An Empirical Study of Feature Selection for Text Categorization based on Term Weightage

Bong Chih How, Narayanan K.
Faculty of Computer Science and Information Technology,
Universiti Malaysia Sarawak,
Kota Samarahan 94300, Sarawak, Malaysia
chbong@fit.unimas.my, nara@fit.unimas.my

Abstract

This paper proposes a local feature selection (FS) measure namely, Categorical Descriptor Term (CTD) for text categorization. It is derived based on classic term weighting scheme, TFIDF. The method explicitly chooses feature set for each category by only selecting set of terms from relevant category. Although past literatures have suggested that the use of features from irrelevant categories can improve the measure of text categorization, we believe that by incorporating only relevant feature can be highly effective. The experimental comparison is carried out between CTD and five well-known feature selection measures: Information Gain, Chi-Square, Correlation Coefficient, Odd Ratio and GSS Coefficient. The results also show that our proposed method can perform comparatively well with other FS measures, especially on collection with highly overlapped topics.

1. Introduction

Text categorization (TC) involves the assignment of categories to natural language documents based on the assessment of their contents. Traditionally, term weighting scheme employs the bag-of-words approach in document indexing in the area of information retrieval (IR). It has been used to weight representative term used to describe and summarize document content based on a term's importance.

Term Frequency Inversed Document Frequency, or abbreviately known as TFIDF is one of the most popular term weighting schemes in IR [1]. TFIDF assumes that "multiple appearances of a term in a document are more important than single appearances" and "rare terms are more important than frequent terms". It has gained popularity in text categorization assuming that the index terms are mutually independent.

Feature selection (FS) is generally used for the purpose of dimensionality reduction. FS identifies the subset of the original set of terms that effectively characterizes the categories. Each term is weighted and scored for its "importance" based on its entropy or overall collection statistics. In a vector space model with r terms, n top ranked terms, $r`$, are chosen, where $r`$<=r. Usually, $r`$ consists of terms from both relevant (positive) and irrelevant (negative) categories. Particular FS measures have been used to reduce vector space up to 90% [4].

Generally, FS can be based on cluster analysis, distance measures, statistical or information-theoretic based. FS can be performed in two ways depending on the particular task. For localized FS, a set of terms is defined based on the relevant and irrelevant documents in the category. Each category is represented by a set of unique terms. On the other hand, in globalized FS, a set of terms is chosen for all the categories. Study has proven that globalilzed FS tends to generate better performance when compared to localized FS. This is because localized FS does not effectively collect categories' characteristics due to the presence of scarcely populated categories in the collection [9].

Research on incorporating of terms from both relevant and irrelevant categories also attracts some attentions especially in the area of localized FS [8][10]. Feature terms can be constructed either from relevant categories (local dictionary) or from both relevant and irrelevant categories (universal dictionary) [11]. Although literature has emphasized that irrelevant features can be useful to build evidence in negating irrelevant document, we believe by incorporating only terms from relevant category can be highly effective.

This paper presents a newly derived term weighting scheme, which is used to explicitly select feature terms solely on the relevant categories. We validate its performance with five standard feature selection measures, which are described in the next section.

Table 1 Feature selection measures

Description	Formula	References
Information Gain	$IG(t_k,c_i)=\sum_{c\in\{c_i,\bar{c}_i\}}\sum_{t\in\{t_k,\bar{t}_k\}}P(t,c)\log_2\frac{P(t,c)}{P(t)P(c)}$	[4][7][9]
Chi-Square	$CHI(t_k,c_i)=\frac{N[P(t_k,c_i)\cdot P(\overline{t}_k,\overline{c}_i)-P(t_k,\overline{c}_i)\cdot P(\overline{t}_k,c_i)]^2}{P(t_k)\cdot P(\overline{t}_k)\cdot P(c_i)\cdot P(\overline{c}_i)}$	[4][8][9][12]
Corelated Coefficient	$CC(t_k,c_i)=\frac{\sqrt{N}[P(t_k,c_i)\cdot P(\overline{t}_k,\overline{c}_i)-P(t_k,\overline{c}_i)\cdot P(\overline{t}_k,c_i)]}{\sqrt{P(t_k)\cdot P(\overline{t}_k)\cdot P(c_i)\cdot P(\overline{c}_i)}}$	[8]
Odd Ratio	$OR(t_k,c_i)=\frac{P(t_k\mid c_i)\cdot[1-P(t_k\mid\overline{c}_i)]}{[1-P(t_k\mid c_i)]\cdot P(t_k\mid\overline{c}_i)}$	[8]
GSS Coefficient	$GSS(t_k,c_i)=P(t_k,c_i)\cdot P(\overline{t}_k,\overline{c}_i)-P(t_k,\overline{c}_i)\cdot P(\overline{t}_k,c_i)$	[5][8]

2. Related Work

In this section, we present five commonly known feature selection (see Table 1) measures we evaluated. In the interest of brevity, we have omitted their mathematical justification.

Subsequently, we proposed our term weighting scheme as feature selection measure which only incorporates feature term from relevant category.

2.6. Category Term Descriptor (CTD)

We adopted TFIDF in a category perspective instead of document perspective. We have TFICF, where TF refers to term frequency in category *c* and ICF is interpreted as inverse category frequency or more accurately, relative frequency of a term in category *c*. However, TFICF scheme shows no way of discriminating between terms that occur frequently in a small subset of documents and terms that are present in a large number of documents throughout a category. The formula scores the weights equally throughout the documents in the specific category with no bearing of term discriminative power among documents. According to the ICF formula, these terms will not be treated differently. Thus, we believe that the lesser a term occurs across documents, the higher is its discriminative power. The factor IDF(*t*) in TFIDF can be applied to normalize the weight. Term's discriminative capability among documents in a category is relatively important to represent the category. Hence, we defined our method

$CTD(t_k,c_i)=TF(t_k,c_i)\cdot IDF(t_k,c_i)\cdot ICF(t_k)$, where

$$ICF(t_k)=\log\left(\frac{|C|}{CF(t_k)}\right),\; IDF(t_k,c_i)=\log\left(\frac{|D(c_i)|}{DF(t_k,c_i)}\right)$$

where

$D(c_i)$ is the number of document in category c_i

C is number of category in the collection

$CF(t_k)$ is the category frequency for term t_k

$DF(t_k,c_i)$ is the document frequency for term t_k in category c_i

2.5. Multinomial Naive Bayes (MNB) Classifier

MNB has been considered an alternative to model text representation with bag of words. A term is captured with its value indicating the frequency of occurrence of a particular term in the document. It assumes that a term that occurs many times in a particular category will have more influence on future classification than a term that makes few appearances. In our experiment, we assume there are *c* models, one for each category. Thus, we only consider binary classification employing MMB in this study.

3. Experimental Setting

To verify our study, we performed the FS measures on three distinct text collections, namely Reuters-21578[1], 20 newsgroup[2] and conference proceedings from Technology and Teacher Education Annual[3]. We have experimented comparative approaches on both benchmark (Reuters) and operational datasets.

3.1. Reuters-21578

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/

[2] http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html

[3] http://www.coe.uh.edu/insite/elec_pub