International Journal of Advance in Computer Science & its application – IJCSIA 2018
Copyright © Institute of Research Engineers and Doctors , SEEK Digital Library
Volume 7 : Issue 2 [ISSN : 2250-3765] - Publication Date: 25 June , 2018

An Analysis of Feature Selection Methods for Multiclass Text Classification

[Sanjay Agarwal, Mayank Kalbhor]

Abstract: **-To classify objects into different classes, feature plays a vital role. So identification of best features is a backbone of classification process. In text classification, features are simple words, having very large dimension so finding the most appropriate feature set is a big challenge. This paper includes analysis of some feature selection methods for multi class text classification and checks their results on different classifier for an email classification. We run our experiments on 20NewGroups and PU corpora datasets. Experiments are done on some well-known feature selection method like Term Selection, Document Frequency, Mutual Information, Odds Ratio, Chi square and etc. This paper concludes that Mutual Information and Chi square are most appropriate for text classification.**

Keywords: - Feature Selection, Text Classification, Multi-Class Classification

I-INTRODUCTION

Classification is a supervised learning process in which number of classes for some objects is known and some training and testing dataset is also given. Goal of classification process is to teach machine so that machine can be capable of classifying object into different classes. Features are very important in classification. Feature is a property of those objects by which they can be classified, example classification of balls into different colour ball. Here in previous example features may be a colour of ball. The second important thing is how we train our machine i.e. training algorithm. If we have some dataset of chair and tables so features may be height and weight. Calculation of height and weight or knowledge of colours can be used to build the classification algorithm.

For multiclass classification of text document aim is to categorize a simple text document into one of the predefined category. Automatic text classification is very important for any web based service. Consider someone wants to buy "smartphone" as gift, so consider the search results of Google and yahoo search engine. In yahoo result you will get some categorization of similar products which is appropriate for shopping. Here broad indexing & speedy search alone are not enough. But organizational view of data is critical for effective retrieval. Some leading companies use more than 100 manpower for manual text classification.

Text classification is also used to handle spam emails, classify large text collections into typical categories, used to manage knowledge and also to help Internet search engines. A major characteristic of text categorization is high dimensionality of the feature space; the native feature space consists of hundreds of thousands of terms for even a moderate sized text collection [18]. This paper considers two main problems first email classification and second automatic text classification.

A Problem Statement: -

E-mail classification a supervised learning problem. It can be formally stated as follows. Given a training set of labeled e-mail documents

Dn={(di,ci)}i=1-n,

Where di is an e-mail document from a document set D and ci is the label chosen from a predefined set of categories C, the goal is to prepare a classifier or hypothesis

H1: D → C

That can correctly classify new, unseen e-mail documents Dtest; Dtest ⊄Dtrain. Here, task is either binary classification or multi class classification. In second problem we have some raw text document and goal is to classify them into different folders that may be sports, social, national and many more. The whole paper generalizes the problem of email classification and also provides review of some well-known feature selection methods.

B Structure of a usual Email Classifier: -

The information contained in an email can be divided into the header i.e. general information on the message, such as the subject, sender and recipient and body i.e. contents of the message. Pre-processing steps are required before the available information can be used by a classifier. The steps involved in the extraction of data from a message can be illustrated as: -

Step 1. Feature selection which includes tokenization of words, lemmatization of words, stop word removal and lastly select most appropriate features.

Step 2. Feature extraction is a process which reduces the dimensionality of the input feature vectors. One simple method for reducing the feature space is by building a dictionary of word's and replaces word with some integer.

Dr. Sanjay Agarwal
Professor, NITTTR Bhopal
INDIA
sagrawal@nitttrbpl.ac.in
agsanju@gmail.com

Mr. Mayank Kalbhor
Research Scholar, RGPV, Bhopal
INDIA
mnk.kalbhor@gmail.com

International Journal of Advance in Computer Science & its application – IJCSIA 2018
Copyright © Institute of Research Engineers and Doctors , SEEK Digital Library
Volume 7 : Issue 2 [ISSN : 2250-3765] - Publication Date: 25 June , 2018

Step 3. Email Classification is supervised learning problem where some training data (labelled class email) is given and aim is to prepare a classifier which can correctly classify test data correctly.

