Feature Selection: An Important Issue in Text Categorization

*Deepshikha Patel
*Rajesh Nigam
*Vijay K. Chaudhari
* Bhupendra Verma

Abstract

Text categorization is a problem of assigning a document into predefined classes. Feature selection is one of the important issues in text categorization. Wide variety of feature selection methods exist for text categorization like Information Gain (IG), Document Frequency (DF),Term Strength (TS), Mutual Information (MI) etc. Feature selection methods can improve the efficiency and performance of text categorization. This paper reports a controlled study on a large number of feature selection techniques for text classification. We also discuss some variation and combinations of these feature selection methods.

Index Terms—Classification, Feature Extraction, Feature Selection, Text Categorization.

1. Introduction: With the growth of the internet and advancement of computer technologies more textual documents have been digitized and stored electronically. Thus text classification became an increasingly important task. The goal of text categorization is to assign a new document into predefined category by identifying discriminating features. A document could fall into one class or many. As the volume of text content grows continuously online, effective retrieval is difficult without good indexing and summarization of document content. Categorization of document is one solution to this problem. A growing number of statistical classification methods have been applied to text categorization such as Naive Bays [3, 4], Decision Tree [7], Neural Network [6], Linear Regression [11], k-Nearest neighbor [5], Support Vector Machine [1].A comparative study of text categorization methods is reported in [2] against the Reuters corpus.

The text categorization problem normally involves an extremely high dimensional feature space [8]. The performance of classification algorithms will decrease dramatically due to the problems of high dimensionality of feature space. Therefore there is a high requirement to reduce feature space. Feature selection is a standard procedure to reduce features dimensionality, which selects “good” features for a classifier. Many Feature Selection methods such as Document Frequency (DF), Term Strength (TS), Mutual Information (MI), CHI Statistics, and Information Gain have been applied to Text categorization [9]. Some variants of feature selection methods are also used by Yang [23].

2. Feature Selection: Almost every popular classifier accepts as input a feature vector that characterizes the document to be classified. Clearly construction of these features vector is very important to the successful operation of the classifier. Selection of a subset of features to be used in inductive learning has already been addressed in machine learning. In order to transform a document into a feature vector, preprocessing is needed. This includes feature extraction, feature selection and feature weighting calculation.

a) **Feature Extraction:** Feature extraction is a process that extracts a set of new features from original features through some functional mapping such as PCA and word clustering. The mapping of approaches use a simple ‘bag of words’ approach that include all the words in a document except stop list, a list of the most common words that are unlikely to be distinguishing features.

b) **Feature Selection:** Feature Selection is a process that chooses a subset from the original set that is formed by feature extraction process. The number of features identified by Feature extraction may be extremely large. Generally high dimensionality of the term space can make the classifier run slowly and increase over fitting, i.e. the phenomenon by which the classifier perform well on reclassifying the documents of training set and perform badly on classifying new documents. Hence Feature selection, which aims to reduce the dimensionality of the feature vector by only retaining those feature that are most informative or distinguish. There are several effective feature selection methods which are discussed below.

i) 2 statistic (CHI)

The □2 statistic measures the association between the term t and category c [9].It is defined to be

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \qquad (1)$$

Using the two way contingency table of a term t and category c, where A is the number of times t and c co-occur, B is the number of times c occurs without c, C is the number of times c occurs without t, D is the number of times neither c nor t occurs, and N is the total number of documents.

ii) Information gain (IG): Information gain [9], an information theoretic function that tries to keep only the terms distributed in the sets of positive and negative examples of the categories. Let m be the number of categories. The information gain of a term t is defined as

This definition is more general than the one employed in binary classification models [10, 12]

iii) Document Frequency (DF): Document Frequency is the number of document in the training corpus in which a term occurs. It is the simplest criterion for term selection a variation of document frequency is document frequency thresholding in which we compute the document frequency for each unique term in the training corpus and remove from the feature space those terms whose document frequency was less than some predetermined threshold.

Document Frequency thresholding is the simplest techniques for vocabulary reduction. It easily scales to very large corpora, with a computational complexity approximately linear in the number of training document.

iv) Term strength (TS) :

Term strength is originally proposed and originally proposed and evaluated for vocabulary reduction in text retrieval [14]. And later applied by Yang and Wilbur to text categorization [13]. It is computed based on how commonly a term is likely to appear in closely related documents.

