A Hybrid Neural Network for Automated Classification

Samea A. Wood *Tamás D. Gedeon*

School of Information Technology
Murdoch University
South St, Perth 6150, Australia

swood@it.murdoch.edu.au *tgedeon@murdoch.edu.au*

Abstract

The swift and ever increasing growth of the World Wide Web and other online information sources such as search engines, databases, digital libraries and newsgroups, has given rise to a new and growing problem: information overabundance. Existing information systems, including classification and retrieval systems, are struggling to cope with the large amounts of information now being produced, stored and accessed [6].

This research focuses on one aspect of this problem, automated document classification. In order to achieve this, the research is centered on the classification of newsgroup documents (postings) to relevant newsgroups using neural networks. Newsgroup documents were chosen due to their abundance and variety.

We will show that our hybrid connectionist approach has the capacity to outperform some more conventional classification techniques.

Keywords Information Retrieval; Document Management.

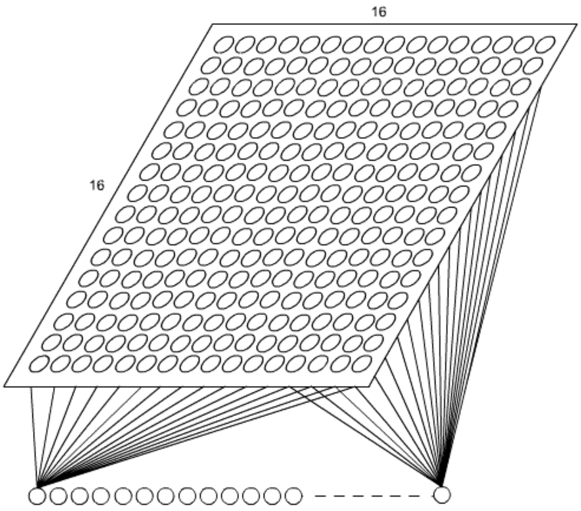
1 Introduction

This connectionist approach to automated document classification uses a self-organising map to generate document-word clusters. These document-word clusters are then used to train a back-propagation network to classify documents based on normalised word-frequency profiles.

2 Self-Organising Map (SOM)

The self-organising feature map (SOM) is a neural network architecture presented by Kohonen, although seeds of the same idea did appear elsewhere [1]. This neural network paradigm is often called the Kohonan Feature Map.

The self organising map is a two layer network consisting of an input layer and a competitive layer organised into a two dimensional grid. These networks have the ability to find the organization of relationships among patterns. Incoming patterns are classified by the units they activate in the competitive layer, with similarities among patterns mapped to closeness relationships on the competitive layer grid [1].



3 Back-propagation neural net (BPNN)

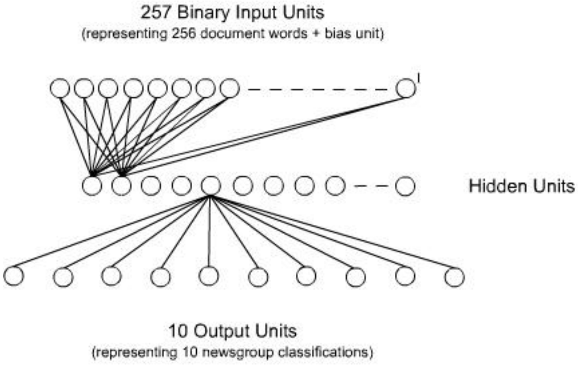
The back-error propagation concept was first presented by Paul Werbos in 1974, and then independently presented by David Parker in 1982. Also known as the generalised delta rule, it is a learning algorithm that can be applied to multi-layer networks.

Back-propagation networks have been applied to a broad range of application areas including image recognition, military applications, medical applications and speech recognition. It is the most widely used of the neural network paradigms.

Back-propagation networks are usually layered, with each layer fully connected to the one below and above it. When the network is given an input, the

Proceedings of the 6th Australasian Document Computing Symposium,
Coffs Harbour, Australia,
December 7, 2001.

updating of activation values propagates forward through the network to the output layer. During training, the error is calculated at the output layer and propagated backward, the weights updated as it progresses. Back-error propagation networks are best suited to problems that require pattern mapping: given an input pattern, the network produces an associated output pattern [1].


