Enhancing Feature Selection Using Statistical Data with Unigrams and Bigrams

M.Janaki Meena
Department of CSE
PSG College of Technology
Coimbatore,Tamil Nadu,India

K.R.Chandran
Department of IT
PSG College of Technology
Coimbatore,Tamil Nadu,India

J.Mary Brinda
Student, Department of CSE
PSG College of Technology
Coimbatore,Tamil Nadu,India

P.R.Sindhu
Student, Department of CSE
PSG College of Technology
Coimbatore,Tamil Nadu,India

ABSTRACT

Feature selection is an essential preprocessing step for classifiers with high dimensional training corpus. Features for text categorization include words, phrases, sentences or distribution of words. The complexity of classifying documents to related categories is on higher scale in comparison with unrelated categories. A feature selection algorithm based on chi-square statistics, have been proposed for Naïve Bayes classifier. The proposed feature selection method identifies the related features for a class and determines the type of dependency between the feature and category. In this paper, the proposed method ascertains related phrases and words as features. A comparison of the conventional chi-square method is made with the proposed method. Experiments were conducted with randomly chosen training documents from one unrelated and five closely related categories of 20Newsgroup Benchmarks. It is observed that the proposed method has better precision and recall.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information Filtering;* 1.5.4 [Pattern Recognition]: Applications – Text processing.

General Terms

Algorithms, Performance, Experimentation

Keywords

Text Classification, Naïve Bayes Classifier, Supervised learning, Feature selection, chi-square statistics.

1. INTRODUCTION

Text classification is the task of automatically sorting a set of documents into categories from a predefined set. Efficient text categorization systems are beneficial for many applications such as information retrieval, classification of news stories, text filtering, spam email filtering and content management in industries. A number of machine learning, knowledge engineering, and probabilistic-based methods have been proposed for text classification. The most popular methods include Bayesian probabilistic methods, regression models, example-based classification, decision trees, decision rules, Rocchio method, support vector machines and association rule mining [3], [11], [12].

Feature selection is the technique used in machine learning for selecting a subset of relevant features for building robust models. Feature selection gives a better understanding of data by giving their important features. Features for text categorization problems could be anything like words, phrases or sentences. Feature selection is an important step in text categorization problems; not all the features in a document are required to classify it. Feature selection eliminates irrelevant and redundant words of text and thereby reduces the dimensionality of documents. A good feature selection technique can computationally improve the learning algorithms [13]. A number of feature selection metrics based on information theory and statistics have been explored. Feature selection techniques could be classified as one sided metrics and two sided metrics. One sided metrics relate the features of a category with sign, positive and negative whereas two sided metrics relate features without sign, all features are considered equal. A term t_1 is a positive feature for a category C_1 when its presence increases the probability of the document to be in the category C_1 and t_1 is a negative feature for a category C_1 when its absence in a document increase the probability of the document to be in C_1 [12]. Examples for one sided metrics are Correlation Coefficient and Odds Ratio. Some of the well known two sided metrics are Information Gain and Chi Square methods referred as CHI in this paper.

Generally, all the text categorization algorithms are tested with documents from some of the categories in 20newsgroup or Reuters benchmark set. It is observed that the classifying accuracy is more when experiments were carried out with categories that are not related whereas the accuracy is less when the chosen categories are related. In our experiments, when the categories chosen were alt.athesim, comp.graphics, rec.motorcycles, rec.sport.hockey and talk.politics.mideast from 20newsgroup the classification accuracy was higher. However, the results are not superior in the experiments with the categories alt.atheism, comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware and comp.windows.x.

Performance of the naïve Bayes classifier, ever popular for its ease of programming is highly sensitive to feature selection [2], [13]. In this paper, the performance of the classifier is analyzed with features selected by CHI and CHIR algorithms. CHIR algorithm could discriminate positive and negative features for a class, experiments were conducted with only positive and some negative features selected by CHIR.

