Selecting Features with Class Based and Importance Weighted Document Frequency in Text Classification

Baoli Li
College of Information Science and Engineering
Henan University of Technology
100 Lotus Street, High & New Industrial Development Zone
Zhengzhou, Henan, P.R. China
csblli@gmail.com

ABSTRACT
Document Frequency (DF), which counts how many documents a feature appears in, is reported by Yang and Pedersen [1] to be quite effective for feature selection in text classification. Features with the same DF value are likely to have different appearance distribution over categories, and demonstrate quite different discriminative powers for classification. However, the original DF metric is class independent and does not consider features' distribution over classes. On the other hand, different features play different roles in delivering the content of a document. The chosen features are expected to be the important ones, which carry the main information of a document collection. However, the traditional DF metric considers features equally important. To overcome simultaneously the above two problems of the original document frequency metric, we propose a class based and importance weighted document frequency measure. Preliminary experiments on two text classification datasets do validate the effectiveness of the proposed metric.

Keywords
Feature Selection; Document Frequency; Text Classification; Text Categorization; Feature Filtering.

1. INTRODUCTION
To process a large document collection, text classification is usually a necessary step to assign one or more pre-defined categories to each document. In general, we may have at least thousands of candidate features in a text classification problem, and, at the same time, these features are not equally effective in text classification. Therefore, feature selection, which aims at finding the most effective feature subset, is usually a must step [2]. It can not only reduce the time and space costs but also boost the system's performance. Different feature selection measures, such as document frequency, information gain, chi-square, bi-normal separation, odds ratio, mutual information, etc., have been put forward for ranking features in the past years [3]. Document Frequency (DF), among these measures, counts how many documents a feature appears in, and has been reported as a simple yet quite effective measure in solving different text classification problems [1].

DocEng '16, September 12-16, 2016, Vienna, Austria
© 2016 ACM. ISBN 978-1-4503-4438-8/16/09…$15.00
DOI: http://dx.doi.org/10.1145/2960811.2967164

As pointed in [4], the original DF metric does not care about class information. It cares only about whether a feature appears in a document. Thus derived DF values cannot differentiate two features having the same DF value. Those features, which distribute evenly over different categories, may have less discriminative power than those features with biased distribution. On the other hand, features in a document may have different importance in delivering the document's content, but the traditional DF regards features equally [5]. We do expect in the feature selection step to filter out those unimportant features, which are usually considered as noise.

To overcome both the above two weaknesses of the original document frequency measure at the same time, we propose a class based and importance weighted document frequency measure to revise the original DF to some extent. Basically, for each feature, we count its document frequencies in each category and then choose the maximal class document frequency for ranking. To give important features more weights, we add a value between 0 and 1 rather than always 1 when we find a feature appears in a document. The real value indicates how important the feature is in that document. Experiments on two publicly available datasets show that the proposed class based and importance weighted document frequency metric (CBIWDF) performs consistently better than the traditional DF, and achieves at least as good results as Chi-Square and information gain, which are two popular state-of-the-art feature selection measures.

The rest of this short paper is organized as following: section 2 summaries related work; section 3 presents the proposed class based and importance weighted document frequency (CBIWDF) measure; section 4 gives experimental results on two text classification problems and analyzes the results; finally, section 5 concludes the paper with some possible investigation in the future.

2. RELATED WORK
Feature Selection has been widely investigated in machine learning community in recent decades. The successful application includes but is not limited to: gene microarray analysis, combinatorial chemistry, image classification, face recognition, text clustering, spam detection, and text classification. The advent of big data era demands more for feature selection. Literature [6-9] presents excellent review about feature selection in dealing with different problems.

We normally have two alternative feature selection strategies: choosing a subset from a candidate set or deriving a new compact set from all candidate features. In this research, we concentrate on the first strategy, which is further split into two categories: wrapper and filter. Wrapper methods expect to obtain the ideal feature set by evaluating the performance of each candidate subset,

where filter methods rank features independently. Filter methods are more popular than wrapper ones, because of their lower computation cost. Forman [3] and Yang and Pedersen [1] empirically compare different feature filtering methods for text classification, including document frequency, information gain, chi-square, bi-normal separation, odds ratio, mutual information, power, and so on. Yang and Pedersen [1] conclude that the DF metric can perform as excellent as chi-square and information gain metrics. In this study, we concentrate on how to further improve Document Frequency (DF) metric.

In [4], we take class based document frequency as measure for ranking features, while an importance weighted document frequency strategy is reported in [5]. In this research, we explore how to solve the two problems of the traditional DF metric simultaneously: neglecting the difference of features' distribution over categories and counting each feature equally. We thus propose a class based and importance weighted document frequency feature selection metric.

