Feature Selection and Document Classification in a Massive Dataset

Categorizing Documents for the BBC Website

Peter Zeidman, University of Birmingham School of Computer Science

pzeidman@gmail.com | supervisor: Dr. Peter Tino

INTRODUCTION

Introduction

• Whenever a journalist adds a page to the BBC website, he or she must assign keywords ("metadata") to the page so that it may be filed appropriately.

• A computer system automatically suggests keywords to the journalist. This is based on hand-written rules, for example "If the journalist's page contains the word 'score' then suggest the keyword 'sport' ".

• This process of hand-writing rules is laborious. In addition, the journalist may assign keywords which are not appropriate.

Project Objectives

1. Investigate approaches to analysing large sets of data, including representation, feature selection and automatic classification.

2. Build an automatic document classifier for the content on the BBC website.

3. Compare the performance of the classifier against others reported in the academic literature.
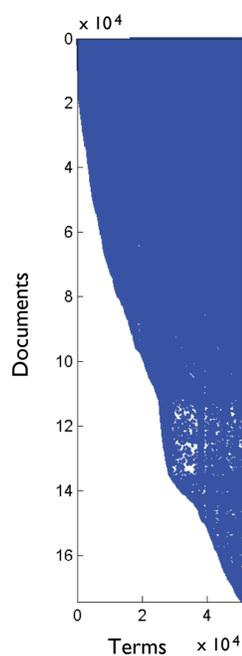

Figure 1. Sparsity diagram showing the distribution of documents and terms in the BBC dataset. Part of Objective 1, above.

The Data / Problem Domain

• The BBC dataset consists of 51,075 documents (every article from 32 English Regions websites). There were 174,456 unique terms (words) prior to processing.

• Initial experiments were conducted on a smaller dataset - six categories of the 20-Newsgroup set, which was limited to the first 200 documents per category.

The data was imported into a term-document matrix with binary coding for computational efficiency.

FEATURE SELECTION

Overview

• The pages of the BBC website were imported and represented as vectors - each dimension representing a word in the vocabulary.

• We were working with a very sparse high-dimensional document space - presenting us with **the curse of dimensionality**. This refers to the problem that the volume of a space increases massively as the number of dimensions increases.

• **Feature Selection** is the process of choosing only those features (words in the vocabulary) which are useful for classifying documents. By reducing the number of features, we are reducing the dimensionality of the search space, and improving performance.

Methods

Two approaches were compared:
• **Model-based**, where the reduction in dimensionality is based on transforming the document vectors. The technique employed selected was Random Projection, which projects the term-document matrix to a lower dimensional subspace.
• **Model-free**, where some independent test is applied to each feature to assess its value. Information Gain was used, which compares the entropy of the dataset with and without the term in question.
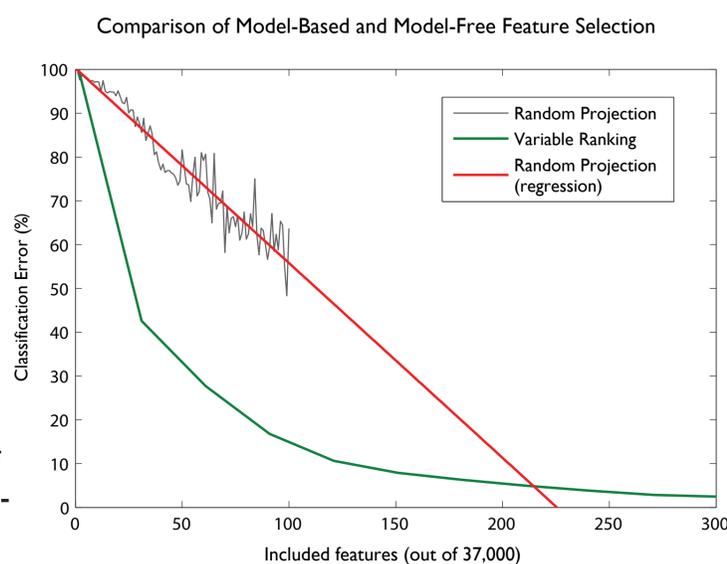
Results

Figure 2. Comparison of feature selection approaches

Conclusions

The model-free approach (Variable Ranking based on information gain) outperformed the model-based approach (Random Projections). The stochastic nature of Random Projection means that it takes considerably more computational power to find a strong projection than for variable ranking. The downside to Variable Ranking is that each feature is considered individually; it is possble, however, that certain features only have value when considered in conjunction with others.

CLASSIFICATION

Overview

• Once feature selection had been applied, reducing the dimensionality of the dataset to a workable size, classification of the documents could begin.

• Classification is the process of deciding on the most likely class, or set of classes, for a variable. In this case, the system decides which keywords (categories) will be appropriate for each web page.

Methods

The classification approach was based on hierarchical neural networks. These networks:

• Break down the classification problem into smaller sub-problems.

• Allow for individual classifiers to be re-trained when the dataset is altered, without needing to retrain the whole system.

• Allows for simple hardware distribution in a multi-CPU environment.
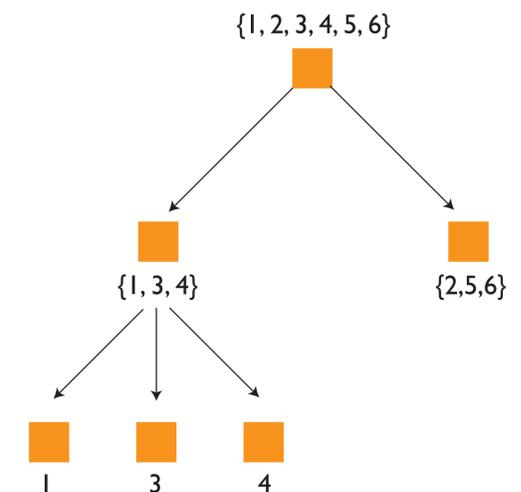

Figure 3. Hierarchical classifiers. Each box represents a classifier, labelled with the set of classes it is trained to recognise. The training data is initially re-labelled to give a branching factor of six.

Results

Formal experiments, together with comparisons against traditional neural network classifiers, are still running at the time of going to press. We predict that hierarchical classifiers will outperform any individual classifier, with improved generalization ability.

Further Reading

D'Alessio, S., Murray, K., Kershenbaum, A. & Schiaffino, R. The Effect of Using Hierarchical Classifiers in Text Categorization. Proceedings of RIAO-2000, April 2000.

Acknowledgements