2. Dimensionality Reduction and its Importance

In most existing document classification algorithms, documents are represented as vector space model, which treats a document as a "bag of terms" [1]. A major characteristic feature of this representation is the high dimensionality of the feature space. The classification algorithm could not work efficiently in high dimensional spaces due to the inherent sparseness of the data [1]. All the features in the document are not important for classification; classification algorithms may even be misguided when there are more irrelevant features than relevant ones. Feature selection algorithms aims at selecting a subset of features for improving prediction accuracy and decreasing the size of the structure without significantly decreasing prediction accuracy of the classifier built using only the selected features [14]. Traditional feature selection methods are either supervised or unsupervised. Supervised selection methods using information gain and χ^2 statistics perform better than the unsupervised methods using document frequency and term strength [7].

One of the major problems of all the classifiers lies in handling rare categories that contain only a few documents. When a category has only a few training documents, informative words that are useful to determine the category are buried in the relatively large number of noise terms [2].

2.1 χ^2 max Feature Selection(CHI)

CHI is a supervised learning technique that determines the features set by ranking their χ^2 statistics values, which is done by comparing the observed co-occurrence frequencies in a 2 way contingency table with the expected frequencies, when they are assumed to be independent. Suppose that the corpus contains n labeled documents, and they fall in m categories, after stop word removal and stemming, distinct terms are extracted from the corpus and χ^2 value is determined. To determine the χ^2 value a 2X2 contingency table is formed for each term as in Table I. For example, when there are six categories with 500 documents of each class in training corpus, and the term w occurs as in Table I, CHI method works as discussed.

TABLE I

A 2X2 TERM-CATEGORY CONTINGENCY TABLE

	C	⌐c	∑
w	450	500	950
⌐w	50	2000	2050
∑	500	2500	3000

Expected frequency E (i, j), where i representing the presence or absence of a feature and j representing whether the document belongs to a category can be calculated as [1]:

$$E(i,j)=\frac{\sum_{a\in\{w,\neg w)}O(a,j)\sum_{b\in\{c,\neg c)}O(i,b)}{n} \qquad (1)$$

The χ^2 statistics is defined as:

$$\chi^{2}_{w,c}=\sum_{i\in\{w,\neg w\}}\sum_{j\in\{c,\neg c\}}\frac{(O(i,j)-E(i,j))^{2}}{E(i,j)} \qquad (2)$$

The "degrees of freedom" for a 2X2 contingency table is calculated as (r-1) X (c-1) where r is the number of rows in the contingency table and c is the number of columns in the table. The value obtained with (2) is compared with the value in the standard χ^2 tabulation for the determined degrees of freedom with confidence level 0.1%. The value in the table for the degrees of freedom 1 with 0.1% confidence level is 10.83. If the determined value is lesser than the value in the tabulation then null hypothesis is used to decide that the term and category are independent. Otherwise it is assumed that, there is a dependency between the term and the category (i.e.) alternative hypothesis is considered.

When it is assumed that there is a dependency between the term and the category, goodness-of-fit is used to decide the dependency. For a corpus with m classes, CHI defines the term-goodness of a term as the maximum among the categories defined in (3) [1]:

$$\chi^{2}{}_{\max}(w)=\max_{j}\{\chi^{2}{}_{w,cj}\} \qquad (3)$$

Where $p(c_j)$ is the probability of the documents to be in the category c_j. Drawback of CHI algorithm is that, it has determined only whether there is a dependency between a term and a category and not the type of dependency [1]. Same χ^2 value is obtained for a term occurred as in Table II. It may be noted that Table II is complement of Table I.

TABLE II

A 2X2 TERM-CATEGORY CONTINGENCY TABLE

	c	⌐c	∑
w	50	2000	2050
⌐w	450	500	950
∑	500	2500	3000

It may be observed from Table I that w is a positive feature for the category c, since 9/10 of the documents in c contains the term w and 4/5 of the documents not in c do not contain w. When Table II is analyzed it is observed that only 1/9 of the documents in C contain the term W and 8/9 of the documents with term W are not in category C. From Table II one may deduce W as a negative feature for category C rather than as a positive feature. Sign of dependency either positive or negative is not determined by CHI algorithm, terms with negative dependency are also considered equally important and selected as features for a category.