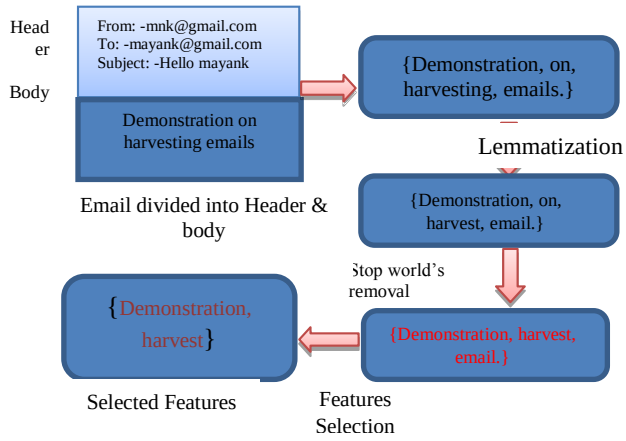

Figure 1 Feature selection from email

II-FEATURE SELECTION METHODS: -

In email messages, generally the info contained in messages are terribly complex (i.e. with different attachments like images and document file etc., and also having many languages), multidimensional (single feature with different categories), and depicted by an out sized range of features. As we are using words as our features, so there are many categories of those words. As a result, the employment of dimensionality reduction strategies is useful among the classification task to avoid the curse of dimensionality [1]. In general, the dimensionality reduction ways are divided into two categories. The primary one is Feature selection (FS), where the dimensionality is reduced by selecting a set of original features, and also the removed features don't seem to be utilized in the computations any further. The aim of FS methods is to compute a group of k features from a group of d, by increasing a given criterion.

Consider text dataset corpora having three classes' C ∈{c1, c2, c3} and aim here to classify a bunch of document D ∈{D1, D2, D3… Dn} in one of possible class. Here words are considered as feature vector. To achieve better results of text categorization first stop words of text documents are removed. This can be done by making a list of stop words and eliminate each stop word from every text documents. Then appropriate lemmatization method is applied for taking words to its original form. Now, each document has structure having some collection of words ex. D1 = {Hello, late, night, party, free, invitation}. Goal of feature selection method is to find best features that can define a class or that can distinguish message from other classes. Below discussion of different feature selection method on above assumed text corpora is presented.

A. Document Frequency (DF): -

Document frequency is a very simple feature selection method. Document frequency for any term can be calculated by counting number of documents in which a particular term or feature occurs [2]. Here we set a threshold of Document Frequency and all the terms whose document frequency count is below of this predefined threshold are removed from the set of terms in the dataset. The DF of a term can be calculated by below given formula:

$$D(t_i) = |\{m_j, m_j \in M, \text{and } t_i \in m_j\}| \quad (1)$$

In this equation M resemble the complete training set of messages (part of dataset used for training purpose), and mj is a message in M. Here we have a assumption that rare terms contains very less information about classification. So, importance of DF is to get rid of rare terms, therefore if we remove them it doesn't have any major effect on overall performance.

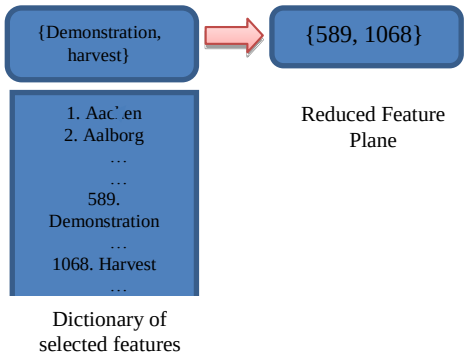

Figure 2 illustration of Feature Extraction

Suppose threshold frequency is set to j so all the words having document frequency less then j are removed. This way we will get some collection of words as our feature set.

$$f = \{f1, f2 \ldots fn\}$$

These feature set can be used for text classification and feature and documents are provided into the next step that is feature extraction method.

