Neural Net Learning Issues in Classification of Free Text Documents

Venu Dasigi and Reinhold C. Mann*

Abstract

In intelligent analysis of large amounts of text, not any single clue indicates reliably that a pattern of interest has been found. When using multiple clues, it is not known how these should be integrated into a decision. In the context of this investigation, we have been using neural nets as parameterized mappings that allow for fusion of higher level clues extracted from free text. By using higher level clues and features, we avoid very large networks. By using the dominant singular values computed by Latent Semantic Indexing (LSI) [Deerwester, et al., 90] and applying neural network algorithms for integrating these values and the outputs from other "sensors", we have obtained preliminary encouraging results with text classification.

1 Introduction

The ever-increasing volume of information from various scientific, commercial, industrial and intelligence sources is currently overwhelming the best available methods for information processing. Millions of bytes of data are commonplace, and most practical systems should realistically be expected to handle billions or even trillions of bytes of data [Harman, 93]. The objective of the work summarized here was to develop a system that would allow proof-of-principle experiments using an approach that incorporates standard machine learning technologies with standard methods for text retrieval, such as LSI, key word searches, and others. Specific issues to be addressed include:

- The selection features to be included in the classification, e.g., how many dominant singular values computed by LSI? how many additional clues?
- The size of the training set required to achieve acceptable performance
- Experiments that allow for quantitative benchmarking of system performance

A large class of text indexing and retrieval methods (including LSI) is based on vector space representations where documents and queries are represented as vectors in a generally very high-dimensional space. Dimensions of the order of several thousand are not uncommon. Applying neural networks to these representations directly would result in impractical algorithms due to the enormous number of parameters to be estimated in the network. Therefore, our approach to text document analysis outlined in this paper incorporates the following key elements: (1) extraction of multiple features from the text documents, using an LSI-based sensor (to be described) as a primary feature extractor, and (2) the application of machine learning methods and recent finite sample results for empirical estimation in order to estimate the parameters of a fusion mapping that integrates these multiple features. In the current system we are using a simple backpropagation network for fusion.

2 A Multi-Sensor Neural Net Approach - Background

This work was, in part, motivated by the success of Gene Recognition and Analysis Internet Link (GRAIL), a pattern recognition system which used a multi-layer, feed-forward neural network that receives inputs from several sensors that measure different characteristics of the signals or data sets to be analyzed [Xu, et al., 94]. The net acts as a classifier and assigns the input pattern to a given number of classes, after being trained. The neural net represents a reliable mechanism to integrate the information from multiple sources to form a combined best estimate of the true classification decision. The term "sensor" is interpreted in a broad sense. It can encompass a real physical sensor device or a "logical sensor", that is, an algorithm that computes a feature [Uberbacher, et al., 95].

We start with the hypothesis that a GRAIL-like system would be very appropriate for classification and filtering of English text documents. We expect the system to be capable of integrating in a systematic way existing and new algorithms as required by the application. The GRAIL-type system can integrate different kinds of sensors, e.g., statistical and syntactic sensors as well as simple keyword sensors, and other standard techniques already in use by document analysis community.

A major stumbling block in applying neural networks to most IR applications has been that the size of a typical IR problem results in impractically large neural networks. Typically, documents to be classified as well as

*Authors' addresses: **Venu Dasigi**, Dept. of Computer Science and Information Technology, Sacred Heart University, Fairfield, CT 06432-1000, e-mail: dasigi@shu.sacredheart.edu. **Reinhold C. Mann**, Intelligent Systems Section, Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6364, e-mail: mannrc@ornl.gov.

retrieval queries are represented as a set of terms, the size of which is at least in the thousands. In such networks there could be hundreds of thousands of connections, not to mention the complexity when lateral inhibition is added for a winner-takes-all effect, e.g., [Wilkinson and Hingston, 92]. An LSI-based approach may be used to address the issue. In addition to adding trainability, a neural network can integrate the information in inputs coming from other logical sensors into the final outcome.

3 Two Logical Sensors

Specifically, in this initial effort, we focused on two main goals. First, create input to a neural network that is LSI-based, so that the size of the neural net will be practical, and it can be trained without much difficulty. Further, a second goal is to see if additional sensors can be added easily to the neural net input, to give improved results. The relationship between the LSI component and the neural network is symbiotic. Our work attempts to exploit the dimensionality-reduction capability of LSI, and combine it with the powerful pattern-matching and learning capabilities of neural networks. The LSI-based input enables the neural network input to be of a much smaller size than a long term vector. The neural network adds trainability to the LSI-based method, and also makes it possible to integrate other sensors to complement or supplement the LSI-based input.