Let x and y be an arbitrary pair of distinct but related documents, and t be a term then the term strength is defined by

$$S(t) = P(t \in y \mid t \in x) \tag{3}$$

v) Mutual information (MI): Mutual information [15] is a criterion commonly used in statistical language modeling of word associations and related applications [15,16,17]. If one considers the two way contingency table, records co-occurrence statistics for terms and categories, of a term t and a category c, where A is the number of times t and c co-occur, B is the number of times the t occur without c, C is the number of times c occurs without t, and N is the total number of documents, thus the mutual information is given by

$$I(t,c) = \log \frac{P(t \,\&\, c)}{P(t)P(c)} \tag{4}$$

This may be approximated by

$$\frac{A \times N}{(A+C)\times(A+B)} \tag{5}$$

Since mutual information gives values for pairs, rather than individual terms. Yang and Pedersen calculate both the maximum and average mutual information for each term and test both.

$$I_{avg}(t) = \sum_{i=1}^{m} P(c_i) I(t, c_i) \tag{6}$$

$$I_{\max(t)} = \max_{i=1}^{m} \{I(t, c_i)\} \tag{7}$$

vi) Entropy based ranking (En)

Entropy based ranking is proposed by Dash and Liu [18]. In this method, the term is measured by entropy reduction when it is removed. The entropy is defined as:

$$E(t) = -\sum_{i=1}^{N}\sum_{j=1}^{N} (M_{ij} \times \log(M_{ij}) + (1-M_{ij}) \times \log(1-M_{ij})) \tag{8}$$

Where,

M_{ij} =Similarity value between documents $D_i \,\&\, D_j$

This can be formulated as:

$$M_{ij} = e^{-\alpha * Dist_{ij}}, \alpha = -\frac{\ln(0.5)}{\overline{\overline{Dist}}} \tag{9}$$

Where Dist, is the distance between the document D_i and D_j after the term t is removed, Dist is the average distance among the documents after the term *t* is removed.

vii) Term Contribution (TC): Term Contribution is introduced by Liu et al.[19]. This method includes term weight in calculation. Because, the result of classification depends on the similarity of documents. The similarity between two documents can be expressed as :

$$Similarity(D_i, D_j) = \sum w(t, D_i) \times w(t, D_j) \qquad (10)$$

Where,

$w(t, D\ )$ represents the tf*idf [20] weight of term t in document D. so, the contribution of a term in a set of documents is given by the equation(11).

$$TC(t) = \sum_{i,j \cap i \neq j} w(t, D_i) \times w(t, D_j) \qquad (11)$$

viii) Other Dimensionality Reduction Techniques: Apart from the feature selection methods discussed above, some other feature selection methods are also there. A brief introduction of these approaches is discussed below.

Latent Semantic Indexing (LSI)

A different approach for dimensionality reduction of the term space is to infer, from the original term space by document matrix, a new term by document matrix in which terms are no more intuitively interpretable but can express the latent semantics of the documents. The technique used is called *Latent Semantic Indexing (LSI)* [21].

Principal Component Analysis (PCA)

This method is also called Karhunen-Loeve or K-L method. PCA can also be used as a feature selection and reduction method in which original data are projected into much smaller space, resulting in dimensionality reduction. Detailed study of PCA can be found in the work of Calvo et. al.[22]. PCA is computationally inexpensive, can handle sparse and skewed data. Multidimensional data of more than two dimensions can be handled by reducing the problem to two dimensions.

3. Method Variations & Combinations: Feature selection for text categorization is well known problem. Feature selection techniques are used to improve classifier performance and computational efficiency. For this purpose several method variations used by yang [23].some of them are: Combination of IG and CHI with their generalized versions, Eliminating rare words (DF<=5), Combination of both the average and maximum value as the score for IG, CHI etc.

4. Conclusion: Feature selection methods have successfully applied to text categorization for long years. There are various feature selection methods to select good features, extracted by feature extraction method. All feature selection methods are not suitable for every type of classification task. We can say efficiency of feature selection methods vary according to type of data set chosen. To improve efficiency many different combination of feature selection methods are also used. Feature selection can improve dramatically improve the efficiency of text categorization and even improve the categorization accuracy to some extent, so it is an interesting idea to apply feature selection methods to text clustering task to improve the clustering performance.

