A Framework of Feature Selection Methods for Text Categorization

Shoushan Li[1] **Rui Xia**[2] **Chengqing Zong**[2] **Chu-Ren Huang**[1]

[1] Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
{shoushan.li,churenhuang}@gmail.com

[2] National Laboratory of Pattern Recognition
Institute of Automation
Chinese Academy of Sciences
{rxia,cqzong}@nlpr.ia.ac.cn

Abstract

In text categorization, feature selection (FS) is a strategy that aims at making text classifiers more efficient and accurate. However, when dealing with a new task, it is still difficult to quickly select a suitable one from various FS methods provided by many previous studies. In this paper, we propose a theoretic framework of FS methods based on two basic measurements: frequency measurement and ratio measurement. Then six popular FS methods are in detail discussed under this framework. Moreover, with the guidance of our theoretical analysis, we propose a novel method called weighed frequency and odds (WFO) that combines the two measurements with trained weights. The experimental results on data sets from both topic-based and sentiment classification tasks show that this new method is robust across different tasks and numbers of selected features.

1 Introduction

With the rapid growth of online information, text classification, the task of assigning text documents to one or more predefined categories, has become one of the key tools for automatically handling and organizing text information.

The problems of text classification normally involve the difficulty of extremely high dimensional feature space which sometimes makes learning algorithms intractable. A standard procedure to reduce the feature dimensionality is called feature selection (FS). Various FS methods, such as document frequency (DF), information gain (IG), mutual information (MI), χ^2 -test (CHI), Bi-Normal Separation (BNS), and weighted log-likelihood ratio (WLLR), have been proposed for the tasks (Yang and Pedersen, 1997; Nigam et al., 2000; Forman, 2003) and make text classification more efficient and accurate.

However, comparing these FS methods appears to be difficult because they are usually based on different theories or measurements. For example, MI and IG are based on information theory, while CHI is mainly based on the measurements of statistic independence. Previous comparisons of these methods have mainly depended on empirical studies that are heavily affected by the experimental sets. As a result, conclusions from those studies are sometimes inconsistent. In order to better understand the relationship between these methods, building a general theoretical framework provides a fascinating perspective.

Furthermore, in real applications, selecting an appropriate FS method remains hard for a new task because too many FS methods are available due to the long history of FS studies. For example, merely in an early survey paper (Sebastiani, 2002), eight methods are mentioned. These methods are provided by previous work for dealing with different text classification tasks but none of them is shown to be robust across different classification applications.

In this paper, we propose a framework with two basic measurements for theoretical comparison of six FS methods which are widely used in text classification. Moreover, a novel method is set forth that combines the two measurements and tunes their influences considering different application domains and numbers of selected features.

The remainder of this paper is organized as follows. Section 2 introduces the related work on

Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 692–700,
Suntec, Singapore, 2-7 August 2009.

feature selection for text classification. Section 3 theoretically analyzes six FS methods and proposes a new FS approach. Experimental results are presented and analyzed in Section 4. Finally, Section 5 draws our conclusions and outlines the future work.

2 Related Work

FS is a basic problem in pattern recognition and has been a fertile field of research and development since the 1970s. It has been proven to be effective on removing irrelevant and redundant features, increasing efficiency in learning tasks, and improving learning performance.

FS methods fall into two broad categories, the filter model and the wrapper model (John et al., 1994). The wrapper model requires one predetermined learning algorithm in feature selection and uses its performance to evaluate and determine which features are selected. And the filter model relies on general characteristics of the training data to select some features without involving any specific learning algorithm. There is evidence that wrapper methods often perform better on small scale problems (John et al, 1994), but on large scale problems, such as text classification, wrapper methods are shown to be impractical because of its high computational cost. Therefore, in text classification, filter methods using feature scoring metrics are popularly used. Below we review some recent studies of feature selection on both topic-based and sentiment classification.