2.2 CHIR Algorithm

CHIR is a supervised learning algorithm based on χ^2 statistics, which not only determines the dependency between a term and a category but also the type of dependency. To evaluate the type of dependency, a new measure $R_{w,c}$ is defined in CHIR as [1]:

$$R_{w,c}=\frac{O(w,c)}{E(w,c)} \qquad (4)$$

When there is no dependency between the term w and the category c, then the value of $R_{w,c}$ is close to 1. If there is a positive dependency then the observed frequency is larger than the expected frequency, hence value of $R_{w,c}$ is larger than 1 and when there is a negative dependency $R_{w,c}$ is smaller than 1. Based on χ^2 statistics and $R_{w,c}$ a new definition for term-goodness for a corpus with m classes is given in CHIR algorithm as [1]:

$$r\chi^2(w) = \sum_{j=1}^{m} p(R_{w,cj})\chi^2_{w,cj} \text{ with } R_{w,c_j} > 1 \qquad (5)$$

where $p(R_{w,c_j})$ is the weight of $\chi^2_{w,cj}$ in the corpus. In terms of R_{w,c_j}, $p(R_{w,c_j})$ is defined as:

$$p(R_{w,c_j}) = \frac{R_{w,c_j}}{\sum_{j=1}^{m} R_{w,c_j}} \text{ with } R_{w,c_j} > 1 \qquad (6)$$

Larger value of $r\chi^2(w)$ indicates that the term w is more relevant to the category. The CHIR feature selection algorithm consists of the following steps:

1. For each distinct term in each category, determine its $r\chi^2$ value.

2. Sort the terms in descending order of their $r\chi^2$ value.

3. Select the top ‘q’ terms as feature for the current category.

When documents from closely related categories are taken for classification, it is common that some of the positive features of a category are negative features for some other categories. For example, the term “RAM” may be a positive feature for the category comp.sys.ibm.pc.hardware, since this term is common to appear in most of the documents related to hardware. Whereas if the same term “RAM” appears in few documents of the category comp.windows.x, then it must be a negative feature for the category. Feature selection by CHI, ignores the type of dependency and the term appears as a feature in both the categories, which misleads the classifier and decreases the classification accuracy.

3. Proposed Algorithm

Naïve Bayes classifier is a well-known practical probabilistic classifier that has been employed in many applications. It assumes that all attributes (i.e.) features of the training documents are independent of each other given the context of the class [2]. It has been shown that naïve Bayes under zero-one loss performs surprisingly well in many domains in spite of the independence assumption [10]. From Bayes’ theorem, the probability that a document with vector $X =< x_1,\ldots..,x_m >$ (where $x_1, x_2,\ldots,x_m$ are features in the document) belongs to category C is [4], [5], [6]:

$$p(C \mid X) = \frac{p(C).p(X \mid C)}{p(X)} \qquad (7)$$

Since the denominator does not depend on the category, naïve Bayes classifies each document that maximizes $p(C).p(X \mid C)$. Two popular versions of naïve Bayes classifiers are multivariate and multinomial model, the results given in this paper are output of multivariate Bernoulli model [9], [10].