3. CLASS BASED AND IMPORTANCE WEIGHTED DOCUMENT FREQUENCY

As pointed out in section 1, the original document frequency metric has two problems: one is ignoring the class distribution of a feature over different categories, and the other is regarding each feature equally. Obviously, features with imbalanced class distribution may have more discriminative capacity than those with balanced distribution, and the chosen features are expected to carry the main content of documents. We, thus, propose a class based and importance weighted document frequency measure to overcome the two weaknesses of the original DF metric. For a feature, we accumulate its importance value for each document it appears over different categories and then choose the maximal importance weighted class document frequency for ranking.

To evaluate the discriminative capacity of a feature t for class CLS_i, we need to count the following numbers:

A_i: how many documents of class CLS_i contain feature t;

B_i: how many documents with feature t do not belong to class CLS_i;

C_i: how many documents of class CLS_i do not contain feature t;

D_i: how many documents without feature t do not belong to class CLS_i.

Suppose that we totally have M categories in a classification problem. Then, the traditional document frequency (DF) measure can be calculated as follows:

$$DF = \sum_{i=1}^{M} A_i \tag{1}$$

, and the popular Chi-Square feature selection metric can be computed as follows:

$$Chi\text{-}Square = \sum_{i=1}^{M} CHI_i \tag{2}$$
$$= \sum_{i=1}^{M} \frac{(A_i + B_i + C_i + D_i) \times (A_i \times D_i - C_i \times B_i)}{(A_i + C_i) \times (B_i + D_i) \times (A_i + B_i) \times (C_i + D_i)}$$

A simple class based document frequency metric (CBDF), which is proposed in [4], use the following formula:

$$CBDF = \max_{i=1}^{M} A_i \tag{3}$$

, which chooses the maximal class document frequency of a feature for ranking.

If there are totally $D(CLS_i)$ documents of class CLS_i in the training data set, A_i can also be calculated as follows:

$$A_i = \sum_{j=1}^{D(CLS_i)} f(t, d_j) \tag{4}$$

, where $f(t, d_j)$ is defined as follows:

$$f(t, d_j) = \begin{cases} 1, & t \text{ is in } d_j \\ 0, & \text{no } t \text{ in } d_j \end{cases} \tag{5}$$

In formulas (4) and (5), $f(t, d_j)$ cares only about whether feature t appears in d_j, but fails to consider how important t is in d_j. Important features are expected to have more discriminative power than others. We then replace $f(t, d_j)$ with the following formula:

$$f(t, d_j) = \frac{TFIDF_t}{\sum_{f \in d_j} TFIDF_f} \tag{6}$$

, where $TFIDF_t$ is the $TFIDF$ value of feature t in document d_j. Formula (6) gives the relative importance of feature t in document d_j.

By combining formulas (1), (4), and (6), we obtain a variant of importance weighted document frequency (IWDF) as reported in [5].

$$IWDF = \sum_{i=1}^{M} \left(\sum_{j=1}^{D(CLS_i)} \left(\frac{TFIDF_t}{\sum_{f \in d_j} TFIDF_f} \right) \right) \tag{7}$$

From formulas (3), (4), and (6), we derive the class based and importance weighted document frequency metric as follows:

$$CBIWDF = \max_{i=1}^{M} \left(\sum_{j=1}^{D(CLS_i)} \left(\frac{TFIDF_t}{\sum_{f \in d_j} TFIDF_f} \right) \right) \tag{8}$$

4. EXPERIMENTS AND DISCUSSION

In order to evaluate the proposed CBIWDF metric, we conduct extensive experiments on two text classification problems.

4.1 Datasets

The following two datasets are used in our experiment:

20 Newsgroups: it is a balanced dataset, which has 20 different newsgroups, each corresponding to a specific topic [10]. We use the "bydate" version of this dataset, as it has a standard training and test split. The training set has 11,293 samples and the test set 7,528 samples.

Sector: it is an imbalanced dataset, which has 105 categories, 6,412 training samples, and 3,207 test samples. In the training dataset, the largest categories have 80 samples, while the smallest category has only 10 samples. Most categories have around 40-80 samples. This dataset was first used by McCallum and Nigam in their paper [11]. We used a version of this dataset from the LIBSVM data collection, which has removed stop and rare words (DF=1).

4.2 Experimental Settings

According to the vector space model (VSM), we represent a document as a space vector, whose coordinates correspond to the words in the collection. The weight of a feature, i.e. its TFIDF value, is computed as follows:

$$TFIDF_t = \frac{(1+\log(TF_{t,d}))\log(\frac{|D|}{DF_t})}{\sqrt{\sum_t((1+\log(TF_{t,d}))\log(\frac{|D|}{DF_t}))^2}} \quad (9)$$

, which is the standard feature weighting schema "ltc" in Manning and Schutze [11]. D is the document collection, where TF and DF are the frequency of feature t in document d and its document frequency in the collection D respectively. We experiment with four widely used algorithms: Centroid, Multinomial Naive Bayes[11], Linear (Liblinear [13]) and SVM (Libsvm [14]).

