A Feature Selection Method for Improved Document Classification

Tanmay Basu and C.A. Murthy

Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India
mailtanmaybasu@gmail.com, murthy@isical.ac.in

Abstract. The aim of text document classification is to automatically group a document to a predefined class. The main problem of document classification is high dimensionality and sparsity of the data matrix. A new feature selection technique using the google distance have been proposed in this article to effectively obtain a feature subset which improves the classification accuracy. Normalized google distance can automatically extract the meaning of terms from the world wide web. It utilizes the advantage of number of hits returned by the google search engine to compute the semantic relation between two terms. In the proposed approach, only the distance function of google distance is used to develop a relation between a feature and a class for document classification and it is independent of google search results. Every feature will generate a score based on their relation with all the classes and then all the features will be ranked accordingly. The experimental results are presented using knn classifier on several TREC and Reuter data sets. Precision, recall, f-measure and classification accuracy are used to analyze the results. The proposed method is compared with four other feature selection methods for document classification, document frequency thresholding, information gain, mutual information and χ^2 statistic. The empirical studies have shown that the proposed method have effectively done feature selection in most of the cases with either an improvement or no change of classification accuracy.

Keywords: Feature Selection, Document Classification.

1 Introduction

Document classification is the process of automatic grouping of documents to a predefined class. It is extremely difficult to effectively retrieve some particular information from the huge online resources without good indexing of document content [1]. Document classification has become one of the key tools for automatically handling and organizing such a huge document collections [4]. The huge dimensionality of the document collections is the main difficulty of any document classification method. So effectively reduce the dimension is a key part of document classification. The standard procedure for dimensionality reduction is feature selection. Feature selection is a process that chooses a subset from the original feature set according to some criteria. The selected feature subset

S. Zhou, S. Zhang, and G. Karypis (Eds.): ADMA 2012, LNAI 7713, pp. 296–305, 2012.

retains original physical meaning and provides a better understanding for the data and learning process [12]. The feature selection for document classification task use an evaluation function that is applied to a single term [5]. Then all the terms will be sorted according to the score of the evaluation function assigned independently to each feature. Among this sorted list a predefined number of features will be selected as the best feature subset. Among various feature selection methods, Document frequency (DF) thresholding, Information Gain (IG), Mutual Information (MI), χ^2 statistic (CHI) are commonly applied techniques in document classification [1]. Yang et. al. [1] investigated five feature selection methods. They reported that good feature selection methods improve the categorization accuracy with an aggressive feature removal using DF, IG and CHI methods.

A simple technique for vocabulary reduction in document classification [1] is *DF thresholding*. Document frequency refers to the number of documents in which a term occurs. The document frequency of each term in the training documents will be computed and the terms with a document frequency less than a predefined threshold will be discarded from the vocabulary. *Mutual information* based feature selection assumes that the term with higher class ratio is more effective for classification. On the other hand rare terms will have a higher score than common terms for those terms with equal conditional probability. Hence MI might perform badly when a classifier gives stress on common terms. For a training corpus those terms whose MI score is less than a predefined threshold will be removed from the vocabulary. It is to be mentioned here that the same choice is made for other two methods IG and CHI. *Information gain*, measures the number of bits of information obtained for class prediction by knowing the presence or absence of a term in a document [1]. It gives more weight to common terms rather than the rare terms. Hence IG might perform badly when there are scarcity of common terms between the documents of the training corpus. Supervised feature selection using χ^2 *statistic* measures the association between the term and the class. In our experiments we have used the maximum CHI score for comparison. Forman [11] shows a comparative study of twelve feature selection methods on 229 text classification problem instances and proposed a new method called *bi-normal separation*. Their experiments show that bi-normal separation can perform very well in the evaluation metrics of recall rate and f-measure. But for precision, it often loses to IG. Liu et. al. [12] proposed two new unsupervised feature selection methods. First one is *term contribution* which ranks the feature by its overall contribution to the documents similarity in a data set. Another is, *iterative feature selection*, which utilizes some successful feature selection methods, such as IG and CHI, to iteratively select features and perform text clustering at the same time.

