Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

(Special Issue for ITECE 2016)

IMPROVE FILTER METHOD FOR FEATURE SELECTION TECHNIQUE IN TEXT MINING

[1] Mital Vala [2] Prof. Jay Gandhi

[1] *M.E Student Dept of Computer Engineering, B.H.Gardi college of engineering & Technology, Rajkot, India.*

[2]*Assistant Prof Dept of Information Technology, B.H.Gardi college of engineering & Technology, Rajkot, India*

ABSTRACT***: Huge amount data on the internet are in unstructured texts can't simply be used for further processing by computer, therefore specific processing method and algorithm is required to extract useful pattern. Text mining is process to extract information from the unstructured data. A major difficulty of text classification is high dimensionality of feature space. Feature selection method used for dimension reduction. We use filter method for feature selection that use MFD algorithm but main drawback it is it necessary to assign value of F so it's work for only given features F we overcome this by proposed method in which F value select automatically by pre analyses data set extract unique feature and create feature set F so its work for all unique features and also reduce maximum redundancy by imposing threshold value.***

KEYWORDS***: Text mining, Feature selection, Filter method, wrapper method, embedded method, MFD algorithm.***

I INTRODUCTION

Text mining is process to extract information from the unstructured data. Text mining is also known to be data mining, data mining help to mine data of structure pattern while text mining help to handle data with unstructured form.Text mining task including text clustering, text classification, document summarization, entity modeling [1]. Text classification is task of automatically sorting set of document into categories from predefined set.

Text Classification tasks divided into 1) Supervised Document Classification: In Supervised Document Classification some external mechanism provides information on the correct classification for documents or to define classes for the classifier.

2) Unsupervised Classification: In Unsupervised Document Classification, the classification must be done without any external reference and the system do not have predefined classes [2]

high dimensionality of feature space seems to be major challenge in text mining which can be reduce by feature selection method. Our main motive is: "dimension reduction using feature selection method". Feature selection is a process that is based on criteria of data.

Reasons to use feature selection technique:
1. reducing the number of features, to reduce over fitting and improve the generalization of models.
2. To gain a better understanding of the features and their relationship to the response variables.

There are three classes of feature selection algorithms: 1) filter methods, 2) wrapper methods and 3) embedded methods. We are going to use filtration method. The features are ranked by the score and either selected to be kept or removed from the dataset. Example of some filter methods include the Chi squared test, information gain and correlation coefficient scores.

We use filter method for feature selection that use MFD algorithm but main drawback it is it necessary to assign value of F we overcome this by proposed method in which F value select automatically by pre analyses data set extract unique feature and create feature set F and also remove more redundancy by proposed method.

The rest of the paper is as follow: Section II discusses Text mining challenges. Section III Discuss feature selection method for dimension reduction .section IV discuss Maximum feature per document algorithm.. In section V discuss proposed algorithm .section VI discuss related work and section VII gives conclusion.

II TEXT MINING CHELLENGES

1) **Large textual data base:** No clear picture of what document suit the application.
2) **High dimensionality:** Thousand of word, only very small percentages is used in typical document.
3) **Several input modes**: Text intended for different consumer's i.e. Different languages, different format.
4) **Dependency**: Word and phrases create context for each other.
5) **Noisy data**.

III FEATURE SELECTION METHOD

Feature-selection methods used for reduction of the dimensionality of the dataset by removing features that are considered irrelevant for the classification. [5]The aim of feature selection is to select a subset of variables from the input which can efficiently describe the input data while reducing effects from noise or irrelevant variables and still provide good prediction results. One of the applications would be in gene microarray analysis .The standardized gene expression data can contain hundreds of variables of which many of them could be highly correlated with other variables .The dependant variables provide no extra information about the classes and thus serve as noise for the predictor. This means that the total information content can be obtained from fewer unique features which contain maximum discrimination information about the classes.[5] Hence by eliminating the dependent variables, the amount of data can be reduced which can lead to improvement in the classification performance. Feature selection in text classification focuses on identifying relevant information without affecting the accuracy of the classifier. Feature selection techniques can be classified into three categories: filtering techniques ,wrapper techniques and embedded technique.