In the past decade, FS studies mainly focus on topic-based classification where the classification categories are related to the subject content, e.g., sport or education. Yang and Pedersen (1997) investigate five FS metrics and report that good FS methods improve the categorization accuracy with an aggressive feature removal using DF, IG, and CHI. More recently, Forman (2003) empirically compares twelve FS methods on 229 text classification problem instances and proposes a new method called 'Bi-Normal Separation' (BNS). Their experimental results show that BNS can perform very well in the evaluation metrics of recall rate and F-measure. But for the metric of precision, it often loses to IG. Besides these two comparison studies, many others contribute to this topic (Yang and Liu, 1999; Brank et al., 2002; Gabrilovich and Markovitch, 2004) and more and more new FS methods are generated, such as, Gini index (Shang et al., 2007), Distance to Transition Point (DTP) (Moyotl-Hernandez and Jimenez-Salazar, 2005), Strong Class Information Words (SCIW) (Li and Zong, 2005) and parameter tuning based FS for Rocchio classifier (Moschitti, 2003).

Recently, sentiment classification has become popular because of its wide applications (Pang et al., 2002). Its criterion of classification is the attitude expressed in the text (e.g., recommended or not recommended, positive or negative) rather than some facts (e.g., sport or education). To our best knowledge, yet no related work has focused on comparison studies of FS methods on this special task. There are only some scattered reports in their experimental studies. Riloff et al. (2006) report that the traditional FS method (only using IG method) performs worse than the baseline in some cases. However, Cui et al. (2006) present the experiments on the sentiment classification for large-scale online product reviews to show that using the FS method of CHI does not degrade the performance but can significantly reduce the dimension of the feature vector.

Moreover, Ng et al. (2006) examine the FS of the weighted log-likelihood ratio (WLLR) on the movie review dataset and achieves an accuracy of 87.1%, which is higher than the result reported by Pang and Lee (2004) with the same dataset. From the analysis above, we believe that the performance of the sentiment classification system is also dramatically affected by FS.

3 Our Framework

In the selection process, each feature (term, or single word) is assigned with a score according to a score-computing function. Then those with higher scores are selected. These mathematical definitions of the score-computing functions are often defined by some probabilities which are estimated by some statistic information in the documents across different categories. For the convenience of description, we give some notations of these probabilities below.

$P(t)$: *the probability that a document* x *contains term* t *;*

$P(\overline{c_i})$: *the probability that a document* x *does not belong to category* c_i *;*

$P(t, c_i)$: *the joint probability that a document* x *contains term* t *and also belongs to category* c_i *;*

$P(c_i \mid t)$: *the probability that a document x belongs to category* c_i *, under the condition that it contains term t.*

$P(\bar{t} \mid c_i)$: *the probability that, a document x does not contain term t with the condition that x belongs to category* c_i ;

Some other probabilities, such as $P(\bar{t})$, $P(c_i)$, $P(t \mid c_i)$, $P(t \mid \overline{c_i})$, $P(c_i \mid \bar{t})$, and $P(\overline{c_i} \mid t)$, are similarly defined.

In order to estimate these probabilities, statistical information from the training data is needed, and notations about the training data are given as follows:

$\{c_i\}_{i=1}^{m}$ *: the set of categories;*

A_i *: the number of the documents that contain the term* t *and also belong to category* c_i ;

B_i *: the number of the documents that contain the term* t *but do not belong to category* c_i ;

N_i *: the total number of the documents that belong to category* c_i ;

N_{all} *: the total number of all documents from the training data.*

C_i *: the number of the documents that do not contain the term* t *but belong to category* c_i *, i.e.,* $N_i - A_i$

D_i *: the number of the documents that neither contain the term* t *nor belong to category* c_i *, i.e.,* $N_{all} - N_i - B_i$;

In this section, we would analyze theoretically six popular methods, namely DF, MI, IG, CHI, BNS, and WLLR. Although these six FS methods are defined differently with different scoring measurements, we believe that they are strongly related. In order to connect them, we define two basic measurements which are discussed as follows.

The first measurement is to compute the document frequency in one category, i.e., A_i .

The second measurement is the ratio between the document frequencies in one category and the other categories, i.e., A_i / B_i . The terms with a high ratio are often referred to as the terms with high category information.

These two measurements form the basis for all the measurements that are used by the FS methods throughout this paper. In particular, we show that DF and MI are using the first and second measurement respectively. Other complicated FS methods are combinations of these two measurements. Thus, we regard the two measurements as basic, which are referred to as the *frequency measurement* and *ratio measurement*.

