Newspaper Article Classification Contest

Proposal

Terry Li *tianweil@andrew.cmu.edu*
Tao Yu *taoyu@andrew.cmu.edu*

Abstract

This is the proposal for the Newspaper Article Classification Contest project. The content includes how we retrieve the features from articles, which classifier we plan to use and how to train the classifier. We have plan A: Naïve Bayes Classifier and plan B SVM.

1 Project members

Terry Li (Andrew ID: tianweil) and Tao Yu (Andrew ID: taoyu)

2 Plan A: Naïve Bayes Classifier

Use naïve Bayes as the classifier and the frequency of each word happened in each class as the feature.

2.1 Feature Extraction and reduction

Follow the "Bag of words" instruction.

1. Parse the training articles and construct a dictionary for the words.
2. Remove the stop words.
3. Use the words as the features for classification

Feature Reduction: we plan to try three different feature reduction methods

1. Order the words according to their frequency. Instead of using the whole words in the dictionary, select a certain number of highest frequency.
2. Select the words that have the highest mutual information. If one word that has the same probability happened in each class, then it is a noise feature. Noise feature should be removed.
3. Latent Semantic Analysis (LSA). Singular value decomposition (latent semantics) approach which (as far as we understand) needs me to construct that matrix in the first place to calculate the SVD matrix. I found some research of people doing this and was wondering if there was a trick not needing to construct the full matrix I didn't get until now.

2.2 Training Classifier

Find out the frequency of each word happened in each class. (For each class, count the number of a certain word divided by the total number of words). This is the MLE estimation of the probability of each words happened in each class ($p(x_i|y)$).

Then find the MLE probability for each class $p(y)$.

Then for the new articles, we can calculate the $\log(p(y) * \prod p(x_i|y))$ for each class y. And select the class that has the max result.

2.2 Against Overfitting

Using Cross Validation to fight against overfitting. Separate the training set into small training set and test set. Then repeatedly trains the classifier.

3 Plan B (SVM)

3.1 Feature Extraction/Selection

Each distinct word w_i corresponds to a feature, with the number of times word w_i occurs in the document as its value. (words are considered as features only if they occur in the training data at least 3 times except the "stop-words")

Information gain criterion will be used to select a subset of features. Meanwhile, we will scale the dimensions of the feature vector with inverse document frequency(IDF). "tfc" variant is used and each document feature vector will be normalized to unit length.

3.2 Classifier Selection

We plan to try the Support Vector Machines, which prove to be very robust, eliminating the need for expensive parameter tuning.

Reasons: SVM uses overfitting protection, it has the potential to handle large feature spaces. SVM's performance is good when there are few irrelevant features, document vectors are sparse and text categorization problem is linearly separable.

3.3 Training Classifier

We will use simple linear SVMs because they provide good generalization accuracy and because they are faster to learn. We plan to use Platt's Sequential Minimal Optimization (SMO) method to learn the vector of feature weights $\vec{w}$. After the weights are learned, new items can be classified by computing $\vec{w} \cdot \vec{x}$ $\vec{x}$ is the binary vector representing the new document to classify. We can also learn two parameters of a sigmoid function to transform the output of the SVM to probabilities.

References

[1] http://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

[2] http://nlp.stanford.edu/IR-book/html/htmledition/feature-selection-1.html

[3] Joachims, Thorsten. Text categorization with support vector machines: Learning with many relevant features. Springer Berlin Heidelberg, 1998.

[4] Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." ICML. Vol. 97. 1997.