257 Binary Input Units
(representing 256 document words + bias unit)
Hidden Units
10 Output Units
(representing 10 newsgroup classifications)

4 Previous research

Kohonen, Kaski, et al [3] developed a two-level system, using self-organising maps at each level, to facilitate the browsing of newsgroups. The system used an SOM (self-organising map) to generate word clusters for the training documents, from which cluster histograms were generated. These histograms were then taken as input to a second SOM, which generated a document map. Related documents were clustered on the document map. This system was called WEBSOM.

Similarly, Lin et al [4] used a self-organising map for information retrieval. The SOM was used to construct a self-organising representation of the input documents, through the use of a document vector as input to the self-organising map. The output was a set of 140 cells, a 10x14 grid. After 2500 iterations, the system classified 140 and was subsequently used for information retrieval.

Troina & Walker [6] classified and searched documents in a micro-gravity database using a Hebbian Network and SOM. The Hebbian network was used to dynamically generate 'related terms' in a semantic vector form, which could be used to achieve query expansion by directly matching the semantic patterns of the query to the semantic patterns of the documents. The self-organising map received the document's semantic vectors as input and produced a number of document clusters based on these semantic patterns. These clusters may then be further broken down into smaller clusters, upon which queries may be directed, enhancing query performance.

All the research examined so far used self-organising maps to generate the final document clusters, or categories. Kohonen et al [3] also used a SOM to generate word dusters, which was used as input to a second document clustering SOM.

Given the ability of self-organising maps to map relationships between input data, it is of little surprise that this network architecture has become so widely used for such classification tasks [6]. However, a good deal of research has also been conducted into document and textual classification using other network paradigms, such as feed-forward networks. Feed-forward networks, including the back-propagation network, are used for their pattern recognition abilities. The ability to generalise (produce correct results for data instances on which the network was not trained) also makes feed-forward networks a useful tool for many pattern recognition tasks, including categorisation.

An example of the successful use of such a network in document classification is [3], who used an bilinear retrieval model implemented with a three layer feed-forward neural network to compute term associations based on document vectors presented to the network. The nodes in the input layer represented the document terms, while the nodes in the hidden layer represented query terms. The query term nodes were connected to the first layer document nodes, and the output layer was just one node. They showed that a smaller network with 200 terms performs comparably to one using all the terms of the collection.

Gedeon and Bustos [2] also conducted experiments using a three layer feed-forward network trained using error back propagation. The network, trained using input vectors generated from weighted cue word occurrence, title keyword frequency, and relative word location, and output vectors generated using calculated Inverse Document Term Weighting, generated expected Inverse Document Term Weightings (IDTW) for query documents. These IDTWs could then be used to discriminate between documents for retrieval and classification. They point out however, the use of title keyword frequency weighted cue word occurrence and relative word location results in this system's implementations being highly tailored to a specific domain, in the case of his research, the domain of civil law.

5 Document collection

The training and test document collection for this experiment consisted of newsgroup postings from a number of different newsgroups. This source was chosen as it facilitated the collection of a large number of documents, all of which were pre-classified, with classifications derived automatically from the source of the documents. The documents were collected over a period of several months due to low posting rates to some of the newsgroups.

Recent (up to January 2001) postings from ten newsgroups were used as the document set for this research. These ten newsgroups, corresponding to ten classifications, though distinctly different, can be considered similar enough to provide a substantial basis for testing the ability of the system to distinguish between documents of different literary genres.

Advertising documents and multiple instances (Spam) of documents across newsgroups were filtered from the set prior to any further processing of the documents, and all unwanted punctuation characters and NNTP (Network News Transport Protocol) header information was removed, with the exception of the subject line.

The newsgroups were:

1. alt.babylon5.uk
2. alt.books
3. alt.computer
4. alt.movies
5. alt.os.linux
6. alt.os.windows2000
7. alt.tv.farscape
8. alt.startrek
9. rec.humor
10. sci.astro.amateur

Newsgroup	Documents Collected
alt.babylon5.uk	6405
alt.books	3268
alt.computer	6078
alt.movies	5879
alt.os.linux	4911
alt.os.windows2000	4764
alt.tv.farscape	13047
alt.startrek	5357
rec.humor	5191
sci.astro.amateur	6002

Table 1: Newsgroup Documents.

5.1 Training Document Set

Exactly 30000 documents were used as the training document set for this experiment. This figure was decided upon by considering the number of links in the smallest back-propagation network (BP) that was to be used in the system and multiplying this figure by a minimum factor of 3 (with a generally accepted upper bound of roughly 10) before rounding. The following calculation yields the number of interconnections in the smallest and largest back-propagation networks used for this research.