3.1 Document Frequency (DF)

DF is the number of documents in which a term occurs. It is defined as

$$DF = \sum_{i=1}^{m} (A_i)$$

The terms with low or high document frequency are often referred to as rare or common terms, respectively. It is easy to see that this FS method is based on the first basic measurement. It assumes that the terms with higher document frequency are more informative for classification. But sometimes this assumption does not make any sense, for example, the stop words (e.g., the, a, an) hold very high DF scores, but they seldom contribute to classification. In general, this simple method performs very well in some topic-based classification tasks (Yang and Pedersen, 1997).

3.2 Mutual Information (MI)

The mutual information between term t and class c_i is defined as

$$I(t, c_i) = \log \frac{P(t \mid c_i)}{P(t)}$$

And it is estimated as

$$MI = \log \frac{A_i \times N_{all}}{(A_i + C_i)(A_i + B_i)}$$

Let us consider the following formula (using Bayes theorem)

$$I(t, c_i) = \log \frac{P(t \mid c_i)}{P(t)} = \log \frac{P(c_i \mid t)}{P(c_i)}$$

Therefore,

$$I(t, c_i) = \log P(c_i \mid t) - \log P(c_i)$$

And it is estimated as

$$\begin{aligned} MI &= \log \frac{A_i}{A_i + B_i} - \log \frac{N_i}{N_{all}} \\ &= -\log \frac{A_i + B_i}{A_i} - \log \frac{N_i}{N_{all}} \\ &= -\log(1 + \frac{1}{A_i / B_i}) - \log \frac{N_i}{N_{all}} \end{aligned}$$

From this formula, we can see that the MI score is based on the second basic measurement. This method assumes that the term with higher category ratio is more effective for classification.

It is reported that this method is biased towards low frequency terms and the bias becomes extreme when $P(t)$ is near zero. It can be seen in the following formula (Yang and Pedersen, 1997)

$$I(t, c_i) = \log(P(t \mid c_i)) - \log(P(t))$$

Therefore, this method might perform badly when common terms are informative for classification.

Taking into account mutual information of all categories, two types of MI score are commonly used: the maximum score $I_{max}(t)$ and the average score $I_{avg}(t)$, i.e.,

$$I_{max}(t) = \max_{i=1}^{m}\{I(t,c_i)\},$$

$$I_{avg}(t) = \sum_{i=1}^{m} P(c_i)\cdot I(t,c_i).$$

We choose the maximum score since it performs better than the average score (Yang and Pedersen, 1997). It is worth noting that the same choice is made for other methods, including CHI, BNS, and WLLR in this paper.

3.3 Information Gain (IG)

IG measures the number of bits of information obtained for category prediction by recognizing the presence or absence of a term in a document (Yang and Pedersen, 1997). The function is

$$\begin{aligned} G(t) = &\{-\sum_{i=1}^{m} P(c_i)\log P(c_i)\} \\ &+\{P(t)[\sum_{i=1}^{m} P(c_i \mid t)\log P(c_i \mid t)] \\ &+ P(\bar{t})[\sum_{i=1}^{m} P(c_i \mid \bar{t})\log P(c_i \mid \bar{t})]\} \end{aligned}$$

And it is estimated as

$$\begin{aligned} IG = &\{-\sum_{i=1}^{m}\frac{N_i}{N_{all}}\log\frac{N_i}{N_{all}}\} \\ &+(\sum_{i=1}^{m} A_i / N_{all})[\sum_{i=1}^{m}\frac{A_i}{A_i+B_i}\log\frac{A_i}{A_i+B_i}] \\ &+(\sum_{i=1}^{m} C_i / N_{all})[\sum_{i=1}^{m}\frac{C_i}{C_i+D_i}\log\frac{C_i}{C_i+D_i}] \end{aligned}$$

From the definition, we know that the information gain is the weighted average of the mutual information $I(t,c_i)$ and $I(\bar{t},c_i)$ where the weights are the joint probabilities $P(t,c_i)$ and $P(\bar{t},c_i)$:

$$G(t) = \sum_{i=1}^{m} P(t,c_i)I(t,c_i) + \sum_{i=1}^{m} P(\bar{t},c_i)I(\bar{t},c_i)$$

Since $P(t,c_i)$ is closely related to the document frequency A_i and the mutual information $I(t,c_i)$ is shown to be based on the second measurement, we can say that the IG score is influenced by the two basic measurements.