Cilibrasi et. al. [3] developed normalized google distance to measure the semantic relation between two terms/phrases using google page counts. Google page count is the number of hits of a term displayed by google. The google distance determines how close a term x is to another term y on a zero to infinity

scale. A distance zero indicates two terms are exactly same. The distance will be infinity when the terms never occur together. Here we shall use this distance to develop a relation between a term and a class for document classification. Then every feature will be assigned a score depending on their relation with all the classes. All the terms will be ranked according to their individual score. Then a predefined number of terms will be selected as very important features which have high weights. The idea is new in feature selection literature and no google search results are required for the proposed method.

We have applied the method on several text data sets. Knn classifier is used to judge the effectiveness of the proposed feature selection method and precision, recall, f-measure and classification accuracy are used for experimental analysis. The empirical studies show that the proposed method performs better than the other conventional feature selection methods even after 90% terms removal in most of the cases. The proposed method is an alternative approach of the existing feature selection methods and in various situations the proposed method is showing either an improvement or no change in classification performance, when compared to the other methods.

The paper is organized as follows. The proposed method is described in section 2. The experimental results are shown and discussed in section 3. Section 4 presents some conclusions on the proposed method.

2 Feature Selection by Term Relevance

Normalized google distance (NGD) was introduced by Cilibrasi et. al. [3] to utilize the vast knowledge available on the web by using the google page counts. NGD determines the semantic distance between two terms using the number of hits returned by Google search engine as search terms. The value of NGD lies between 0 and ∞. The NGD value 0 indicates that the terms are exactly same. If two terms never appear together then the value will be ∞. This weighting scheme of NGD will be used to derive a relation between a term and a class in the proposed feature selection method. In this method we shall measure the relevance of a term with a class. The class for which the term has the highest score will be the ultimate score of the term over all the classes. The relevance of a term t with a particular class c can be derived from the following distance function *term relevance* (TR).

$$TR(t,c) = \begin{cases} -1, & \text{if } f(t,c) = 0 \\ \frac{\max\{logf(t), logf(c)\} - logf(t,c)}{logN - \min\{logf(t), logf(c)\}}, & \text{otherwise} \end{cases} \quad (1)$$

Here f(t) is total number of documents containing term t and f(c) is total number of documents in class c. f(t, c) denotes the number of documents belonging to class c and containing the term t, and N is the total number of documents. To measure the relevance of a term in global feature space, all the class specific

scores of a term are to be taken into consideration. The score of a term over all the m classes can be obtained in the following way:

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

All the features will be ranked in decreasing order according to their $TR_{max}(t)$ value and then a predefined number of terms will be selected for classification which have high weights. The following are the main properties of the proposed term relevance.

Properties of TR

• The value of TR will be 0, when $f(t) = f(c) = f(t, c)$, for any term t and a class c.

• From the right hand side of equation 1 we have

$$\begin{aligned}&\max\{logf(t), logf(c)\} - logf(t, c) - logN + \min\{logf(t), logf(c)\}\\ &= logf(t) + logf(c) - logf(t, c) - logN\\ &= log\frac{f(t) * f(c)}{f(t, c) * N}\end{aligned}$$

Now $f(t) * f(c)$ may be greater than $f(t, c) * N$. Hence the value of TR may be greater than 1 for any pair of class c and term t. But the value of TR will always be finite.

• $TR(t, c) \geq 0$, if a term t appears in the class c. The only negative value of TR is -1 when t does not belong to c. Hence TR is not a metric.

• TR is symmetric. For every pair of a term t and a class c, we have $TR(t, c) = TR(c, t)$.

3 Empirical Evaluation

This section describes the performance of various feature selection methods in document classification using several document collections. The description of the data sets and performance measures are given below. The effectiveness of different feature selection algorithms is evaluated using the performance of knn classifier. The knn classifier is chosen since it shows very good performance in a previous study by Yang et. al. [2]. 10-fold cross validation method has been used to fix the value of k from the training set. The range of k has been set from 1 to 20. For all the data sets except Reuter have no separate test and training sets. So 10 fold cross validation is performed on the entire data set to split it into training and test set. Knn has been executed for 10 times to reduce the effect of random selection of the documents by cross validation. The average results of this 10 executions is reported in Table 2.