Connections = (Input Nodes * Hidden Nodes) + ((Hidden Nodes + Bias Node) * Output Nodes)
Connections = (257 * 10) + (11 * 10)
or (257 * 25) + (26 * 10)
Connections = 2680 or 6685

As a result, the number of training documents required to adequately train the neural network is estimated to be between 3 * 6685 and 10 * 6685, or 20055 and 66850. The figure of 30000 was decided upon as it is the multiple of the number of neurons in the smallest network and 10 (rounded to the nearest 10000) and as it falls within the acceptable range. This figure gives an even 3000 training patterns per classification. This method of determining training set size is widely accepted within the artificial intelligence field.

These 30000 documents were selected randomly from the spam-filtered documents, 3000 per newsgroup. Random selection was used in order to obtain a training set representative of the full set of documents. The average, minimum and maximum number of lines for each newsgroup document set can be seen in Table 2.

Newsgroup	Average Lines	Min Lines	Max Lines
alt.babylon5.uk	39.14	3	788
alt.books	44.48	3	3516
alt.computer	32.07	3	746
alt.movies	51.12	3	7317
alt.os.linux	33.48	2	2949
alt.os.windows2000	28.23	3	442
alt.tv.farscape	36.73	3	178
alt.startrek	33.35	3	1034
rec.humor	36.21	3	946
sci.astro.amateur	31.76	3	1321

Table 2: Training Newsgroup Document Lines

5.2 Test Document Set

In order to test the system a number of documents not included in the original training set were required to determine the system's ability to generalise. That is, can it classify documents that it was not trained to classify? These documents were collected and filtered in the same manner as the training documents, prior to their use in testing.

The test document set consisted of 10000 documents, 1000 per classification, one third the number of training documents. The average, minimum and maximum number of lines for each newsgroup document set can be seen in Table 3.

Newsgroup	Average Lines	Min Lines	Max Lines
alt.babylon5.uk	34.73	4	507
alt.books	43.78	2	3170
alt.computer	27.54	4	167
alt.movies	33.44	3	561
alt.os.linux	36.16	1	1121
alt.os.windows2000	27.07	3	350
alt.tv.farscape	31.61	3	252
alt.startrek	31.91	4	434

rec.humor	37.63	3	716
sci.astro.amateur	33.29	3	444

Table 3: Test Newsgroup Document Lines

6 Word-Frequency Profiles

Once document collection and filtering was complete, document word-frequency profiles were generated for use in training the self-organising map, and later, the back-propagation network.

This in~~c~~volved counting all instances of all words within each document and generating an associative vector of words to word-frequency generated profiles for each document within the training set. For example:

"Computer science deals with the science of computing" would yield the following word-frequency vector. Assuming: [word, freq, word2, freq2]

[computer, 1, computing, 1, deals, 1,of, 1, science, 2, the, 1, with, 1]

These vectors were then used to determine which words are most representative of the documents as a whole, and would provide the maximum chance of allowing the SOM to classify the documents based on only a portion of the entire word frequency profiles.

For this research, 256 normalised word-frequencies were used to comprise each input vector to the SOM, so 256 words that were representative of the diversity and content of the original training documents had to be selected. The SOM was trained until no further reduction in network error occurred, often when occilation of the error commenced.

The 256 document-words were selected by examining the difference in total word-frequencies between the ten different classifications of documents. Roughly 25 words per comparison were selected manually, with common words excluded from the result. Words occurring in the newsgroup name were also included, with newsgroup type abbreviations being excluded. The figure of 256 for the number of document-words to use for this research was selected based on prior research (Bustos. 1994.) and in anticipation of future work. The largest factor limiting the number of document-words that could be used was processing power.

Trained using the back-error propagation algorithm, the back-propagation networks underwent supervised training with training patterns generated from the self-organising map output. The BP networks were trained for 10,000 epochs, this being determined a suitable figure via prior testing and examination of network error. ~~-~~Each training pattern consisted of 256 binary inputs.

These training patterns were generated with the objective being to provide the back-propagation network with enough information to allow it to classify the documents based on the occurrence and relationships between document words (mapped by the SOM) for documents in different newsgroups.

