Active Learning for Text Classification

CS 365 Project Report
Ankit Bhutani, Y9094
Mentor: Dr. Amitabh Mukherjee

Abstract

With the availability of large volumes of text data on internet and other text databases, it becomes absolutely necessary to be able to classify or label doucments in order to enable easy maintainence of databases and to support various operations like searching, clustering, query reply etc. The labelling of documents can be done either manually or automatically. Manual Labelling is not only time consuming, but also very costly and so, there have been lots of attempts to do it an automated manner. In our project, we have attempted to make a classification system which can learn to classify documents with minimal help from a user (human being). The system, given a large number of unlabelled documents and very few labelled documents, can decide, which unlabelled documents it is able to label on its own and which documents it should request for label from a user(Active Learning), so as to maximize its efficiency of labelling with minimum number of requests.

1 Introduction

As stated above, we cannot undermine the need to have systems that can automatically label documents. Further, these systems should be able to evolve with time with the help of minimal supervision from human beings and human effort in labelling is costly as well as time-consuming. This project presents a technique for using a large pool of unlabelled documents to improve text classification enabling requests to a user for labelling a few documents, which are expected to maximize the information with a decent trade off for the cost incurred due to the request. In section 2, we describe the related work done in this area. In section 3, we describe the framework for our present research along with the avialable data-set. In section 4, we describe the experimental results along with the conclusion.

2 Related Work

Expectation Maximization (EM) is often chosen to make use of the unlabeled data for learning a Multi Nomial Naive-Bayes (MNB) model [3]. The combination of EM+MNB produces a fast semi-supervised learning method. Further, in [2], it was shown that Active Learning + EM + MNB could reduce the requirement of number of labelled examples to little more than half as compared to EM + MNB alone. In [4], a new method for parameter estimation for MNB model was proposed, and it was shown to be not only faster than EM+MNB, but also generated better AUC compared to EM for most of the datasets without any loss in accuracy. In this project, we explore the possibility of combining SFE + Active Learning, and demonstrate that although Active Learning reduces the requirement for the number of labelling examples, but thsi reduction is not as much significant as in the case of Active Learning + EM.

3 Framework

3.1 Document Representation

In text classification, a labeled document d is represented as d = $\{w_1, w_2, \ldots, w_i, c\}$, where variable or feature w_i corresponds to a word in the document d, and c is the class label of d. The set of unique words w appearing in the whole document collection is called vocabulary V . Typically, the value of w_i is the frequency f_i of the word w_i in document d. We use the boldface lower case letters **w** for the set of words in a document d, and thus a document can also be represented as { **w**, c}. We use T to indicate the training data and the d^t for the t_{th} document in a dataset T . Each document d has $\|d\|$words in it. In general, we use a hat $\hat{()}$ to indicate parameter estimates. Text representation often uses the bag-of-words approach. By ignoring the ordering of the words in documents, a word sequence can be transferred into a bag of words. In this way, only the frequency of a word in a document is recorded, and structural information about the document is ignored.

3.2 Multinomial Naive Bayes (MNB) + Semi-Supervised Frequency Estimate (SFE)

The task of text classification can be approached from a Bayesian learning perspective, which assumes that word distributions in documents are generated by a specific parametric model, and the parameters can be estimated from the training data. Equation 1 shows Multinomial Naive Bayes (MNB) model [3] which is one such parametric model commonly used in text classification:

$$P(c/d) = \frac{P(c)\Pi_{i=1}^{n} P(w_i/c)^{f_i}}{P(d)} \qquad \ldots(1)$$

where f_i is the number of occurrences of a word w_i in a document d, P (wi/c) is the conditional probability that a word w_i may happen in a document d given the class value c, and n is the number of unique words in the document d. P(c) is the prior probability that a document with class label c may happen in the document collections.

The parameters in Equation 1 can be estimated by a generative parameter learning approach, called maximum likelihood or frequency estimate (FE) , which is simply the relative frequency in data. FE estimates the conditional probability P (wi/c) using the relative frequency of the word wi in documents belonging to class c.

This rest of the part is same as the implementation in [4] in which the concept of SFE has been introduced.

3.3 Active Learning

Active Learning, as in the present setting, is a form of supervised learning wherein the system can request for labels of some unlabelled documents at the expense of some cost. An optimal active learner should select those documents, that when labelled and incorporated into training, will minimize classification error error over the distribution of future documents. We are using a particular method of active learning called Query By Committee (QBC). It samples several times from the classifier parameter distribution that results fro the training data, in order to create a committee of classifier variants. This committee approximates the entire classifier distribution. QBC then classifies each unlabelled documents with each committee member and measures the disagreement between their classifications-thus approximating the classification variance. Finally, documents on which the committee disagrees strongly are selcted for labelling requests. The newly labelled documents are included in the training data and similar process is carried out for the next set.