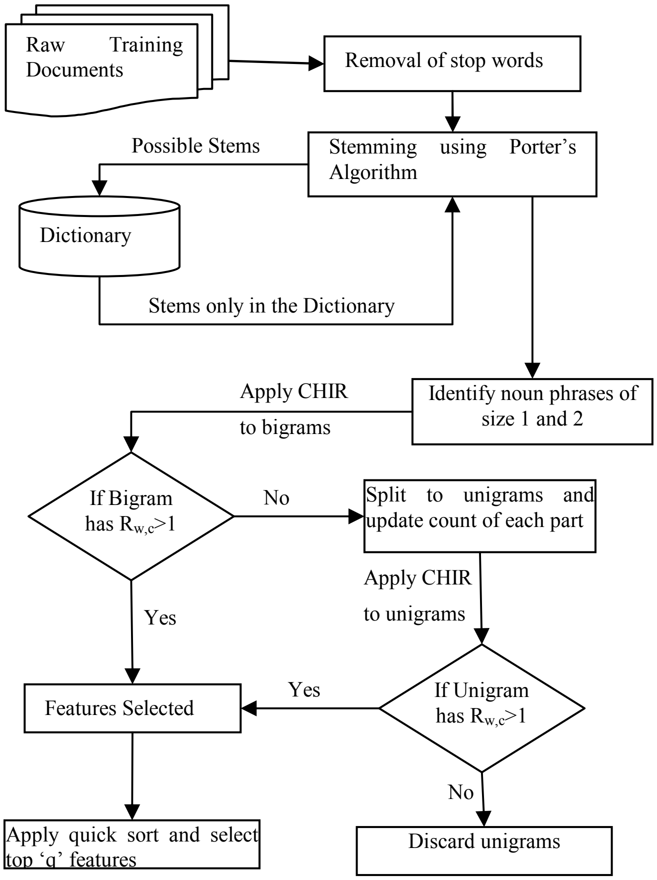
Raw Training Documents
Removal of stop words
Stemming using Porter’s Algorithm
Possible Stems
Dictionary
Stems only in the Dictionary
Identify noun phrases of size 1 and 2
Apply CHIR to bigrams
If Bigram has $R_{w,c}$>1
No
Split to unigrams and update count of each part
Apply CHIR to unigrams
Yes
Features Selected
Yes
If Unigram has $R_{w,c}$>1
No
Apply quick sort and select top ‘q’ features
Discard unigrams

Fig. 1 Flow of the training algorithm

The proposed algorithm preprocesses the training documents by removing stop words and stemming. Existing stemming algorithms does not perform a perfect stemming; they either overstem or understem a word. Most of the stemming algorithm gives more than one stem for a word; these issues are handled by choosing only words that occur in the dictionary. After stemming unigrams and bigrams (noun phrases of size two) are collected from training corpus and their counts are retained independently. The bigram features are selected by applying CHIR algorithm on the bigram collection, the inferior bigrams are splitted into unigrams, and the count of the obtained unigrams is updated as discussed in Table III.

TABLE III

Proposed feature selection algorithm

Input: D: Training set of documents of size N_d

```
        C: Set of document categories in D of size $N_c$
        Dic: Dictionary used.
        q : Required number of features
Output: List of selected features (uni and bigrams)
Procedure: Feature_Selection
    For all $d_i \in$ D do
    $LS_i \leftarrow$ List of sentences(S) and part of sentences of D.
    For each S $\in LS_i$ do
    $LU_s \leftarrow$ List of Unigrams in sentence S
    $LB_s \leftarrow$ List of Bigrams in sentence S
    For each $LU_{si} \in LU_s$ and $LB_{si} \in LB_s$
        LU $\leftarrow$ Insert or Update_Count($LU_{si}$)
        LB $\leftarrow$ Insert or Update_Count($LB_{si}$)
    End
    End
    End
    For all $LB_i \in$ LB do
        Find $R_{w,c}$
        If $R_{w,c} > 1$
            GOT $\leftarrow$ Estimate_Goodness_Of_Term($LB_i$)
            F $\leftarrow$ F $\cup$ $LB_i$
          End
        Else
              $LB_{i1}$ , $LB_{i2}$ $\leftarrow$ Split($LB_i$)
              LU $\leftarrow$ Insert or Update_Count($LB_{i1}$)
              LU $\leftarrow$ Insert or Update_Count($LB_{i2}$)
        End
    End
    For all $LU_i \in$ LU do
        Find $R_{w,c}$
        If $R_{w,c} > 1$
            GOT $\leftarrow$ Estimate_Goodness_Of_Term($LU_i$)
            F $\leftarrow$ F $\cup$ $LU_i$
          End
    End
    Quick_Sort(F)
    return 'q' Top_Features(F)
```

4. Experimental Results and Discussion

The text categorization approach proposed in this paper has been implemented and evaluated with extensive experimentations on six categories of 20 newsgroup benchmarks. Out of the six categories, five are related to computer science and about operating systems and hardware. It is difficult to determine the features for these types of data sets. Negative features are significant when documents are categorized to related categories. As discussed in section II CHI method chooses even sparse terms of a category as its features.