We compare the performance of different classification algorithms with the original DF, Chi-Square, information gain, CBDF, IWDF, and CBIWDF feature selection metrics. The original DF, Information Gain, and Chi-Square are used as baselines. We explore how to revise the original DF to get a better metric, and the Information Gain and Chi-Square metrics have been reported to perform well on many different problems and can be taken as state-of-the-art feature selection metrics.

We use Micro-averaging F1 and Macro-averaging F1 as evaluation metrics.

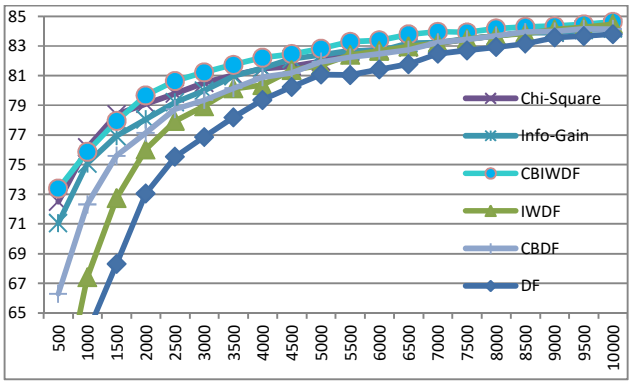
4.3 Results and Discussion

In our experiments, Liblinear achieves the highest scores among the four experimented text classification algorithms. Due to the limited space, we thus only report the results of this algorithm here. For each dataset, we experiment with top N features, where N varies from 500 to 10,000 with interval of 500.

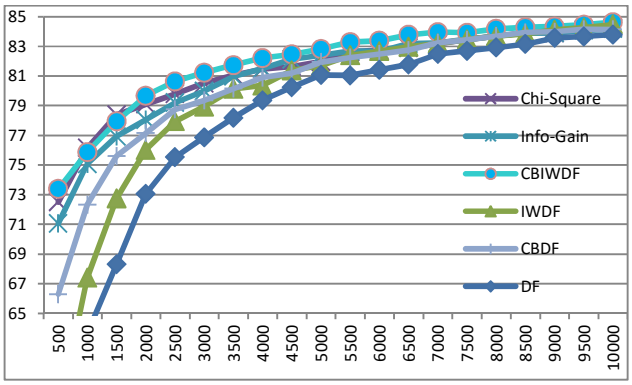
Figure 1 and figure 2 show the averaging Micro-F1 and Macro-F1 results of the DF, CBDF, IWDF, CBIWDF, information gain, and Chi-square as feature selection metrics on two datasets respectively.

On the 20 newsgroups dataset, CBDF, IWDF and CBIWDF constantly perform much better than the original DF. The difference is evident when using fewer features, but tends to be narrower when using more features. DF looks approximately good as others when we use more than 9,000 features (the total number of candidate features is 73,712), although all of them (CBDF, IWDF, and CBIWDF) do beat the original DF.

CBIWDF, which aims at solving the two problems of the original DF, does exhibit advantages over both CBDF and IWDF, as verifies that considering more factors do result in much better performance. On this dataset, CBDF achieves better result than IWDF. The problem of features' imbalanced distribution over categories is much serious in this dataset.



(a) Micro-Averaging F1 Scores



(b) Macro-Averaging F1 Scores

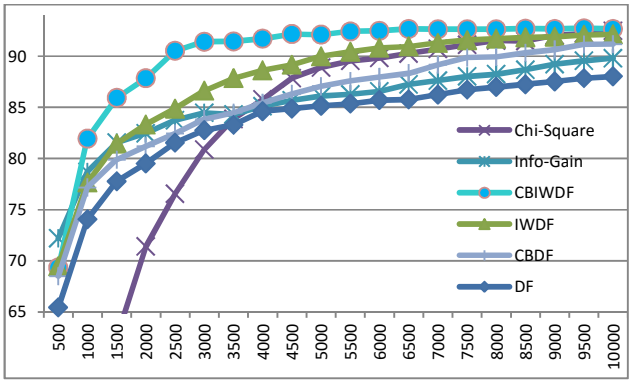
Figure 1. Performance of different feature selection methods on the 20 newsgroup dataset.

CBIWDF performs a little better than both information gain and Chi-square on this dataset, where chi-square obtains a little better result than information gain metric.