In LSI, a large and sparse term-document matrix is reduced into three relatively small matrices (one of which is simply a diagonal matrix) by singular value decomposition (SVD) corresponding to a number of the larger singular values. Instead of representing documents by thousands of possible terms, LSI allows a document to be represented by around a hundred "factors" that are supposed to capture the "significant" term-document associations. This is done by some linear transformations of the much longer term vector, using the constituent matrices that result from the SVD of a "reference matrix". A *reference matrix* is the term-document matrix of a *reference library/collection* of documents. A reference library is simply the collection of documents that "adequately" represents all concepts of interest.

The input to the system is an individual document that needs to be classified into one of several categories. Different logical sensors are applied to the document, constituting different kinds of preprocessing to derive salient features. The first such sensor of interest to this work is based on the term vector representing the input document, which is reduced to a much smaller size using an LSI-based linear transformation. The features derived by the logical sensors constitute input to a neural network that has already been trained. The output is an indication of the category to which the document belongs.

The purpose of the second logical sensor currently used in this work is to allow for simple keyword profiles to be considered in the classification. Each category profile is simply a set of keywords characterizing that particular category. There is one input to the neural network from this logical sensor, corresponding to each category. Each output simply represents what fraction of the terms in the given document match the category profile. Inclusion of more sophisticated algorithms is the subject of ongoing research.

4 Experiments

Our initial focus was exclusively on a number of AP news wire stories from the standard TIPSTER collection [Harman, 93]. The collection contains AP news wire stories for two full years, tens to hundreds of stories per day. The purpose of the multi-sensor neural net is to classify the news stories into one of ten ad hoc categories, such as accidents, crime, business and finance, culture, politics and government, weather, obituary, etc. For the documents used for training and testing purposes, the categories of the news story documents were manually determined, which turned out to be a bottleneck.

Despite dimensionality reduction of the input vector through LSI, the neural network is of a substantial size, with more than one hundred input nodes and about ten output nodes. Such a network typically requires several thousand training inputs, and this requirement increases with the number of hidden units. Within the time frame of this initial effort, we manually categorized a few hundred news stories. Consequently, the following results are far from conclusive. We believe that they do, however, speak for the promise of the approach, in spite of the fact that the neural net is inadequately trained. Further work on training and performance evaluation is underway.

Of the nearly five hundred documents used for training, about three-quarters were used as the "reference library" for all LSI/SVD operations. The number of SVD "factors" used in this work was over one hundred. The neural net was a simple feedforward net with back propagation, and used the delta rule for learning and the tanh transfer function. It was tested in two configurations; one with the (over one hundred) inputs just based on LSI alone (we call this version the single sensor network) and another with the LSI-based inputs *plus* another ten inputs based on simple category profiles (called the two sensor version). Both configurations used ten output units, one for each category.

5 Summary of Results and Analysis

We compared the multi-sensor neural net approach against an LSI-based classification. The original LSI approach was modified to do classification by first identifying the document from the reference library that best matches the input document, and then looking up the category of the reference document. Table 1 summarizes the main results, which are discussed below.

When the LSI method was used by itself to perform classification, the results were somewhat surprising. Although the performance of LSI in classifying the *known reference library* documents was a perfect 100%, the percentage of correct results when *new documents* outside the reference library were used dropped to 54%. For this method, there was essentially just a single experiment, because generating the SVD for each new reference library was computationally very expensive even when several megabytes of main memory were devoted to the program.

Approach Name	Range of Iterations	Correct (Test)	Correct (Training)
LSI alone	N.A.	54%	100%
Single Sensor Net	16-64K	58-72%	76.05-80.7%
Two Sensor Net	16-64K	58-75%	80.47-85.11%

Table 1: Summary of Classification Results with Approaches

For the neural network experiments, the percentages of correct results as cited represent the peak performance that did not get any better with more iterations. Since there were only a limited number of inputs, several different experiments were constructed by cross validation. We cross validated the available data by generating a dozen pairs of files, each pair containing a training file (90% of data) and a test file (10% of data). The same pairs of data sets were used to test both the single sensor neural net and the two sensor one.

