Flexibility Through Incremental Learning: Neural Networks for Text Categorization

P. Geutner, U. Bodenhausen and A. Waibel

Department of Computer Science
University of Karlsruhe
7500 Karlsruhe 1
Germany

School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
USA

Email: pgeutner@ira.uka.de, uli@cs.cmu.edu, ahw@cs.cmu.edu

Abstract

In this paper we show an adaptive incremental learning algorithm that learns interactively to classify text messages (here: emails) into categories without the need for lengthy batch training runs. The algorithm was evaluated on a large database of email messages that fall into five subjective categories. As control experiment best human categorization performance was established at 79.4% for this task. The best of all connectionist architectures presented here achieves near human performance (79.1%). This architecture acquires its language model and dictionary adaptively and hence avoids handcoding of either. The learning algorithm combines an adaptive phase which instantly updates dictionary and weights during interaction and a tuning phase which fine tunes for performance using previously seen data. Such systems can be deployed in various applications where instantaneous interactive learning is necessary such as on-line email or news categorization, text summarization and information filtering in general.

1 Introduction

A multitude of applications deal with categorizing text into subjectively chosen semantic categories. Example applications are:

- categorization of emails
- categorization of news
- categorization of paper abstracts for library retrieval systems
- information filtering
- summarization of texts, etc.

Traditional natural language processing methods usually require extensive development of grammars and dictionaries. This paper proposes the **ACCSYS**, an Adaptive Connectionist Text Categorization SYStem. ACCSYS uses a distributed pattern processing approach instead of rule based natural language techniques. These traditional methods are not useful for many approaches as they require extensive programming efforts and do not adapt when text material and categories change over time. Compared to traditional systems our system acquires its knowledge with every new text encountered by extending the underlying connectionist architecture. Our approach aims for the following properties:

- instant learning (one-shot learning)
- immediate updates during interaction
- no long batch processing on entire database
- no grammar needed
- adaptive vocabulary (instead predefined dictionary)
- easy adaptation to varying categories

2 Training Procedures

Two underlying training procedures were used for the ACCSYS-network:

Adaptive Learning: This algorithm was originally proposed by Gorin [Gor90, Gor91]. The connection weights w_{nk} between the n-th input unit v_n and the k-th output unit c_k of the network are estimated by the mutual information $I(c_k, v_n)$ between the set of input words v_n and the output categories c_k:

$$w_{nk} = I(c_k, v_n) = \log \frac{P(c_k|v_n)}{P(c_k)}$$

Compared to other training algorithms, e. g. gradient-descent learning, this learning procedure has one important advantage: learning can be done in one single step through the whole network. This allows to train the network towards many different tasks in a very short amount of time. Also learning is order-invariant to the presented training data and no step-size has to be determined. Most importantly the mutual information w_{nk} can be updated incrementally by each additional sample (message) without having to revisit previously seen data.

Backpropagation: Using the generalized delta rule [Rum86] as training procedure allows more complex network architectures. Some tasks may require internal representations. Gradient descent permits hidden units and network structures specially designed for spatio-temporal tasks like recurrent networks and Time-Delay Neural Networks [Wai89]. Nevertheless training may take very long until it converges to a minimum and requires the entire database to be on-line.

So adaptive training allows only single-layer networks and no discriminative training. It also yields a large number of parameters through its growing vocabulary. This leads to badly estimated connection weights. Gradient descent on the other hand allows complicated discriminative training and high performance on many tasks. Still it needs the entire database for training, incremental learning is difficult and long training times are expected.

Our final ACCSYS uses a hybrid approach between those two methods and achieves a combination of the advantageous properties of each technique. Adaptive incremental learning takes place in 2 phases:
- In an interactive mode the weights are set and reset by the mutual information criterion as described above, based on each new incoming token.
- Later, in the batch phase, the algorithm revisits all connections and fine tunes their weights discriminatively by using gradient descent and a pool of training data.

In the following experiments we evaluate the components of this approach by a series of experiments. The guiding principles behind these experiments are:

1. The attempt to reduce the number of parameters (to improve generalization performance).
2. The attempt to minimize batch training effort during phase 2 by suitable initialization with the results of phase 1.

3 Experimental Results

The system was trained and tested on a database consisting of 1204 email messages. These messages were taken from the archive of the connectionist mailing list administered by the Carnegie Mellon University. 946 mails were set aside as training material and the remaining 258 for test evaluation. The network was trained to classify the emails and copy them into the following subdirectories accordingly:

- conference announcements
- job offers
- paper announcements
- general discussion
- junk

The 1204 mails were labelled according to this division. All administrative information and routing information in the header was removed automatically, except for the subject and the sender of a message.

For the evaluation of the ACCSYS seven experiments were performed and are summarized in Table 1 and shown in Figure 1.

- As a control experiment (**HUMAN**) **human performance** was established by a group of four test subjects who were asked to classify the test set of 258 emails. Their average performance (with and without prior instructions) was 79.4%.
- In the first adaptive experiment (**ADPT**) baseline was performed by doing **adaptive classification based on entire mail messages**. Vocabulary grows quickly to more than 26000 words resulting in 132350 connections in the network. Given the relatively small database a poor test score of 47% results. Due to overtraining a major goal was to reduce the number of parameters [Moo91] to achieve the highest possible performance.
- **Reducing the messages to the first 50 words** (**ADPT_50**) decreases the number of different words in the lexicon from 26470 to 6428 — a reduction of 75% — and achieves a test performance of 72%.
- Another experiment (**ADPT_WP**) introduces word order by **allowing word pairs to be added**. The network architecture of the ACCSYS was extended by an intermediate layer representing word pairs. Test performance remains the same (training performance improves to 99%).
- To further reduce the number of units in the network **only morphological base forms** are used (**ADPT_MORPH**) reducing the vocabulary to 5694 words. However by preprocessing the data using a morph decomposition program [Hau92] ACCSYS's performance degrades.