In order to achieve this, mapping the highest SOM activation node for specific document words, and setting the corresponding BP input node to "on" produced the set of training patterns. The corresponding input node was set to "on" for all words that occur within the document and the 256 document word set.

This was achieved by passing 256 test vectors through the trained SOM and logging the output. Each of these vectors had one element active, corresponding to a word. The order of words represented by these vectors was the same as the document-word training vectors, being alphabetical. Once the SOM output was logged, the index of the highest active output node for each word was found.

The training patterns were then generated, given that the pattern had the vector element at the index corresponding to a word's high SOM activation index turned on, if that word occurred in the training document. This method was used to generate all BP training and testing patterns. The expected output for the training vectors was a 10 element vector, one element for each newsgroup. A single vector element was turned on representing the newsgroup the document belonged to. The back-propagation network software [4] required training and testing patterns with an expected output vector appended to the input vector. In the case of testing however, this vector is only used for comparison with actual output and not for training.

7 Results

7.1 SOM results

In order to analyse the self-organising map's performance it is necessary to look at the document-word clustering it generated. This can be done using the document-word activation mapping that was used to generate the back-propagation input patterns by examining the highest activation node index for each document word.

Table 4 presents a sample of clusters from the final SOM word-activation mapping results

Word	Cluster	Word	Cluster
guy	7	box	30
service	7	network	30
man	144	cd	164
space	144	monitor	164
file	170	character	245
software	170	chinese	245
		western	245
borg	130		

class	130	enterprise	21
ops	130	mars	21
st	130	observer	21

Table 4: Example SOM Word Clusters

The most apparent attribute of these clusters is their small size. Indeed, the self-organising map clusters range from 2 words to a maximum of 9 words in one cluster. Nonetheless, the majority of clusters that were formed appear reasonable in the context of the document sets.

For example, character, chinese and western were classified as one cluster. Within the context of character-maps, computer hardware, movies and books this cluster is easily justified. Similarly, the larger set of borg, class, ops and st were classified as one cluster. Although not immediately apparent, this cluster is also justified when it is considered that each word has a strong connection to Star Trek, and therefore presumably, the alt.startrek newsgroup. The Borg, are a race from Star Trek. Class, is often used to designate the make of a Star Trek vessel (as in Naval vessel classes). Ops, is a common abbreviation for Operations within the Star Trek universe, and st is the equally common abbreviated form of Star Trek.

7.2 BP results

Evaluations of neural network results can be expressed in terms of recall and precision. In this instance, recall and precision measures were based on the classification performance of the system, with regard to the documents' sources. Hence, a correctly classified document is one where the network classification matches the original newsgroup that was the source of the document. These measures were generated for each of the ten classifications of newsgroup documents.

Table 6 shows the back-propagation network recall results, expressed in percentages, for each of the four different back-propagation network sizes. The results for both the training and testing data are shown; the values indicating what percentage of the total document sets were correctly classified.

	Training Data	Test Data
Hidden Nodes	Recall %	Recall %
10	72.24%	57.67%
12	72.91%	58.62%
15	75.86%	57.09%
25	79.49%	55.90%

Table 6: Recall Results

Table 7 gives the average precision measure across each of the 10 classifications, for each of the four tested network sizes and two sets of data. This measure is the average precision, expressed as a percentage, of the ten precision measures.

	Training Data	Test Data
Hidden Nodes	Avg Precision	Avg Prec
10	82.30%	68.76%
12	90.77%	76.94%
15	92.93%	75.13%
25	95.73%	75.54%

Table 7: Average Precision Results

The most apparent observation to be made from table 06's results is the increase in Training Data Recall accuracy as the number of hidden nodes increases, also indicated by the decrease in total sum squared (TSS) error. This is not unexpected however, as increasing the number of hidden nodes often results in back-propagation networks that are better able to learn their training data. This is due to an increase in the ability of the network to detect "features" in the input patterns. Unfortunately, this may also result in over-fitting: the network learning the training data very specifically, reducing its ability to generalise.

An examination of the Test Data Recall values quickly confirms that over-training has occurred in the larger back-propagation networks. The most obvious example of this can be seen in a comparison of the recall values for the back-propagation networks with 12 and 25 hidden nodes.