Implementation of document frequency is very simple task. Suppose D document set for training is provided, document frequency of particular term is just number of document in which that term occurs. DF is usually employed

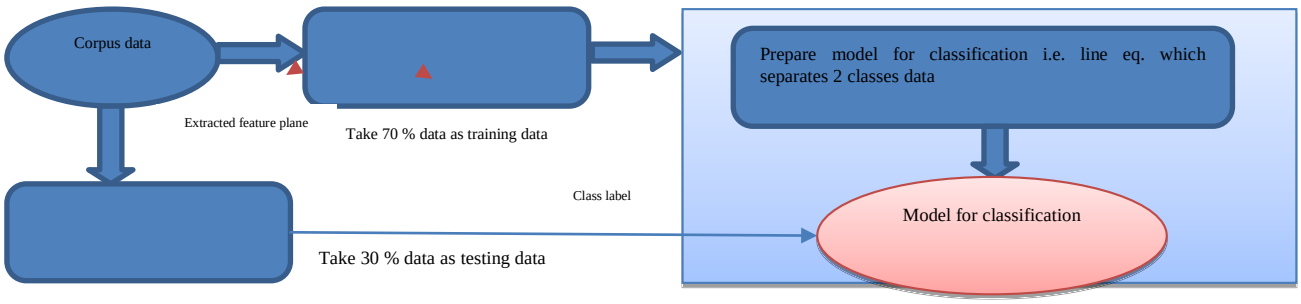

Figure 3 illustration of Text Classification

with Inverse Term Frequency (ITF), can be calculated by number of times that term occurs in particular document [3].

B. Information Gain: -

Information Gain (IG) is generally referred as term goodness criteria in category prediction, splitting criteria and in some other machine learning problem. In DF-ITF no information regarding the term goodness is provided for selecting features, but in IG aim is to find importance of particular word for email categorization [2]. It measures the number of bits of information for category prediction by knowing the presence or absence of a term in document. Let 〖{C_i}〗 _(i=1)^m denotes the set of categories in the target space. So, the IG of term t may be defined as:

IG(t)= -i=1mPr(Ci)log Pr(Ci) + Pr(t)i=1mPr(Ci|t)log Pr(Ci|t) + Pr(t')i=1mPr(Ci|t')log Pr(Ci|t') (2)

Here Pr(Ci) is probability of class Ci, Pr(Ci|t) is probability of class Ci when t is given and Pr(Ci|t') is probability of class Ciwhen t is not present.

For simplification consider binary classification problem where aim to classify unknown email into spam or non-spam. So, IG of particular term t can be defined as:

IG(t)= -Pr(C)log Pr(C) +Pr(t)[Pr(Cs|t)log Pr(Cs|t) +Pr(Cn|t)log Pr(Cn|t)] + Pr(t')[Pr(Cs|t')log Pr(Cs|t') +Pr(Cn|t')log Pr(Cn|t')] (3)

Here C_s and C_n are class spam and non-spam respectively. For a given training corpus IG of each unique term is calculated by given formula and terms whose IG is less than predefined threshold are removed.

C. Mutual Information: -

Mutual information is mostly term in statistical language modelling of word association and connected application [18]. Consider we have table with column consist of term noted as t and class of document noted as c and A is number of time t and c co-occur, B is number of time the t occur while not c, C is range of time c occur while not t, and N is the total number of documents, then we can define mutual information between term t and class c as below given equation.

I(t,c)= log A*N(A+C)*(A+B) (4)

I(t,c)= log Pr(t∧c)Pr(t)*Pr(c) (5)

If term is independent to the class then mutual information is obviously zero. The main drawback of this algorithm is for a term with equal conditional probability rare terms have higher score than common terms [6].

D. Chi Square (χ^2 measures): -

The χ2 statistic measures the independence or absence of term t and class c and might compared to the χ2 distribution with one degree of freedom to judge extremeness [17]. Again let two method contingency table of a term t and class c, wherever A is that the number of time t and c co-occur, B is number of time the t occur while not c, C is number of time c occur while not t, D is that the number of times neither c or t occur, and N is the total number of documents, the term goodness measure is outline to be:

χ2(c,t)=N*(AB-CD)^2/(A+C)*(B+D)*(A+B)*(C+D)

χ2(c,t)max=maxi(χ2(ci,t)) (6)

If t and c has no dependency then χ2 statistic will be 0. The major difference between CHI and MI is that χ2 is a normalized value. Hence, the χ2 statistic is known not to be reliable for the terms which has low occurrence [6].

Again follow the same steps as in information gain just for all set of words first find χ2 statistic value and then short it by ascending order and remove below threshold defined.