3.4 χ^2 Statistic (CHI)

The CHI measurement (Yang and Pedersen, 1997) is defined as

$$CHI = \frac{N_{all}\cdot(A_iD_i - C_iB_i)^2}{(A_i+C_i)\cdot(B_i+D_i)\cdot(A_i+B_i)\cdot(C_i+D_i)}$$

In order to get the relationship between CHI and the two measurements, the above formula is rewritten as follows

$$CHI = \frac{N_{all}\cdot[A_i(N_{all}-N_i-B_i)-(N_i-A_i)B_i]^2}{N_i\cdot(N_{all}-N_i)\cdot(A_i+B_i)\cdot[N_{all}-(A_i+B_i)]}$$

For simplicity, we assume that there are two categories and the numbers of the training documents in the two categories are the same ($N_{all} = 2N_i$). The CHI score then can be written as

$$\begin{aligned} CHI &= \frac{2N_i(A_i-B_i)^2}{(A_i+B_i)\cdot[2N_i-(A_i+B_i)]} \\ &= \frac{2N_i(A_i/B_i-1)^2}{(A_i/B_i+1)\cdot[\frac{2N_i}{A_i}\cdot A_i/B_i-(A_i/B_i+1)]} \end{aligned}$$

From the above formula, we see that the CHI score is related to both the frequency measurement A_i and ratio measurement A_i / B_i. Also, when keeping the same ratio value, the terms with higher document frequencies will yield higher CHI scores.

3.5 Bi-Normal Separation (BNS)

BNS method is originally proposed by Forman (2003) and it is defined as

$$BNS(t,c_i) = \left|F^{-1}(P(t \mid c_i)) - F^{-1}(P(t \mid \bar{c_i}))\right|$$

It is calculated using the following formula

$$BNS = \left|F^{-1}(\frac{A_i}{N_i}) - F^{-1}(\frac{B_i}{N_{all}-N_i})\right|$$

where $F(x)$ is the cumulative probability function of standard normal distribution.

For simplicity, we assume that there are two categories and the numbers of the training documents in the two categories are the same, i.e., $N_{all} = 2N_i$ and we also assume that $A_i > B_i$. It should be noted that this assumption is only to allow easier analysis but will not be applied in our experiment implementation. In addition, we only consider the case when $A_i / N_i \le 0.5$. In fact, most terms take the document frequency A_i which is less than half of N_i.

Under these conditions, the BNS score can be shown in Figure 1 where the area of the shadow part represents $(A_i / N_i - B_i / N_i)$ and the length of the projection to the x axis represents the BNS score.

From Figure 1, we can easily draw the two following conclusions:

1) Given the same value of A_i, the BNS score increases with the increase of $A_i - B_i$.
2) Given the same value of $A_i - B_i$, BNS score increase with the decrease of A_i.

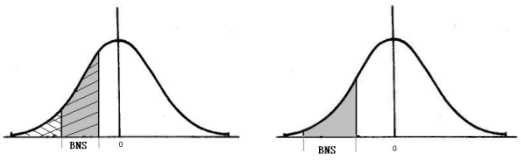

Figure 1. View of BNS using the normal probability distribution. Both the left and right graphs have shadowed areas of the same size.

And the value of $A_i - B_i$ can be rewritten as the following

$$A_i - B_i = \frac{A_i - B_i}{A_i} \cdot A_i = (1 - \frac{1}{A_i / B_i}) \cdot A_i$$

The above analysis gives the following conclusions regarding the relationship between BNS and the two basic measurements:

1) Given the same A_i, the BNS score increases with the increase of A_i / B_i.
2) Given the same A_i / B_i, when A_i increases, $A_i - B_i$ also increase. It seems that the BNS score does not show a clear relationship with A_i.

In summary, the BNS FS method is biased towards the terms with the high category ratio but cannot be said to be sensitive to document frequency.