Table 1. Data Sets Overview

Data Set	No. of Documents	No. of Terms	No. of Classes
la1	3204	31472	6
la2	3075	31472	6
la12	6279	31472	6
Reuter	13324	29944	90
tr21	336	7902	6
tr41	878	7454	10

3.1 Document Collections

Reuters-21578[1] is a collection of documents that appeared on Reuters newswire in 1987. The documents were originally assembled and indexed with categories by Carnegie Group, Inc. and Reuters, Ltd. The corpus contains 21578 documents in 135 categories. In this version of Reuter the documents with multiple class labels were discarded and the largest 90 categories were selected. The corpus was divided into 9583 training documents and 3741 test documents according to the Apte split [6].

The rest of the text data sets have been developed by Karypis and Han[2] [7]. Data sets tr21, tr41 are derived from TREC-5, TREC-6, and TREC-7 collections [8]. Data sets la1, la2 and la12 are from the Los Angeles Times data of TREC-5 [8]. The classes of the tr21 and tr41 data sets were generated from the relevance judgment provided in the TREC collections. The class labels of la1 and la2 were generated according to the name of the news-paper sections that these articles appeared, such as Entertainment, Financial, Foreign, Metro, National, and Sports. The documents that have a single label are selected for the la1, la2 and la12 data sets.

The number of documents, number of terms and number of classes of all the data sets can be found in Table 1. For the above data sets, the stop words have been extracted using the standard English stop word list[3]. Then, by applying the standard porter stemmer algorithm for stemming, the inverted index is developed. We have removed those words which occurred only once, twice or thrice in the corpus at first.

3.2 Performance Measures

In order to evaluate the effectiveness of class assignments, the standard recall, precision and f-measure are used. Precision and recall for a particular class are defined as:

$$\text{recall} = \frac{\text{number of documents found which are correct}}{\text{number of actual documents}}$$

$$\text{precision} = \frac{\text{number of documents found which are correct}}{\text{number of documents found}}$$

[1] `http://www.daviddlewis.com/resources/testcollections/reuters21578/`
[2] `http://www-users.cs.umn.edu/~han/data/tmdata.tar.gz`
[3] `http://www.textfixer.com/resources/common-english-words.txt`

The f-measure combines recall and precision with an equal weight in the following form:

$$\text{f-measure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

These scores are computed for the binary decisions on each individual class and then be averaged over all classes. The closer the values of precision and recall, the higher is the f-measure. The value of f-measure lies between 0 and 1. Classification accuracy is also measured for performance evaluation.

Table 2. Comparison of Various Feature Selection Methods for Document Classification