2.5. Term Frequency Variance (TFV): -

TFV technique is proposed to select features with high variance which has some additional information. Here variance can be defined as difference between two terms or numbers in a dataset. TFV methods is originally proposed by Koprinska et al.[4]. Like information gain, TFV method is category dependent. For each term f we compute w the term document frequency (tf) in every class and variance can be calculated as

TFV (f) = i=1k[tf(f,ci)-mean_tf(f)]^2 (7)

Here as we know features which have high variance in all possible classes will be more informative and so have more chances to get select.

III. SURVEY ON FEATURE SELECTION METHODS FOR TEXT CATEGORIZATION: -

Hamood Alshalabi et al. propose "Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization" in which they work on Malay text categorization problem. They collected categorized Malay documents from online Malay newspapers archives. So, Malay corpus manly made up of Bernama, and Utusan on-line newspaper. This corpus contains some 3040 documents that are different in size. Also, they are divided into eight categories includes Arts, Medicine, Politics, Business, Crime, History, Religion, Healthy and Sport. Experiments were performed on two feature selection methods namely Information Gain and Chi-Square on three classification methods (k-Nearest Neighbor, Naïve Bayes and N-gram). Features were selected from different size feature space i.e. 100 to 600. All algorithms were evaluated using 5-fold cross-validation. Performance of these classification methods were measured by the use of Macro-averaged (Macro-F1) measure, which combines Recall and Precision. Experiment result showed that Chi-square feature selection method performed the best for terms selection [7].

In 2013 Jaesung Lee et al. proposes "Feature selection for multi-label classification using multivariate mutual information" method [12]. In which Authors propose a feature selection method for multi-label classification that naturally derives from mutual information between selected features and the label set. In this paper authors suggested decomposing the joint entropy of two sets: H(S, L) after finding mutual information.

In March 2014 Deqing Wang et al. proposes "t-Test feature selection approach based on term frequency for text categorization" method, which is a variant of term frequency method [14]. Author proves that most of the methods of

feature selection related to document frequency having drawback of not considering of lower frequency terms although those are important. Authors concentrated on how to construct a feature selection function based on term frequency. Here the t-test function is used to measure the diversity of the distributions of a term frequency between the specific category and the entire corpus. So, additionally for each feature they calculated the class diversity to select the best features. They have performed their experiments on dataset Reuters-21578 and 20Newsgroup with SVM, weighted KNN, and centroid based classifier.

IV. DISCUSSION AND CONCLUSIONS: -

In this paper, a comprehensive review of recent feature selection approaches for multiclass text categorization was presented. Also quantitative analysis of the use of feature selection algorithms and datasets was conducted. It was verified that the information gain and Chi measures are the most commonly used method for feature selection. Among the several publicly available datasets, the LingSpam and PU corpora stand as the most popular, while the 20Newsgroup corpora for multi class classification is moderately popular at present. In terms of evaluation measures, precision recall and accuracy, which are given, respectively, by the relative number of Spam and legitimate messages correctly classified, are suggested as the preferred indices for evaluating filters.

TABLE 1 RESULT COMPARISON

Corpora	Classifier	Feature Selection Method	F-measure In %	Accuracy In %
PU1	k-NN	CHI	93.98	94.76
		IG	92.45	94.23
		DF	91.76	93.87
	NB	CHI	94.21	96.98
		IG	94.23	97.45
		DF	93.87	94.76
20Newsgroup	k-NN	CHI	92.78	90.78
		IG	94.91	91.91
		DF	92.12	89.12
	NB	CHI	94.13	92.13
		IG	95.60	90.60
		DF	93.23	91.23

As shown in Table 1, experiments are performed on email corpora PU1, PU2, PU3, PUa and on one newspaper corpora and result were calculated in the form of F-measures, which can be calculated by recall and precision as equation (8).

$$F\beta = (1+\beta 2)\ (Ps*Rs)/(\beta 2\ (Ps+Rs)) \qquad (8)$$

Here, Ps is the precision count and Rs is the recall of the classifier and β is set to 1.