4.1 Evaluation Methodology

A number of metrics used in text categorization are evaluated and measured for categorization effectiveness. The well known precision and recall metrics are used in this paper to analyze the results. Precision is defined as the ratio of correctly assigned category C documents to the total number of documents classified as category C. Recall is the ratio of correctly assigned category C documents to the total number of documents actually in category C. Let a, b, c represent the values as follows:

- a – number of C_i documents classified into C_i
- b – number of non-C_i documents classified into C_i
- c – number of C_i documents classified as non-C_i

Precision = a / (a + b)

Recall = a / (a + c)

Some combinations of precision and recall can be more effective in measuring classifier performance. One of such measures is F-measure, which is used in this paper for evaluation. F-measure is determined by calculating the harmonic mean of precision (P) and recall (R) and is computed as:

$$F = 2PR / (P + R)$$

4.2 Experimental Results and Discussion

Experiments were conducted for 10 trials with randomly chosen training corpus. The size training set was varied from 10 to 75 percentage of the total experiment set. The results shown in the charts are the average of the trials.

Charts in Fig 2 to Fig 7 compares the F-measure of the categories, when only unigrams were chosen by the algorithm and while unigrams and bigrams were chosen as features by the algorithm. Including bigram features in the feature set improves the precision and recall of the classifier.