Data	Vocabulary Size(%)	F-measure					Accuracy (%)				
		CHI	DF	IG	MI	TR	CHI	DF	IG	MI	TR
la1	10%	0.816	0.649	0.815	0.296	**0.817**	81.55	66.15	81.54	34.06	**82.51**
	20%	0.811	0.670	0.822	0.580	**0.830**	81.12	67.24	82.30	58.22	**83.67**
	30%	0.816	0.734	0.820	0.682	**0.834**	81.71	74.10	82.05	68.66	**83.77**
	50%	0.821	0.789	0.814	0.777	**0.836**	82.24	79.13	81.51	78.15	**83.94**
	100%	0.809	0.809	0.809	0.809	0.809	80.97	80.97	80.97	80.97	80.97
la2	10%	0.844	0.660	0.834	0.512	**0.854**	84.52	66.81	83.64	52.77	**85.68**
	20%	0.835	0.731	0.835	0.738	**0.864**	83.65	73.38	83.53	73.64	**86.85**
	30%	0.832	0.742	0.838	0.785	**0.858**	83.12	74.10	83.89	78.77	**86.16**
	50%	0.835	0.798	0.833	0.824	**0.859**	83.55	79.84	83.31	82.78	**86.09**
	100%	0.831	0.831	0.831	0.831	0.831	83.09	83.09	83.09	83.09	83.09
la12	10%	0.841	0.653	0.835	0.402	**0.847**	84.15	66.62	83.56	41.62	**84.83**
	20%	0.838	0.711	0.843	0.654	**0.859**	83.87	72.13	84.34	65.28	**86.22**
	30%	0.841	0.754	0.842	0.748	**0.859**	84.15	75.96	84.23	75.20	**86.27**
	50%	0.842	0.786	0.841	0.825	**0.860**	84.24	78.60	84.14	82.89	**86.35**
	100%	0.840	0.840	0.840	0.840	0.840	84.06	84.06	84.06	84.06	84.06
Reuter	10%	0.642	0.500	0.659	0.306	**0.660**	68.05	54.10	**69.09**	34.96	**69.09**
	20%	0.644	0.573	**0.656**	0.482	**0.656**	68.29	60.94	**68.72**	52.12	68.56
	30%	0.645	0.586	0.656	0.539	**0.659**	67.76	61.96	68.75	57.84	**69.17**
	50%	0.633	0.600	0.656	**0.672**	0.640	66.02	62.76	68.16	**69.60**	66.77
	100%	0.634	0.634	0.634	0.634	0.634	65.97	65.97	65.97	65.97	65.97
tr21	10%	0.873	0.781	**0.894**	0.817	0.881	88.12	79.78	**89.87**	83.28	88.27
	20%	0.875	0.812	0.886	0.859	**0.913**	88.42	84.20	90.29	87.18	**91.93**
	30%	0.876	0.831	0.877	0.866	**0.910**	88.81	84.43	89.35	88.54	**91.51**
	50%	0.867	0.851	0.862	0.866	**0.880**	88.59	85.74	88.85	86.91	**89.33**
	100%	0.857	0.857	0.857	0.857	0.857	87.80	87.80	87.80	87.80	87.80
tr41	10%	0.919	0.761	0.924	0.399	**0.931**	92.23	76.27	92.82	43.44	**93.46**
	20%	0.926	0.855	0.921	0.666	**0.927**	92.78	86.10	92.13	66.91	**92.93**
	30%	0.922	0.877	**0.924**	0.789	0.923	92.50	87.94	**92.76**	79.03	92.49
	50%	0.919	0.904	**0.925**	0.921	0.923	92.18	90.54	92.82	92.43	**92.71**
	100%	0.922	0.922	0.922	0.922	0.922	92.60	92.60	92.60	92.60	92.60

3.3 Analysis of Results

Table 2 shows the performance of various feature selection methods in document classification using f-measure and classification accuracy. The results are shown using several thresholds on the total number of unique terms i.e., the performance of the knn classifier is reported after removing 50%, 70%, 80%, and 90% unique terms. The vocabulary size in Table 2 indicates that the experiments are performed when there are 10%, 20%, 30%, 50% and 100% unique terms in the vocabulary. The proposed TR is compared with CHI, DF, IG and MI methods.

If significant amount of terms were removed from the vocabulary at high levels (after 90% terms removal) by TS then knn would not provide improved classification performance which becomes clear from the experimental results. But the experimental results show that TR is able to detect significant terms even after 90% or more terms removal. Let us take the example of la2 data set using TR thresholding when the total number of unique terms are reduced from 100% to 10%, the f-measure is improved from 0.831 to 0.854. The same thing can be observed in all the data sets for TR. The performance of CHI and IG are also not degraded after 90% or more terms removal in most of the data sets. The DF and MI methods are not able to provide better classification performance even after removal of more than 50% terms in most of the cases.

The proposed method is compared with other four methods and six data sets have been used for analysis. So there are 192 comparisons in Table 2 for the proposed method for 50% to 90% removal of terms. TR performed better than the other methods in 179 cases and in the rest 13 cases other methods (e.g., CHI, IG) have an edge over TR. Consider two such cases where the other methods have an edge over TR. For data set reuter, when there are 20% terms in vocabulary, IG (68.72%) has an edge over TR (68.56%) in classification accuracy and for data set tr41, when there are 30% terms in vocabulary, IG (0.924) has an edge over TR (0.923) in f-measure.