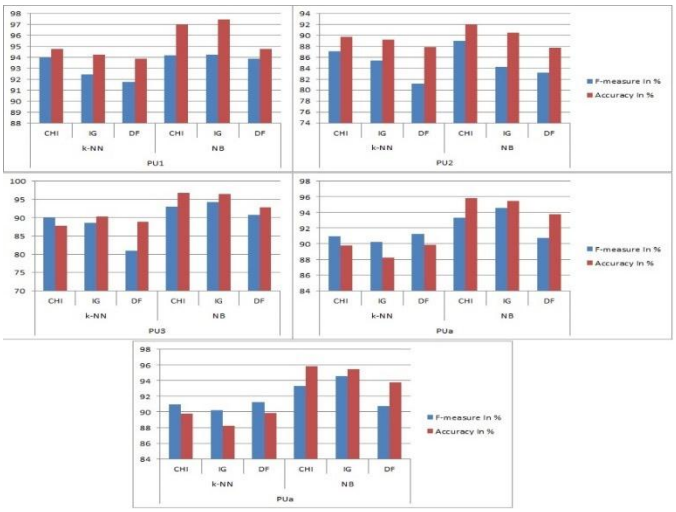

Figure 4 Result comparison

χ2 statistic found as the most useful method for feature selection from text document. Finally conclusion is automatic text categorization is need for betterment of web services. For text classification words are considered as the features so selection of most appropriate features is very important.

References

1. Michael W. Berry and Jacob Kogan[2010]. Text Mining: Applications and Theory John Wiley & Sons, Ltd.
2. Ren WangI, Amr M. Youssef, Ahmed K. Elhakeem "On Some Feature Selection Strategies for Spam Filter Design", IEEE, MAY 2006.
3. Yiming Yang and Jan O. Pedersen "A Comparative Study on Feature Selection in Text Categorization" 1997.
4. I.Koprinska, J. Poon, J. Clark, and J. Chan, "Learning to classify e-mail," Inform. Sci., vol. 177, pp. 2167–2187, 2007.
5. Y. Yang, J. Pedersen, A comparative study on feature selection in text categorization, in: Proc. 4th International Conference on Machine Learning, 1997.
6. T. M. Cover, J. A. Thomas. Elements of information theory. New York: John Wiley and Sons, 1991.
7. Hamood Alshalabi, Sabrina Tiun, Nazlia Omar, Mohammed Albared, "Experiments on the Use of Feature Selection and Machine Learning Methods in Automatic Malay Text Categorization", the 4th International Conference on Electrical Engineering and Informatics (ICEEI 2013), Procedia Technology 11 (2013) 748 – 754.
8. CAO Yinga, Duan run-yingb, "Novel top-down methods for Hierarchical Text Classification", International Conference on Advances in Engineering, Procedia Engineering 24 (2011) 329 – 334.
9. Kunlun Li, "Multi-class text categorization based on LDA and SVM", Advanced in Control Engineering and Information Science, Procedia Engineering 15 (2011) 1963 – 1967.
10. Harun Uğuz, "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm", Knowledge-Based Systems Volume 24, Issue 7, October 2011, Pages 1024–1032.
11. Xi Zhao, "Feature Selection with Attributes Clustering by Maximal Information Coefficient", Information Technology and Quantitative Management, ITQM, Procedia Computer Science 17 (2013) 70 – 79.
12. Jaesung Lee, "Feature selection for multi-label classification using multivariate mutual information", Pattern Recognition Letters Volume 34, Issue 3, 1 February 2013, Pages 349–357.
13. Xiaofei Zhoua, "Text Categorization Based on Clustering Feature Selection", Procedia Computer Science Volume 31, 2014, Pages 398–405.
14. Deqing Wang, "t-Test feature selection approach based on term frequency for text categorization" Pattern Recognition Letters Volume 45, 1 August 2014, Pages 1–10.
15. Tom Fawcett, "An introduction to ROC analysis", Pattern Recognition Letters Volume 27, Issue 8, June 2006, Pages 861–874.

16. yuanchun zhu and ying tan "A Local-Concentration-Based Feature Extraction Approach for Spam Filtering" ,IEEE in VOL. 6, NO. 2, JUNE 2011.
17. Xia Huosong. "The Research of Feature Selection of Text Classification Based on Integrated Learning Algorithm", 2011 10th International Symposium on Distributed Computing and Applications to Business Engineering and Science.
18. Guiying Wei. "Study of text classification methods for data sets with huge features", 2010 2nd International Conference on Industrial and Information Systems.

About Author (s):

Dr. Agrawal, Sanjay
Professor
Department of Computer Engineering and Applications, N.I.T.T.T.R., Bhopal (MP) INDIA

Mr. Mayank Kalbhor
Asst. Professor
Department of Information Technology, S. S. G. M. C. E.., Shegaon, (MH) INDIA