To capture the information regarding disagreement between committee members, we use the Kullback-Leibler (KL) divergence to the mean. Each committee member m produces a posterior class distribution, $P_m(C/d_i)$, where C is a random variable over classes. KL divergence to the mean is an average of the KL divergence between each distribution and the mean of all the distributions:

$$\frac{1}{k}\Sigma_{m=1}^{k} D(P_m(C/d_i)||P_{avg}(C/d_i))$$

where $P_{avg}(C/d_i)$ is the class distribution mean over all committee members, m: $P_{avg}(C/d_i) = (\Sigma_m P_m(C/d_i))/k$
The KL divergence between two distributions $P_1(C)$ and $P_2(C)$ is:

$$D(P_1(C)||P_2(C)) = \Sigma_{j=1}^{|C|} P_1(c_j) log\left(\frac{P_1(c_j)}{P_2(c_j)}\right)$$

3.4 Combining SFE and Active Learning

Active learning can be combined with SFE by running SFE to convergence after actively selecting all the training data that will be labelled.

4 Experimental Results

Finally, this approach of Active Learning + SFE along with MNB as the classifier was tried upon the datasets RCV1-v2 made available through [1]. Out of the available datasets only 4 classes under the broad heading of ”market” were chosen for the experiment- M11, M12, M13 and M14. This meant that out of the available 8 lakh documents in the corpus, only around 50,000 were used. This set of 50,000 documents were divided into 3 parts. 2 were used for training, and the 3rd one was used for testing. A vocabulary of 47,236 words was used for all the runs. Experiments were conducted starting with different number of labelled examples i.e. labelled documents. Starting with 10, 100, 250 500 initially labelled documents. In each iteration, 10 labelling requests were allowed, and once these 10 labels are provided, they were incorporated in the labelled set, SFE was run, and finally it's accuracy on the training set was checked. The resulting plots are shown in the fig As can be seen from the plots, using active learning + SFE requires use of less number of labelled examples as compared to SFE alone. However, Active learning reduced this requirement for EM by a factor of nearly half in [2]. But Active Learning has been able to reduce this requirement for SFE only by a factor of one-fourth i.e. Active learning + SFE requires around 75 percent of labelled examples as required by SFE alone to reach the same accuracy.

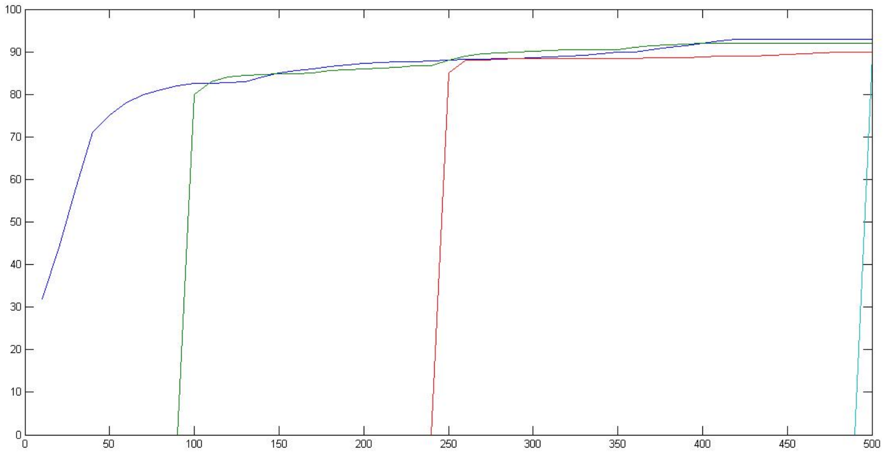

Figure 1: Plot showing classification accuracy vs the number of labelled documents used

Thus, we conclude that though active learning reduces the requirement for labelling to a good extent, but is not as fruitful as it had been for EM. A theoretical analysis into the reasons of the same can be pursued in future works.

References

[1] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li, *Rcv1: A new benchmark collection for text categorization research*, J. Mach. Learn. Res. **5** (2004), 361–397.

[2] Andrew McCallum and Kamal Nigam, *Employing em and pool-based active learning for text classification*, Proceedings of the Fifteenth International Conference on Machine Learning (San Francisco, CA, USA), ICML '98, Morgan Kaufmann Publishers Inc., 1998, pp. 350–358.

[3] Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell, *Learning to classify text from labeled and unlabeled documents*, Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence (Menlo Park, CA, USA), AAAI '98/IAAI '98, American Association for Artificial Intelligence, 1998, pp. 792–799.

[4] Jiang Su, Jelber Sayyad Shirab, and Stan Matwin, *Large scale text classification using semisupervised multinomial naive bayes*, in *ICML* [4], pp. 97–104.