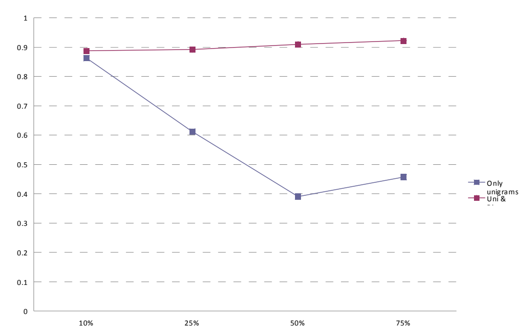

Fig 2 F Measure for alt.atheism

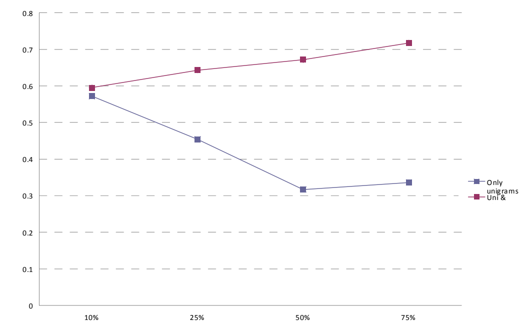

Fig 3 F Measure for Comp.graphics

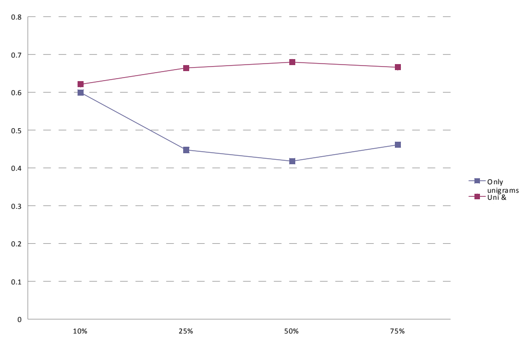

Fig 4 F Measure for Comp.os.ms.windows.misc

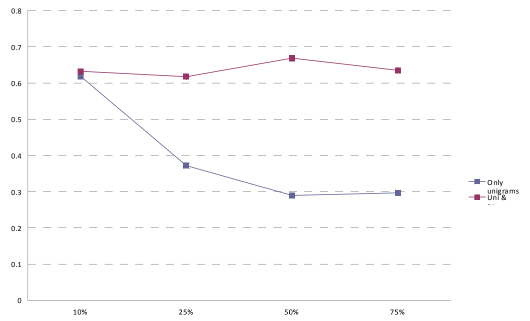

Fig 5 F Measure for Comp.sys.ibm.pc.hardware

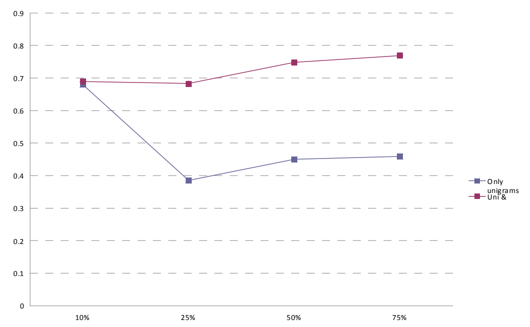

Fig 6 F Measure for Comp.sys.mac.hardware

5. CONCLUSION

Most of the recent text classification research focuses on addressing specific issues such as feature selection, clustering and dimensionality reduction. This paper proposes a novel TC approach with features selected by CHIR algorithm, a statistical based approach. It has been observed that, unigram and bigram features selected by this method improve the accuracy of the naïve Bayes classifier. Results of the classifier could be improved for smaller training sets.

6. REFERENCES

[1] G Yanjun Li, Congnan Luo, and Soon M. Chung, "Text Clustering with Feature Selection by Using Statistical Data," IEEE Transactions on Knowledge and Data Engineering, Volume 20, Issue 5, pp 641 – 652, May 2008.

[2] Sang-Bum Kim, Kyoung-Soo Han, Hae-Chang Rim, and Sung Hyon Myaebg, "Some Effective Techniques for Naïve Bayes Text Classification," IEEE Transactions on Knowledge and Data Engineering, Volume 18, No. 11, pp 1457 – 1466, November 2006.

[3] Hisham Al-Mubaid and Syed A. Umair, "A New Text Categorization Technique using Distributional Clustering and Learning Logic," IEEE Transactions on Knowledge and Data Engineering, Volume 18, No. 9, pp 1156 – 1165, September, 2006.

[4] Andrew McCallum and Kamal Nigam, "A Comparison of Event Models for Naïve Bayes Text Classification," in AAAI-98 Workshop on Learning for Text Categorization, 1998.

[5] Vangelis Metsis, Ion Androutsoplos and Georgios Paliouras, "Spam Filtering with Naïve Bayes – Which Naïve Bayes?," in Proc. CEAS 2006, Third Conference on Email and Anti-Spam, Mountain View, California USA, July 27-28,2006.

[6] Jason D. M. Rennie, Lawrence Shih, Jaime Teevan, and David R. Karger, "Tackling the Poor Assumptions of Naïve Bayes Text Classifiers," in Proc. of the twentieth International Conference on Machine Learning, 2003.

[7] Ciya Liao, Shamim Alpha, Paul Dixon "Feature Preparation in Text Categorization", Oracle Corporation, 1997.

[8] Yiming Yang and Jan O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," in Proc. 1997. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.32.9956.

[9] Jason D. M. Rennie, "Improving Multi-class Text Classification with Naïve Bayes", M. S. Thesis, Submitted at Dept. of EE and CS, Massachusetts Institute of Technology, 2001.

[10] P. Domingos and M.J. Pazzani, "On the optimality of the Simple Bayesian Classifer under zero-One Loss", Machine Learning, volume 29, nos. 2/3, pp. 103 – 130, 1997.

[11] Fabrizio Sebastiani "Text Categorization", in Proc. Text Mining and its Applications to Intelligence, CRM and Knowledge Management, 2005.

[12] Zhaohui Zheng, Xiaoyun Wu and Rohini Srihari, "Feature selection for Text Categorization on Imbalanced Data", ACM SIGKDD Explorations Newsletter, Special Issue on learning from Imbalanced Data, Volume 6, Issue 1, pp. 80 – 89, June 2004.

[13] George Forman, "Feature Selection : We've barely scratched the surface" An essay requested for IEEE Intelligent Systems, Trends and Controversies, 2005.

[14] M. Dash, H. Liu, "Feature Selection for Classification" , Intelligent Data Analysis, 1997 pp 131-156.