Test Data – 10000 Documents
Correctly Classified: 5862
Total Recall: 58.62%

		Recall		Prec
alt.babylon5.uk	446	44.6%	528	84.5%
alt.books	587	58.7%	772	76.0%
alt.computer	462	46.2%	765	60.4%
alt.movies	538	53.8%	651	82.6%
alt.os.linux	715	71.5%	813	88.0%
alt.os.windows2000	599	59.9%	882	67.9%
alt.tv.farscape	624	62.4%	709	88.0%
at.startrek	724	72.4%	1013	71.5%
rec.humor	585	58.5%	728	80.4%
sci.astro.amateur	582	58.2%	829	70.2%
Average:		58.6%		76.9%

Table 8: Results for BP net with 12 hidden nodes

Table 9 shows to which newsgroup the highest number of misclassified documents from any single newsgroup were classified, including the number of documents. As can be seen, in many instances these misclassifications were made to similar newsgroups. For example, 155 documents misclassified to alt.computer instead of alt.os.windows2000, which is a related newsgroup.

Testing Data Set	No. Docs	Classification
alt.babylon5.uk	79	alt.startrek
alt.books	48	alt.startrek
alt.computer	149	alt.os.windows2000
alt.movies	34	alt.books
alt.os.linux	63	alt.os.windows2000
alt.os.windows2000	155	alt.computer
alt.tv.farscape	46	alt.startrek
alt.startrek	68	sci.astro.amateur
rec.humor	46	sci.astro.amateur
sci.astro.amateur	50	alt.computer

Table 9: Misclassified Document Results

Conclusion

The results obtained have shown that the presented system of neural networks and document processing, with a back-propagation network of 12 hidden nodes, is capable of classifying approximately 59% of newsgroup documents presented to it, provided these documents are from the 10 trained classifications. The system can also recall 73% of documents on which it was trained. Although using a larger back-propagation network results in a higher training document recall, results have shown that this reduces the system's ability to generalise.

Results have also shown that word clustering was performed by the self-organising map with justifiable results.

The results of 59% recall and 76.95% average precision for the hybrid network with 12 hidden nodes compares favourably with results for nearest neighbourhood clustering of 59.3% recall and 59.9% average precision. Althought a small difference in recall exists, amounting to roughly 30 documents out of the 10,000, the hybrid system of neural networks performed with 17.05% greater precision, indicating a higher capacity to differentiate between clusters.

This conclusion is further supported by results for a stand-alone BP network classification of the test data using the same parameters as the BP network from the hybrid system, but trained on the training set of normalised word frequency vectors directly. Results of 40.81% recall and 76.45% average precision on the testing data clearly demonstrate the greater ability of the hybrid system to successfully classify testing patterns on which it was not trained.

These two comparisons clearly demonstrate that the hybrid network is capable of achieving a higher level of recall and precision, as opposed to more conventional systems, ~~which tend to trade~~rather than just one or the other. ~~one against the other.~~

Given the generally small size of newsgroup documents, and a limit of 256 document words for classifying these documents, many of which had only a small number of document word occurrences, these results are an encouraging outcome. Additionally, the interrelated nature of many of the newsgroups further highlights the ability of the system to distinguish between document sets.

~~**Results gathered on this same data using other clustering methods have demonstrated that this hybrid connectionist approach may have the ability to significantly outperform more traditional classification methods, although further research and refinement of the system is necessary to validate this hypothesis.**~~

References

[1] Dayhoff, J. 1990. Neural Network Architectures. Van Nostrand Reinhold, New York.

[2] Gedeon, T.D. and Bustos, R.A. (1996) "Word-Concept Clusters in Document Collections", Australian Document Computing Symposium, Melbourne, vol 1, pages 21-24.

[3] Kohonen, T. Kaski, S. Lagus, K. and Honkela, T. 1996. Very Large Two-Level SOM for The Browsing of Newsgroups. In Proceedings of International Conference on Artificial Neural Networks, 1996.

[4] Lin, X. Soergel, D. Marchionini, G. 1991. A Self-Organizing Semantic Map for Information Retrieval. Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 262-269.

[5] McClelland, J, L. Rummelhart, D, E. 1991. Explorations In Parallel Distributed Processing. A Handbook Of Models, Programs and Exercises. A Bradford Book. MIT Press. London.

[6] Troina, G. Walker, N. 1996. Document Classification and Searching - A Neural Network Approach. ESA Bulletin N87, Frascati, Italy.

[7] Wong, SKM. Cai, YJ. and Yao, YY. 1993. Computation of Term Association by Neural Network. SIGIR '93 Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.