A Neural Network Approach to Topic Spotting

Authors: Wiener, Pedersen and Weigend

Introduction

• What topics is a document about?
• Topic spotting is useful for information filtering and organization.
• A Neural Network can be used to classify the topic of a document.
• Reuters newswire corpus used to evaluate the success of topic spotting.
• Over 90 topics and about 12,000 documents - 2/3rds of which are used for training
Basics of the Approach

• Construct a Neural Network for each topic:
  – accepts a representation of a document
  – outputs a number between 0 and 1
  – output close to 1 => topic present, close to 0 =>
topic absent
• Need to encode the document so that it can
be used by the NN
• Need to design the architecture of the NN

Encoding the Documents

• Use the Bag of Words approach
• doc=<weight w1, weight w2, …weight wn>
  – where there are n unique words in the training
corpus.
  – weight of a word in a document is basically its
normalized frequency in the document (many
variations).
  – Over 10,000 unique words in the Reuters
corpus => our document vectors contains over
10,000 attributes.
Reducing the Dimensionality

- Remove stop words e.g the, it, at, etc. these words do not indicate a document’s topic.
- Apply stemming e.g runs, running, runner are all mapped to their stem “run”
- Still over 10,000 words!
- Feature selection technique used to select the most informative words for each topic
- Feature extraction - Latent Semantic Indexing

Architecture of the NN

- Backpropagation with gradient descent
- Inputs: vector of word weights (bag of words)
- Single hidden layer with 6 neurons
- Transfer function: $1/(1+e^{-n})$
- Output in (0,1) - threshold to a hard classification
Summary of the Classification Process

DOC

Vector

NN - Topic A

(0,1)

If 0 then doc is predicted not to contain topic A, if 1 then doc is predicted to contain topic A.

Results

• 0.82 (achieved with LSI feature extraction)
• Nonlinear network only slightly better than linear network.
• Too few positive examples to support nonlinear fits that generalize well?
• Performance can be slightly improved by using many more hidden neurons.
Evaluation of the Approach

• Advantages
  – enough power to represent the concept
  – noise resistant
  – good performance
  – humans do not have to construct topic spotting rules

• Disadvantages
  – can’t deal with a very large number of attributes
  – not scrutable
  – number of hidden neurons is not a solved problem
  – can get stuck in local minima

Future

• State of the art: high 80’s (boosting, SVM)
• Significant improvements most likely to come from advances in the area of Natural Language Processing - extracting high level features
• Smaller improvements may come from better Machine Learning algorithms and improved dimensionality reduction techniques.