3.6 Weighted Log Likelihood Ratio (WLLR)

WLLR method (Nigam et al., 2000) is defined as

$$WLLR(t, c_i) = P(t \mid c_i) \log \frac{P(t \mid c_i)}{P(t \mid \overline{c_i})}$$

And it is estimated as

$$WLLR = \frac{A_i}{N_i} \log \frac{A_i \cdot (N_{all} - N_i)}{B_i \cdot N_i}$$

The formula shows WLLR is proportional to the frequency measurement and the logarithm of the ratio measurement. Clearly, WLLR is biased towards the terms with both high category ratio and high document frequency and the frequency measurement seems to take a more important place than the ratio measurement.

3.7 Weighed Frequency and Odds (WFO)

So far in this section, we have shown that the two basic measurements constitute the six FS methods. The class prior probabilities, $P(c_i)$, $i = 1, 2, ..., m$, are also related to the selection methods except for the two basic measurements. Since they are often estimated according to the distribution of the documents in the training data and are identical for all the terms in a class, we ignore the discussion of their influence on the selection measurements. In the experiment, we consider the case when training data have equal class prior probabilities. When training data are unbalanced, we need to change the forms of the two basic measurements to A_i / N_i and $A_i \cdot (N_{all} - N_i) / (B_i \cdot N_i)$.

Because some methods are expressed in complex forms, it is difficult to explain their relationship with the two basic measurements, for example, which one prefers the category ratio most. Instead, we will give the preference analysis in the experiment by analyzing the features in real applications. But the following two conclusions are drawn without doubt according to the theoretical analysis given above.

1) Good features are features with high document frequency;
2) Good features are features with high category ratio.

These two conclusions are consistent with the original intuition. However, using any single one does not provide competence in selecting the best set of features. For example, stop words, such as 'a', 'the' and 'as', have very high document frequency but are useless for the classification. In real applications, we need to mix these two measurements to select good features. Because of different distribution of features in different domains, the importance of each measurement may differ a lot in different applications. Moreover, even in a given domain, when different numbers of features are to be selected, different combinations of the two measurements are required to provide the best performance.

Although a great number of FS methods is available, none of them can appropriately change the preference of the two measurements. A better way is to tune the importance according to the application rather than to use a predetermined combination. Therefore, we propose a new FS method called Weighed Frequency and Odds (WFO), which is defined as

when $P(t \mid c_i) / P(t \mid \overline{c_i}) > 1$

$$WFO(t, c_i) = P(t \mid c_i)^{\lambda} [\log \frac{P(t \mid c_i)}{P(t \mid \overline{c_i})}]^{1-\lambda}$$

else

$$WFO(t, c_i) = 0$$

And it is estimated as

$$WFO = (\frac{A_i}{N_i})^{\lambda} (\log \frac{A_i \cdot (N_{all} - N_i)}{B_i \cdot N_i})^{1-\lambda}$$

where λ is the parameter for tuning the weight between frequency and odds. The value of λ varies from 0 to 1. By assigning different value to λ we can adjust the preference of each measurement. Specially, when $\lambda = 0$, the algorithm prefers the category ratio that is equivalent to the MI method; when $\lambda = 1$, the algorithm is similar to DF; when $\lambda = 0.5$, the algorithm is exactly the WLLR method. In real applications, a suitable parameter λ needs to be learned by using training data.

4 Experimental Studies

4.1 Experimental Setup

Data Set: The experiments are carried out on both topic-based and sentiment text classification datasets. In topic-based text classification, we use two popular data sets: one subset of Reuters-21578 referred to as R2 and the 20 Newsgroup dataset referred to as 20NG. In detail, R2 consist of about 2,000 2-category documents from standard corpus of Reuters-21578. And 20NG is a collection of approximately 20,000 20-category documents [1]. In sentiment text classification, we also use two data sets: one is the widely used Cornell movie-review dataset[2] (Pang and Lee, 2004) and one dataset from product reviews of domain DVD[3] (Blitzer et al., 2007). Both of them are 2-category tasks and each consists of 2,000 reviews. In our experiments, the document numbers of all data sets are (nearly) equally distributed cross all categories.