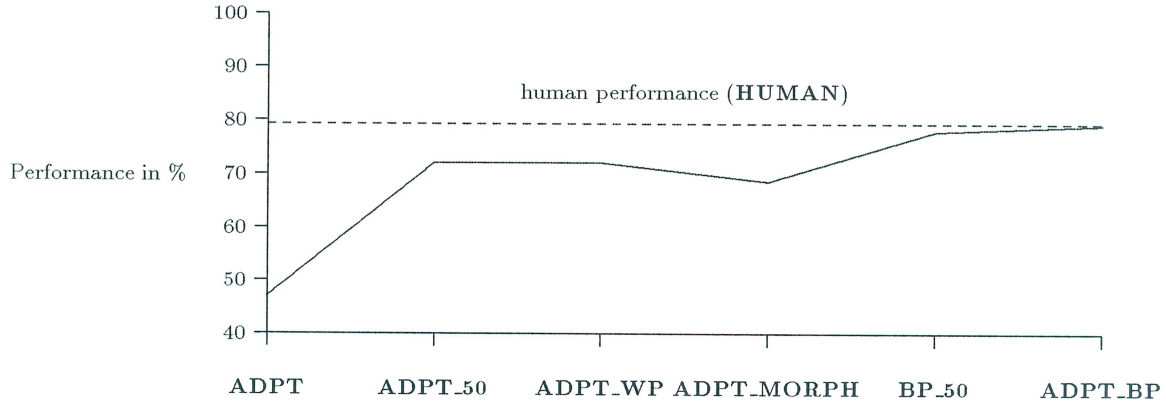

Figure 1: Test Results using Different Architectures

- Backpropagation on the entire database is too large to be handled by a work station. Hence the **backpropagation algorithm** was carried out only on the first 50 words of a message (**BP_50**). With initial random weights 77% test performance was achieved.
- The last experiment (**ADPT_BP**) combined the advantages of **gradient descent and adaptive learning**: here the network was trained adaptively first using **ADPT_50** and subsequently (overnight) fine tuned by backpropagation. After only 9 epochs backprop improves the **ADPT_50** result (72.1%) to 79.1%, a result that is comparable to human performance.

4 Conclusions

As can be seen, the combination of both adaptive learning and gradient descent achieves high class performance. Our system now performs text categorization near human performance. Moreover it offers the

Description	#Words	#Connections	#Epochs	Training Performance	Test Performance
HUMAN	?	$\approx 10^{14}$	?	?	79.4%
ADPT	26470	132350	1	76.7%	47.1%
ADPT_50	6428	32140	1	96.0%	72.1%
ADPT_WP	6428	42970	1	99.5%	72.1%
ADPT_MORPH	5694	40315	1	99.3%	68.6%
BP_50	6428	32140	26	99.9%	77.9%
ADPT_BP	6428	32140	9	99.5%	79.1%

Table 1: Simulation Results (see text for details)

desired flexibility and quick adaptation towards changing categories through a classifier that improves instantaneously through one-shot learning.

The ACCSYS represents a hybrid text categorization system that delivers:

- human-like performance
- on-line adaptive incremental and instantaneous learning
- adaptive dictionary
- no need for complex grammar rules

Additional work is underway to introduce word order into the classification process to put more emphasis on the semantics of an email. Experiments with constructive/destructive architectures will be pursued.

5 Acknowledgements

The authors wish to thank Dave Touretzky for giving us access to the connectionist archive and Roland Hausser for providing the LA-MORPH system.

References

[Gor90] A. L. Gorin, S. E. Levinson, A. N. Gertner, A. Ljolje and E. R. Goldman. *On Adaptive Aquisition of Language.* Proceedings of the IEEE 1990 International Conference on Acoustics, Speech and Signal Processing (ICASSP), Albuquerque, NM, pp. 601-604, April 1990.

[Gor91] A. L. Gorin, S. E. Levinson, A. N. Gertner and E. R. Goldman. *Adaptive Aquisition of Language.* Computer, Speech and Language, Vol. 5, pp. 101-132, April 1991, Academic Press.

[Hau92] Roland Hausser. *Principles of Computational Morphology.* submitted for publication, Computational Linguistics, 1992.

[Rum86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning Internal Representations by Error Propagation.* in *Parallel Distributed Processing.* Vol. 1, pp. 318-362, MIT Press, 1988.

[Moo91] John E. Moody. *The Effective Number of Parameters: An Analysis of Generalization and Regularization in Nonlinear Learning Systems.* in *Advances in Neural Information Processing Systems 4*, Edts. John E. Moody, Steven J. Hanson, Richard P. Lippman, pp. 847-854, 1991.

[Wai89] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang. *Phoneme Recognition using Time-Delay Neural Networks.* Proceedings of the IEEE 1989 International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 1989.