods, that address the specific user's preferences and deal with his/her varying subjective judgment

We used a set of manually categorized document from a specific domain of interest (documents about companies and financial news). We trained a SOM and LVQ ANN to automatically categorize the documents. Then, we measured the distance between the automatically generated set of clusters (or categories) and the pre-defined manual clustering. Our assumption is that if the distances are marginal, then, we can use the following model for automatic clustering using that ANN.

The following method was implemented: A user provides a training set of pre-clustered documents, this set is used for training the system, after training is completed, the system provides an automatic clustering based on the user preferences.

The present study deals with the following questions:

How close can come automatic to personal, subjective user categorization.

What is the effect of the training set size on automatic categorization performance (learning curve)?

What is the difference between supervised (LVQ) and unsupervised (SOM) training on the above questions?

Background

Clustering is defined as unsupervised classification of patterns into groups. Several clustering methods were presented and applied in a variety of domains, such as image segmentation, object and character recognition, data mining, and information retrieval [Jain et al, 1999].

Artificial neural networks (ANN) have been used in recent years for modeling complex systems where no explicit equations are known, or the equations are too ideal to represent the real world. The ANN can form predictive models from data available from past history. ANN training is done by learning from known examples. A network of simple mathematical "neurons" is connected by weights. Adjusting the weights between the "neurons" does the training of the ANN. Advanced algorithms can train large ANN models, with thousands of inputs and outputs. Analysis of the trained ANN may extract useful knowledge from it. Information retrieval and filtering is among the various applications where ANN was successfully tested [Boger et al, 2000]. Two main branches of ANN are in use, differentiated by their training methods: supervised and unsupervised:

a) The supervised ANN branch uses a "teacher" to train the model, where an error is defined between the model outputs and the known outputs. Error back-propagation algorithm adjust the model connection weights to decrease the error, by repeated presentations of inputs vectors.

b) The unsupervised ANN branch tries to find clusters of similar inputs when no previous knowledge exists about the number of the desired clusters.

In both cases, once the ANN is trained, and verified by presenting inputs not used in the training set.

Self-Organizing Maps (SOM), a specific kind of ANN, is a tool that may be used for the purpose of automatic document categorization [Kohonen, 1997; Honkela et al., 1997; Rauber & Merkel, 1999].

The SOM is an unsupervised competitive ANN, which transforms highly dimensional data into a two dimensional grid, while keeping the data topology by mapping similar data items to the same cell on the grid (or to neighboring cells). A typical SOM is made of a vector of nodes for the Input, an array of nodes as the Output map, and a matrix of connections between each Output unit and all the Input units. Thus each vector of the Input dimension can be mapped to a specific unit on a two-dimensional map. In our case each vector represents a document, while the output unit represent the category that the document is assigned to.

Closely related to the SOM algorithm is the Learning Vector Quantization (LVQ) algorithm. LVQ is a supervised competitive ANN which also transforms high dimensional data to a two dimensional grid, without regarding data topology. It uses pre-assigned cluster labels to data items, to facilitate the two dimensional transformation minimizing the average expected misclassification probability. Unlike SOM, where clusters are generated automatically based on items similarity, here the clusters are predefined. In our case the cluster labels are the subjective categorization of the various documents supplied by the user.

LVQ training is somewhat similar to SOM training. An additional requirement of the LVQ is that each Output unit will have a cluster label, a priori to training [Kohonen, 1997, p. 204].

SOM and LVQ may get as input a training set of documents, train the specific ANN for that set and, later on, cluster an incoming stream of new documents to the automatically generated categories.

In order to use any clustering mechanism, and specifically an ANN based approach, an appropriate document representation is required. The vector space model, in which a document is represented by a weighted vector of terms, is a very popular model among the research community in information retrieval [Baeza-Yates and Ribiero-Neto 1999]. This model suggests that a document may be represented by all meaningful terms included in it. Weight assigned to a term, represent the relative importance of that term in the representation. One common approach for term weighting is TF where each term is assigned a weight according to it's frequency in the related document [Salton & McGill, 1983]. This method was adopted in this study in order to generate documents representation for both ANN processing.

The rest of this paper is structured as follows: The next section describes the experiment performed for the evaluation of the training methods, then, follows presentation of the experimental results, concluding with a discussion and planned future work.

Experinent

Data

Data set containing 1073 economics and financial news-items, was used. The data was extracted from an online Internet based source (Yahoo.com). An information specialist read each item and manually clustered it into one of 15 possible clusters. 13 of them had exact topic definition while two others were general in nature. Manual clusters sizes varied from 1 to 125 documents per cluster. This approach represents a daily, normal operation of information seeking and categorization in organizations today, where vast amount of information is gathered and categorized.