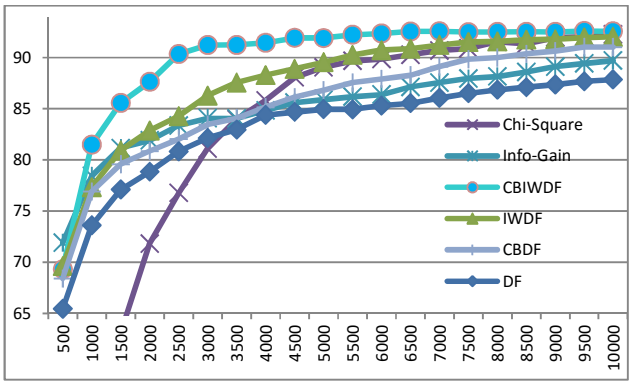
Figure 2 shows the results on the sector dataset. There are totally 48,988 candidate features in this dataset. When using less than 1,000 features, information gain metric obtains the best results, chi-square performs the worst, and the original DF does beat chi-square. Similarly as on the 20 newsgroups dataset, all the three variants of DF (CBDF, IWDF, and CBIWDF) demonstrate strong advantages over DF. When using more than 1,000 features, CBIWDF metric shows big advantages over all other metrics. When using less than 3,500 features, the preference order is as follows: IWDF > Info-Gain > CBDF > DF > Chi-square, where when using more than 3,500 features, this order is changed to be: IWDF > Chi-square > CBDF > Info-Gain > DF. Comparatively speaking, Information Gain, CBIWDF, IWDF, and CBDF metrics perform stably. On this dataset, IWDF achieves better result than CBDF, which means that features' importance plays key role in differentiating different features.

Chi-Square, one of popular feature selection metrics, obtains the poorest results when using less than 3,500 features. We attribute it to the fewer samples for each categories in this dataset, which make the class based Chi-Square metric less accurate and stable.

Overall, CBIWDF achieves at least as good results as information gain and Chi-Square do across all the two datasets. Compared to Chi-Square and information gain, the calculation and implementation of CBIWDF is much straightforward and trivial.



(a) Micro-Averaging F1 Scores



(b) Macro-Averaging F1 Scores

Figure 2. Performance of different feature selection methods on the sector dataset.

5. CONCLUSIONS AND FUTURE WORK

As an unsupervised and class independent metric, Document Frequency, is reported as a simple yet quite effective feature selection measure in text categorization. However, the original DF measure has two problems: one is ignoring the class distribution of a feature over different categories, and the other is regarding each feature equally. Targeting at solving these two problems simultaneously, we propose a class based and importance weighted document frequency measure for selecting features in text classification. Experiments on two publicly available datasets demonstrate that: 1) the proposed CBIWDF metric does perform better than the traditional DF metric, and two previously revised metrics, CBDF and IWDF; 2) CBIWDF can achieve at least as good results as Chi-Square and information gain, which are two popular state-of-the-art feature selection metrics.

In the future, we plan to experiment with more datasets and apply the proposed feature selection metrics into other text mining applications, e.g. text clustering, sentiment analysis, and so on. We also consider revising other existing feature selection metrics with similar strategies.

6. ACKNOWLEDGMENTS

This work was supported by the High-level Talent Foundation of Henan University of Technology (No. 2012BS027), and the Henan Provincial Research Program on Fundamental and Cutting-Edge Technologies (No. 112300410007).

7. REFERENCES

[1] Yang Y. and Pedersen J. O.1997. A comparative study on feature selection in text categorization. In *Proceedings of Fourteenth International Conference on Machine Learning*. 412-420.

[2] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* 3 (Mar. 2003), 1289-1305.

[3] Forman, G. 2007. Feature Selection for Text Classification. Technical Report (No. HPL-2007-16R1), HP Laboratories Palo Alto.

[4] Li B., Yan Q., and Han L. 2016. Using Class Based Document Frequency to Select Features in Text Classification. In Proceedings of the First National Conference on Big Data Technology and Applications (BDTA-2015): 200-210.

[5] Li B., Yan Q., Xu Z., and Wang G. 2015. Weighted Document Frequency for Feature Selection in Text Classification. In Proceedings of 2015 International Conference on Asian Language Processing: 132-135.

[6] Guyon, I., & Elisseeff, A. 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3 (Mar. 2003), 1157-1182.

[7] Chandrashekar, G., & Sahin, F. 2014. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), 16-28.

[8] Tang, J., Alelyani, S., & Liu, H. 2014. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*. Editor: Charu Aggarwal, CRC Press In Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

[9] Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., & Nowe, A. 2012. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (TCBB), 9(4), 1106-1119.

[10] Lang, K.1995. Newsweeder: learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, 331-339.

[11] McCallum A. and Nigam K. 1998. A Comparison of Event Models for Naive Bayes Text Classification. In *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*.

[12] Manning C. D. and Schutze H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

[13] Fan R.-E., Chang K.-W., Hsieh C.-J., Wang X.-R., and Lin C.-J.2008. LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.*9(2008), 1871-1874.

[14] Chang C.-C. and Lin C.-J.. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1--27:27.