Classification Algorithm: Many classification algorithms are available for text classification, such as Naïve Bayes, Maximum Entropy, k-NN, and SVM. Among these methods, SVM is shown to perform better than other methods (Yang and Pedersen, 1997; Pang et al., 2002). Hence we apply SVM algorithm with the help of the LIBSVM [4] tool. Almost all parameters are set to their default values except the kernel function which is changed from a polynomial kernel function to a linear one because the linear one usually performs better for text classification tasks.

Experiment Implementation: In the experiments, each dataset is randomly and evenly split into two subsets: 90% documents as the training data and the remaining 10% as testing data. The training data are used for training SVM classifiers, learning parameters in WFO method and selecting "good" features for each FS method. The features are single words with a bool weight (0 or 1), representing the presence or absence of a feature. In addition to the "principled" FS methods, terms occurring in less than three documents ($DF \leq 3$) in the training set are removed.

4.2 Relationship between FS Methods and the Two Basic Measurements

To help understand the relationship between FS methods and the two basic measurements, the empirical study is presented as follows.

Since the methods of DF and MI only utilize the document frequency and category information respectively, we use the DF scores and MI scores to represent the information of the two basic measurements. Thus we would select the top-2% terms with each method and then investigate the distribution of their DF and MI scores.

First of all, for clear comparison, we normalize the scores coming from all the methods using Min-Max normalization method which is designed to map a score s to s' in the range [0, 1] by computing

$$s' = \frac{s - Min}{Max - Min}$$

where *Min* and *Max* denote the minimum and maximum values respectively in all terms' scores using one FS method.

Table 1 shows the mean values of all top-2% terms' MI scores and DF scores of all the six FS methods in each domain. From this table, we can apparently see the relationship between each method and the two basic measurements. For instance, BNS most distinctly prefers the terms with high MI scores and low DF scores. According to the degree of this preference, we

[1] http://people.csail.mit.edu/~jrennie/20Newsgroups/

[2] http://www.cs.cornell.edu/People/pabo/movie-review-data/

[3] http://www.seas.upenn.edu/~mdredze/datasets/sentiment/

[4] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

FS Methods	Domain							
	20NG		R2		Movie		DVD	
	DF score	MI score	DF score	MI score	DF score	MI score	DF score	MI score
MI	0.004	0.870	0.047	0.959	0.003	0.888	0.004	0.881
BNS	0.005	0.864	0.117	0.922	0.008	0.881	0.006	0.880
CHI	0.015	0.814	0.211	0.748	0.092	0.572	0.055	0.676
IG	0.087	0.525	0.209	0.792	0.095	0.559	0.066	0.669
WLLR	0.026	0.764	0.206	0.805	0.168	0.414	0.127	0.481
DF	0.122	0.252	0.268	0.562	0.419	0.09	0.321	0.111

Table 1. The mean values of all top-2% terms' MI and DF scores using six FS methods in each domain

can rank these six methods as $\text{MI, BNS} \succ \text{IG, CHI, WLLR} \succ \text{DF}$, where $x \succ y$ means method x prefers the terms with higher MI scores (higher category information) and lower DF scores (lower document frequency) than method y. This empirical discovery is in agreement with the finding that WLLR is biased towards the high frequency terms and also with the finding that BNS is biased towards high category information (cf. **Section 3** theoretical analysis). Also, we can find that CHI and IG share a similar preference of these two measurements in 2-category domains, i.e., R2, movie, and DVD. This gives a good explanation that CHI and IG are two similar-performed methods for 2-category tasks, which have been found by Forman (2003) in their experimental studies.

According to the preference, we roughly cluster FS methods into three groups. The first group includes the methods which dramatically prefer the category information, e.g., MI and BNS; the second one includes those which prefer both kinds of information, e.g., CHI, IG, and WLLR; and the third one includes those which strongly prefer frequency information, e.g., DF.

4.3 Performances of Different FS Methods

It is worth noting that learning parameters in WFO is very important for its good performance. We use 9-fold cross validation to help learning the parameter λ so as to avoid over-fitting. Specifically, we run nine times by using every 8 fold documents as a new training data set and the remaining one fold documents as a development data set. In each running with one fixed feature number m, we get the best $\lambda_{i,m-best}$ (i=1,..., 9) value through varying $\lambda_{i,m}$ from 0 to 1 with the step of 0.1 to get the best performance in the development data set. The average value λ_{m-best}, i.e.,

$$\lambda_{m-best} = (\sum_{i=1}^{9} \lambda_{i,m-best}) / 9$$

is used for further testing.