Method

All 1073 documents were analyzed, using classical text analysis methodology. “Stop words” were removed. The resulted terms went under further stemming processing, using Porter stemming algorithm [Frakes and Beiza-Yates 1992]). Finally, normalized weighted term vectors representing the documents, were generated. At last, non-informative terms, which appeared in less then 100 documents, were excluded from the matrix. The SOM and LVQ simulations were implemented using the SOM Toolbox for Matlab 5 [Vesanto et al., 1999].

Training-sets of two different sizes were generated including 30% and 70% of the original data set.

The labeling process of the auto-generated cluster was as follows: we looked upon the original manual classification of the documents clustered at each output unit (automatic cluster). The manual cluster with the highest frequency, meaning where most of the documents belong to, rendered its name to the automatic output unit.

In order to evaluate the ANN performance, both automatic clustering results were compared to the manual categorization.

Initial experimental results

The experimental results are summarized in table 1. For the SOM algorithms, two simulations were conducted. In each simulation a SOM was generated, trained and tested. One simulation used 70% of the documents as a training set and the rest 30% were used for testing. In the other simulation 30% of the documents were used as a training set and the rest 70% were used as a test set. The results of the first simulation (70%-30%, learn-test, respectively) were as follows: During the training process, 71% of the training items were clustered correctly (e.g. a match was found between the manually generated clusters and the automatic clusters). Using the trained ANN for categorization of the test set, 67% of the test-set items were categorized correctly. The second SOM simulation (30%-70%, learn-test, respectively) yielded 73% correct classification for items of the training set and 57% correct classification for items of the test set.

Similar two simulations were made using the LVQ algorithm as well. Using 70%-30% simulation (train-test respectively) 75% of the items during the training phase were clustered correctly, and 67% of the test set items were categorized correctly, using the trained ANN. With the 30%-70% simulation, 77% of the items during the training phase were categorized correctly, and 65% of the test-set items were categorized correctly.

Table 1

	SOM (Unsupervised)			**LVQ** (Supervised)		
	100%	**70%-30%**	**30%-70%**	**100%**	**70%-30%**	**30%-70%**
Learn	72.69	70.84	72.98	74.74	74.43	76.71
Test	NA	67.08	56.72	NA	67.39	64.45

Discussion and further research

Considering the initial results, it looks like supervised learning yields better results when used with small training set then the unsupervised learning (which is expected). However, as the size of the training set grows, the advantage of the supervised learning tends to disappear which is somewhat unexpected. The former finding is leading to the initial conclusion that given enough training examples, automatic classification gets close to user subjective classification. Both methods yielded about 67% of overall success in categorizing the documents. Considering that the number of documents per category provided by the user was considerably different (from 1 item to 125 items), these results are very encouraging. In order to evaluate more accurately the performance of both methods, a more thorough analysis is needed. Besides the overall categorization success there is a need to evaluate each cluster separately, for precision and recall of specific clusters. The influence of the size of various clusters needs to be looked at. The size of the training set needs also to be better understood.. Above all, categorization results need to be evaluated by the original users, including re-evaluating the original categorization of data items that were wrongly categorized by the system, Since some times the documents may belong to more then one category, the different categorization between the system and the user may not necessarily be an error. All above-mentioned aspects are currently under research.

References

1. Baeza-Yates and Ribiero-Neto (1999) Modern Information Retrieval, Addison-Wesley, 1999.
2. Boger, Z. Kuflik, T., Shapira, B. and Shoval, P. (2000) Information Filtering and Automatic Keywords identification by Artificial Neural Network *Proccedings of the 8th europian Conference on Information Systems.* pp. 46-52, Vienna, July 2000.
3. Chakrabarti, S., Dom, B. et al. (1999). Hypersearching the Web. Scientific American, June, 54-60.
4. Honkela, T., Kaski, S., Lagus, K., and Kohonen, T. (1997). WEBSOM - Self-organizing maps of document collections. In *Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4-6*, pages 310-315. Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland.
5. Jain, A. K., Murty, M. N., Flynn, P. J. (1999) Data Clustering: A Review, ACM Computing Surveys, Vol 31, No. 3 pp. 264-323, September 1999
6. Kohonen, T. (1997). *Self-Organizing Maps*. 2nd ed., Springer-Verlag, Berlin.
7. Rauber A. and Merkl. D. (1999). Using self-organizing maps to organize document archives and to characterize subject matters: How to make a map tell the news of the world Proceedings of the 10th Intl. Conf. on Database and Expert Systems Applications (DEXA'99), Florence, Italy.
8. Salton, G., McGill, M. *Introduction to Modern Information Retrieval.* McGraw-Hill New-York (1983).
9. Vesanto, J., Alhoniemi, E., Himberg, J., Kiviluoto, K., & Parviainen, J. (1999). Self-Organizing Map for Data Mining in Matlab: The SOM Toolbox. Simulation News Europe, (25):54.