5. References:

1. T. Joachims, (1998)“Text categorization with support vector machine: learning with many relevant features,” In10th European Conference on Machine learning (ECML-98), pp. 137-142.
2. Y Yang, (1999)“An Evaluation of statistical approaches to text categorization,” Journal of Information Retrieval, Vol.1,No. ½,
3. P. Frasconi, G. Soda and A. Vullo, (2001) “Text categorization for multi page document: a hybrid naive Bayes HMM approach”, In proceeding of 1st ACM/IEEE-CS joint conference on Digital libraries; ACM Press New York, NY, USA, pp. 11-20.
4. A.M. Kibriya, E. Frank, B. Pfahringer and G. Holmes.(2004) “Multinomial naive bayes for Text categorization” revisited. AI 2004: Advances in Artificial Intelligence, 3339, pp. 488–499.
5. G. D. Guo, H. Wang, D. Bell, Y. X. Bi, and K. Greer.(2006) ”Using kNN model for automatic text categorization”. Soft Computing, 10(5), pp. 423–430.
6. R. N. Chau, C. S. Yeh, and K. A. Smith.(2005)”A neural network model for hierarchical multilingual text categorization”. Advances in Neural Networks, LNCS, 3497, pp. 238–245.
7. S. Gao, W. Wu, C. H. Lee, and T. S. Chua.(2006) “A maximal .gure-of-merit (MFoM)-learning approach robust classifier design for text categorization”. ACM Transactions on Information Systems, 24(2), pp. 190–218.
8. Wai, Lam, and Yiqiu Han.(2003)“Automatic Textual Document Categorization Based on Generalized Instance Sets and a Metamodle,” IEEE Transaction on Pattern Analysis and Machine Intelligence.vol.25, no.5, pp.628-633.
9. Yang, Yiming, (1997) ” A Comparative Study on Feature Selection in Text Categorization,” In Proceeding of the Fourteenth International Conference on Machine Learning(ICML'97),pp.412-420.
10. D.D Lewis and M. Ringuette.(1994) “Comparison of two learning algorithms for text categorization,”. Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR’94).
11. D. Lewis and J. Catlett.(1994) “Heterogeneous uncertainty sampling for supervised learning”. In Proceedings of the Eleventh International Conference on Machine Learning, pp. 148–156.
12. I. Moulinier, G. Raskinis and J. Ganascia. (1996) “Text categorization: a symbolic approach,”In Proceeding of Annual Symposium on Document Analysis and Information Retrieval.
13. Y Yang and W. J. Wilbur.(1996) ”Using corpus statistics to remove redundant words in text categorization, ”, In J Amer Soc Inf Sci.
14. J.W. Wilbur and K. Sirotkin.(1992) “The Automatic identification of stop words,”,J. Inf. Sci. 18:45-55.
15. R. Fano ?(1961)“Transmission of Information,” IT Press, Cambridge, MA.

16. K.W. Church and P Hanks.(1989) “Word Association norms, mutual information and lexicography,” In Proceeding of ACL 27, pp. 76-83, Vancouver, Canada
17. E. Wiener, J.O. Pedersen, and A.S. Weigend. (1995) “A Neural Network approach to topic spotting,” In Proceeding of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR’95), 1995.
18. M. Dash and H. Liu, (2000) “Feature Selection For clustering,” In Proceeding of PAKDD-00, pp.110-121.
19. T. Liu, S. Liu and, Z. Chen, (2003)“An Evaluation of Feature Selection for Text Clustering,” In Proceeding of the Twentieth International Conference on Machine Learning (ICML-2003),Washington DC.
20. G. Salton, (1989) ”Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer,” Addison-wesley, Reading, Pennsylvania.
21. .S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman,(1990)” Indexing by Latent Semantic Analysis,” Journal of the American Society of Information Science, pp.391-407.
22. R. A. Calvo, M. Partridge, and M. A. Jabri, (1998) A Comparative Study of Principal Component Analysis Techniques,” In Proc. Ninth Australian Conf. on Neural Networks, Brisbane.
23. M. Rogati and Y. Yang, (2002) ”High Performance Feature Selection for Text Classification,” In Proceeding of CIKM’02,Melean, Virginia, USA, pp.659-661.