Single sensor neural net: With each training set, the neural net was trained for between 16,000 and 64,000 iterations. On the test set, the correctness percentage ranged from a minimum of 58% to a maximum of 72% for the dozen sets. When the training set itself was used as a data set, the performance was between 76.05% to 80.7% correct (contrasted to 100% in the LSI-only method).

Two sensor neural net: Again, with each training set, the neural net was trained for between 16,000 and 64,000 iterations. On the test set, the correctness percentage ranged from a minimum of 58% to a maximum of 76% for the dozen sets. When the training set itself was used as a data set, the performance was between 80.47% to 85.11% correct.

When results with the individual data sets are closely studied, a clear improvement in performance of 2 to 4 percent was noted with 9 out of the 12 data sets in the two sensor neural net. One of the other cases showed no change, but reached that level of performance with 16,000 iterations rather than 48,000. In the remaining two cases, a marginal decrease of performance compared to the single sensor version was observed. But in one of these two cases, the superficially better performance of the single sensor neural net decreased a few percent after more iterations. These anomalies can perhaps be attributed to the simple-mindedness of the second sensor that was employed.

We believe that one thing the results conclusively indicate is that the neural nets need more training inputs. There is a clear improvement of classification results in the neural net approach compared to the LSI method by itself. And the two sensor version, even with a very simple-minded second sensor, seems to do better in most cases than the single sensor version.

Other researchers appear to have evaluated LSI-based approaches, too [Schuetze, Hull and Pederson, 95]. Our approach differs from theirs in using a reference library, and in employing multiple sensors. Some details of our approach may be found in [Dasigi and Mann, 95]. It may be noted that SVD is computationally very expensive, but our approach performs an SVD just once - on the reference collection. This can be done just once, before training or testing begins. Of course, the onus is on us to make sure that the reference library indeed represents all the concepts adequately. The impact of the choice of the reference library on the overall performance still needs to be studied.

This work is far from complete. Clearly the neural networks need more training, and more training requires more data, which in turn requires news stories to be manually categorized. An important extension to the work here would be in the area of other input sensors. LSI is powerful, but is limited by the vocabulary seen so far in the reference library. The second sensor that was implemented attempted to avoid this limitation using a very simple technique. The slight improvement of results even with such a simple addition is very encouraging, but also points to the need for more informative sensors.

6 Acknowledgments

Venu Dasigi thanks the Oak Ridge Institute for Science and Education for the fellowship that made this work possible. Both authors thank Mike Berry for distributing the SVDPACKC software.

References

[Dasigi and Mann, 95] Dasigi, V. and R. C. Mann, Toward a Multi-Sensor-Based Approach to Automatic Text Classification, *ORNL/TM-13094*, Oak Ridge National Laboratory, Oak Ridge, TN 37831, October, 1995.

[Deerwester, et al., 90] Deerwester, S., S. Dumais, G. Furnas, T. Landauer and R. Harshman, Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, 41(6), pp. 391-407, 1990.

[Harman, 93] Harman, D., Overview of the First TREC Conference, *Proc. of SIGIR-93*, pp. 36-47, 1993.

[Schuetze, Hull and Pederson, 95] Schuetze, H., D. Hull and J, Pedersen, A Comparison of Classifiers and Document Representations for the Routing Problem, *Proc. of SIGIR-95*, pp. 229-237, 1995.

[Uberbacher, et al., 95] Uberbacher, E. C., Y. Xu, R.W. Lee, C.W. Glover, M. Beckerman, R. C. Mann, Image Exploitation Using Multi-Sensor/Neural Network Systems, *Proceedings of the 1995 Applied Imagery and Pattern Recognition Workshop*, Washington DC, October, 1995, SPIE, in press.

[Wilkinson and Hingston, 92] Wilkinson, R. and P. Hingston, Incorporating the Vector Space Model in a Neural Network used for Document Retrieval, *Library Hi Tech*, 10(1-2), pp. 69-75, 1992.

[Xu, et al., 94] Xu, Y., R. Mural, M. Shah and E. Uberbacher, Recognizing Exons in Genomic Sequence using GRAIL II, *Genetic Engineering, Principles and Methods*, Plenum Press, 15, June, 1994.