A statistical significance test is needed to check whether these differences are significant, e.g., whether 0.924 is significantly different from 0.923. A generalized version of paired *t-test* is used for testing the equality of means when the variances are unknown. This problem is the classical Behrens-Fisher problem in hypothesis testing and a suitable test statistic[4] is described and tabled in [9] and [10], respectively.

It has been found using t-test that out of those 179 cases where TR performed better than the other methods, in 164 cases the difference was statistically significant for the level of significance 0.05. Out of the rest 13 cases where other methods have an edge over TR, all the differences were statistically significant for the same level of significance. Thus from the experimental results it can be observed that the proposed TR will be very useful for document classification.

Figure 1 show the precision-recall curve of CHI, IG and TR for all the data sets. CHI and IG have comparable performances with TR for precision and

[4] The test statistic is of the form $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$, where $\bar{x}_1, \bar{x}_2$ are the means, s_1, s_2 are the standard deviations and n_1, n_2 are the number of observations.

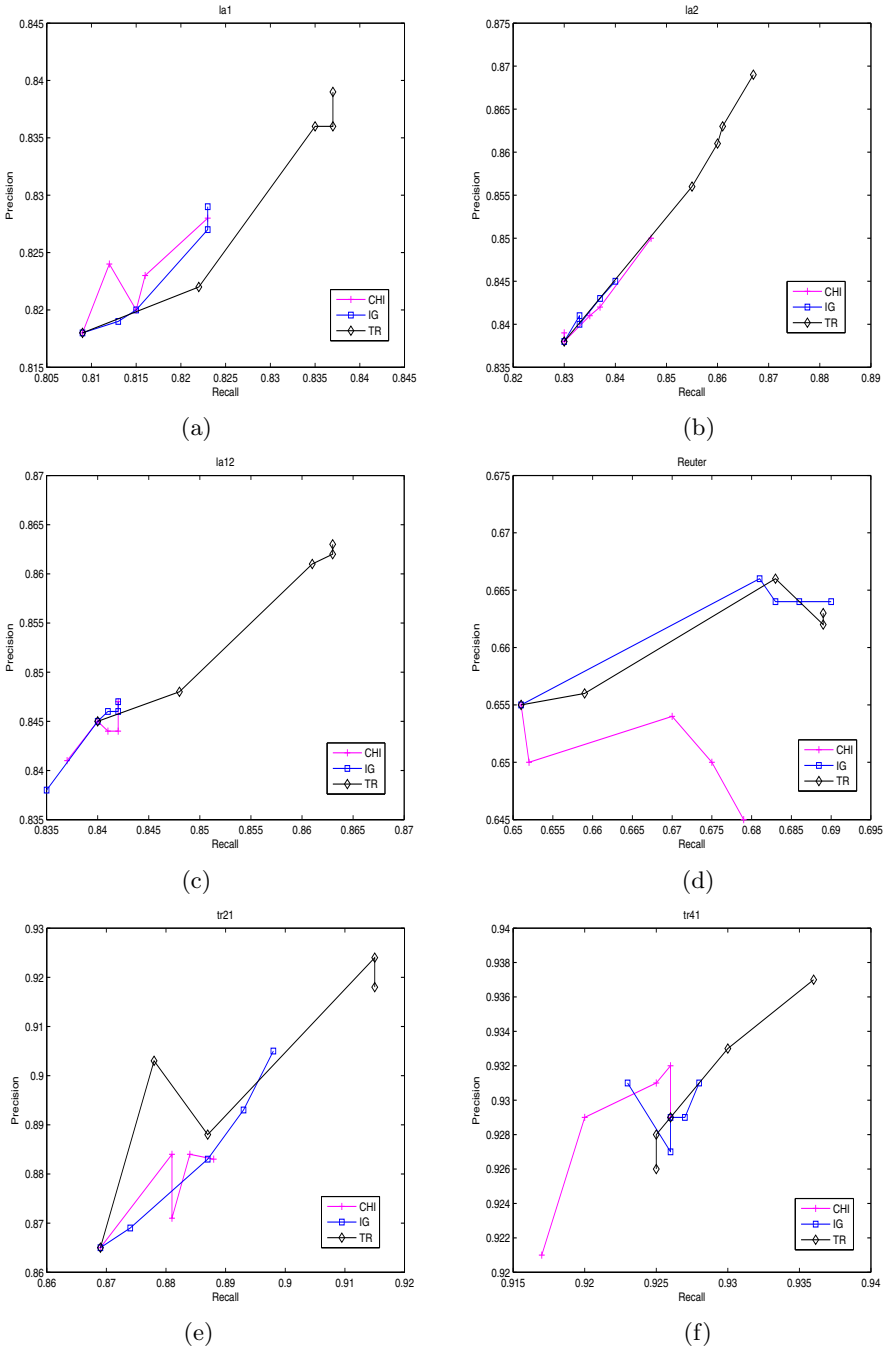

Fig. 1. Precision Recall Curve by Varying the Number of Terms for all the Data Sets

recall. DF never performed better than TR and MI performed better than TR only once. So that DF and MI are not included in the figures and the precision-recall curves are developed for CHI, IG and TR only. It is observed from figure 1(a), figure 1(b), figure 1(c) and figure 1(e) that TR is the best performer. TR shows better performance than CHI, but IG performs comparably better than TR in figure 1(d). In figure 1(f), for recall rate below 93% CHI and IG perform better than TR, but for rest of the cases TR performs better than CHI and IG. Thus the superiority of TR can be observed from the precision-recall curves of all the data sets.

4 Conclusions

Effectively managing the high dimensionality of data is a difficult task for document classification. An efficient feature selection method is needed to improve the performance of document classification. In this article a new feature selection method is proposed for document classification using the distance function of google distance. The google distance was developed to extract semantic distance between two terms