Figure 2 shows the experimental results when using all FS methods with different selected feature numbers. The red line with star tags represents the results of WFO. At the first glance, in R2 domain, the differences of performances across all are very noisy when the feature number is larger than 1,000, which makes the comparison meaningless. We think that this is because the performances themselves in this task are very high (nearly 98%) and the differences between two FS methods cannot be very large (less than one percent). Even this, WFO method do never get the worst performance and can also achieve the top performance in about half times, e.g., when feature numbers are 20, 50, 100, 500, 3000.

Let us pay more attention to the other three domains and discuss the results in the following two cases.

In the first case when the feature number is low (about less than 1,000), the FS methods in the second group including IG, CHI, WLLR, always perform better than those in the other two groups. WFO can also perform well because its parameters λ_{m-best} are successfully learned to be around 0.5, which makes it consistently belong to the second group. Take 500 feature number for instance, the parameters $\lambda_{500-best}$ are 0.42, 0.50, and 0.34 in these three domains respectively.

In the second case when the feature number is large, among the six traditional methods, MI and BNS take the leads in the domains of 20NG and Movie while IG and CHI seem to be better and more stable than others in the domain of DVD. As for WFO, its performances are excellent cross all these three domains and different feature numbers. In each domain, it performs similarly as or better than the top methods due to its well-learned parameters. For example, in 20NG, the parameters λ_{m-best} are 0.28, 0.20, 0.08, and 0.01 when feature numbers are 10,000, 15,000, 20,000, and 30,000. These values are close to 0

(WFO equals MI when $\lambda = 0$) while MI is the top one in this domain.

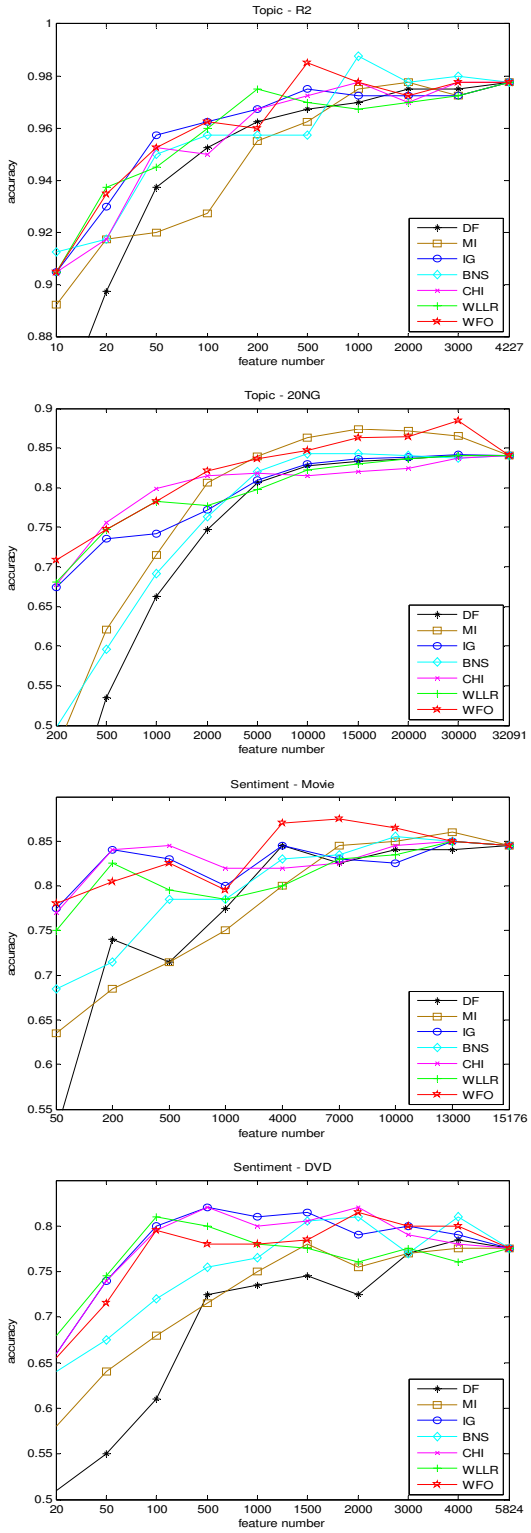

Figure 2. The classification accuracies of the four domains using seven different FS methods while increasing the number of selected features.

