Multi-Label Text Classification

Huangxin Wang, Zhonghua Xi[*]

Jul 18, 2014

1 Problem Statement

2 Algorithm

2.1 ReWeight

We use a variant of TF.IDF model to reweight the word vector of each document. The TF.IDF variants is defined as:

$$w_{t,d} = log(tf_{t,d} + 1) * idf_t$$

where $w_{t,d}$ is the reweight word frequency, $tf_{t,d}$ is the original frequency of term t in document d, and idf_t is inverse document frequency which is defined as follows,

$$idf_t = \frac{|D|}{df_t}$$

where $|D|$ is the number of documents in the training set, and df_t is number of documents contains term t.

2.2 Build Classifiers

We build classifier for each class, the classifier is calculated by using the average of each instances in the class.

2.3 Similarity Calculation

We calculate the similarity of a test instance with the classifier using the cosine similarity.

$$cossim(d_i, d_j) = \frac{d_i * d_j}{|d_i| * |d_j|} = \frac{\sum_{k=1}^{N} w_{k,i} * w_{k,j}}{\sqrt{\sum_{k=1}^{N} w_{k,i}^2 * \sqrt{\sum_{k=1}^{N} w_{k,j}^2}}}$$

[*] Department of Computer Science, George Mason University. Fairfax, VA 22030. Email: hwang14@gmu.edu, zxi@gmu.edu

where d_i is feature vector of document i and d_j is the feature vector of class j. We denote $cossim(d_i, d_j)$ as the score document i achieves in class j.

2.4 Multi-Label Classify

First Label The first label document i assigned is the class in which it gets the highest score.

Second Label Similar to the chosen of label 1, for the second label, we choose the class have the second highest score of document i. However, we accept the second label only when $score(label_1)/score(label_2) \leq \alpha$, which means document i have almost the same degree of similarity to label 2 as to label 1.