using google search results. The proposed term relevance is developed using this distance to derive a relation between a term and a class and then rank the terms according to their scores over all the classes. The experimental results show that term relevance can produce better classification accuracy even after removing 90% unique terms. It can also be seen from the experiment that term relevance outperforms the other methods. It is to be noted that the proposed term relevance is not more computationally expensive than the existing feature selection methods in document classification. Hence the proposed term relevance can be applied to any real life document collection for improved classification.

References

1. Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization. In: Proc. of the Fourteenth International Conference on Machine Learning (ICML 1997), pp. 412–420 (1997)
2. Yang, Y., Liu, X.: A Re-examination of Text Categorization Methods. In: Proc. of the Twenty-Second International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999), pp. 42–49 (1999)
3. Cilibrasi, R.L., Vitanyi, P.M.: The Google Similarity Distance. IEEE Transactions on Knowledge and Data Engineering 19(3), 370–383 (2007)
4. Li, S., Xia, R., Zong, C., Huang, C.: A Framework of Feature Selection Methods for Text. In: Proceedings of ACL-IJCNLP 2009 (2009)
5. Novovicova, J., Malik, A.: Information-Theoretic Feature Selection Algorithms for Text Classification. In: Proceedings of International Joint Conference on Neural Networks, Montreal, Canada, July 31-August 4 (2005)
6. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998)

7. Karypis, G., Han, E.H.: Centroid-Based Document Classification: Analysis and Experimental Results. In: Zighed, D.A., Komorowski, J., Żytkow, J.M. (eds.) PKDD 2000. LNCS (LNAI), vol. 1910, pp. 424–431. Springer, Heidelberg (2000)
8. TREC, Text REtrieval Conference, http://trec.nist.gov
9. Lehmann, E.L.: Testing of Statistical Hypotheses. John Wiley, New York (1976)
10. Rao, C.R., Mitra, S.K., Matthai, A., Ramamurthy, K.G. (eds.): Formulae and Tables for Statistical Work. Statistical Publishing Soc., Calcutta (1966)
11. Forman, G.: An Extensive Empirical Study of Feature Selection Metrics for Text Classification. The Journal of Machine Learning Research 3(1), 1289–1305 (2003)
12. Liu, T., Liu, S., Chen, Z., Ma, W.: An Evaluation on Feature Selection for Text Clustering. In: Proc. International Conference on Machine Learning (ICML 2003) (2003)