From Figure 2, we can also find that FS does help sentiment classification. At least, it can dramatically decrease the feature numbers without losing classification accuracies (see Movie domain, using only 500-4000 features is as good as using all 15176 features).

5 Conclusion and Future Work

In this paper, we propose a framework with two basic measurements and use it to theoretically analyze six FS methods. The differences among them mainly lie in how they use these two measurements. Moreover, with the guidance of the analysis, a novel method called WFO is proposed, which combine these two measurements with trained weights. The experimental results show that our framework helps us to better understand and compare different FS methods. Furthermore, the novel method WFO generated from the framework, can perform robustly across different domains and feature numbers.

In our study, we use four data sets to test our new method. There are much more data sets on text categorization which can be used. In additional, we only focus on using balanced samples in each category to do the experiments. It is also necessary to compare the FS methods on some unbalanced data sets, which are common in real-life applications (Forman, 2003; Mladeni and Marko, 1999). These matters will be dealt with in the future work.

Acknowledgments

The research work described in this paper has been partially supported by Start-up Grant for Newly Appointed Professors, No. 1-BBZM in the Hong Kong Polytechnic University.

References

J. Blitzer, M. Dredze, and F. Pereira. 2007. Biographies, Bollywood, Boom-boxes and Blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL-07, the 45th Meeting of the Association for Computational Linguistics*.

J. Brank, M. Grobelnik, N. Milic-Frayling, and D. Mladenic. 2002. Interaction of feature selection methods and linear classification models. In *Workshop on Text Learning held at ICML*.

H. Cui, V. Mittal, and M. Datar. 2006. Comparative experiments on sentiment classification for online product reviews. In *Proceedings of AAAI-06, the 21st National Conference on Artificial Intelligence*.

G. Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 3(1): 1289-1305.

E. Gabrilovich and S. Markovitch. 2004. Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with C4.5. In *Proceedings of the ICML, the 21st International Conference on Machine Learning*.

G. John, K. Ron, and K. Pfleger. 1994. Irrelevant features and the subset selection problem. In *Proceedings of ICML-94, the 11st International Conference on Machine Learning*.

S. Li and C. Zong. 2005. A new approach to feature selection for text categorization. In *Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*.

D. Mladeni and G. Marko. 1999. Feature selection for unbalanced class distribution and naive bayes. In *Proceedings of ICML-99, the 16th International Conference on Machine Learning*.

A. Moschitti. 2003. A study on optimal parameter tuning for Rocchio text classifier. In *Proceedings of ECIR*, *Lecture Notes in Computer Science*, vol. 2633, pp. 420-435.

E. Moyotl-Hernandez and H. Jimenez-Salazar. 2005. Enhancement of DTP feature selection method for text categorization. In *Proceedings of CICLing*, *Lecture Notes in Computer Science*, vol.3406, pp.719-722.

V. Ng, S. Dasgupta, and S. M. Niaz Arifin. 2006. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the COLING/ACL* Main Conference Poster Sessions.

K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3): 103-134.

B. Pang, L. Lee, and S. Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP-02, the Conference on Empirical Methods in Natural Language Processing*.

B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL-04, the 42nd Meeting of the Association for Computational Linguistics*.

E. Riloff, S. Patwardhan, and J. Wiebe. 2006. Feature subsumption for opinion analysis. In *Proceedings of EMNLP-06, the Conference on Empirical Methods in Natural Language Processing*,.

F. Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys,* 34(1): 1-47.

W. Shang, H. Huang, H. Zhu, Y. Lin, Y. Qu, and Z. Wang. 2007. A novel feature selection algorithm for text categorization. *The Journal of Expert System with Applications*, 33:1-5.

Y. Yang and J. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of ICML-97, the 14th International Conference on Machine Learning*.

Y. Yang and X. Liu. 1999. A re-examination of text categorization methods. In *Proceedings of SIGIR-99, the 22nd annual international ACM Conference on Research and Development in Information Retrieval*.