(1) Filter method: Filter methods analyze intrinsic properties of data, ignoring the classifier. This methods can perform two operations namely ranking and subset selection. In ranking independent feature is evaluated by neglecting interactions among the elements while in subset selection, the final subset of features to be selected is provided. In some cases, these two operations are performed sequentially; in other cases, only the selection is carried out Filter methods suppress the least interesting variables.

Examples:

a) ***X^2 statistic***: Measures lack of independence between t and c and can be compared to the X^2distribution with one degree of freedom to judge extremeness [5].

$$X^2_{(t,c)} = \frac{D\times(PE-MQ)^2}{(P+M)\times(Q+N)\times(P+Q)\times(M+N)}$$

Where D = total number of documents

P = the number of documents of class c containing term t

Q = the number of documents containing t occurs without c.
M= the number of documents class c occurs without t.
N=the number of documents of other class without t.

b) ***Information gain:*** Information gain is frequently employed as a term goodness criterion in machine learning. The prediction of category is done by knowing presence or absence of term in document and by measuring number of bits of information.

c) ***Mutual information****:* Mutual information is a criterion commonly used in statistical language modeling of, word associations and related applications .This is able to provide a precise statistical calculation that could be applied to a very large corpus to produce a table of association of words [5]

2) Wrapper methods: Wrapper method use a predictive model to score feature subsets. Each new subset is used to train a model, which is tested on a hold-out set. Counting the number of mistakes made on that hold-out set gives the score for that subset. As wrapper methods train a new model for each subset, they are very computationally intensive, but usually provide the best performing feature set for that particular type of model[14].

3) Embedded method: Embedded method are a catch-all group of techniques which perform feature selection as part of the model construction process. Recently, embedded methods have been proposed to reduce the classification of learning. They try to combine the advantages of both previous methods. The learning algorithm takes advantage of its own variable selection algorithm. So, it needs to know preliminary what a good selection is, which limits their exploitation[14]

IV MAXIMUM FEATURE PER DOCUMENT ALGORITHM

MFD algorithm steps:

1) A training set dtr loaded .The set composed of d document.
2) For each feature the FEF values are calculated and stored in Sh , thus sh represent the FEFs Values are calculate.
3) The new set of features FS is computed. The hth feature is inserted in FS (Sh is the highest value among all features). However, if this feature is already in FS, it is ignored and the algorithm increments nf to search for another feature or continues to the next document (if nf > f). Sbestfeature assumes a negative value. This avoids future selection of this feature in the document di under analysis. After, the S vector has its values restored because the next document must deal with the original values. At the end of this phase, FS should be a vector with m values, and these and stored in Sh. Thus, Sh represents the importance of the hth feature.

V PROPOSED ALGORITHM

- In existing MFD algorithm we need to assign value of feature F every time and thus we get result only for given feature F but in our proposed algorithm we are going to implement in such way that it's scan the value of feature F automatically by pre analyze data set and find value of F create feature set F so its work for all unique features.

- In proposed algorithm we are going to Remove maximum redundancy by imposing threshold values

Proposed algorithm steps:

1) Pre analysis data sets, Extract unique feature and create feature set f
2) Load training set dtr
3) Decide threshold value
4) The new set of features FS is computed. The hth feature is inserted in FS (Sh is the highest value among all features). However, if this feature is already in FS, it is ignored and the algorithm increments nf to search for another feature or continues to the next document (if nf > f).
5) Check threshold value remove feature that below in threshold value
6) Sbest feature assumes a negative value. This avoids future selection of this feature in the document di under analysis. After, the S vector has its values restored because the next document must deal with the original values. At the end of this phase, FS should be a vector with m values, and these and stored in Sh. Thus, Sh represents the importance of the hth feature.

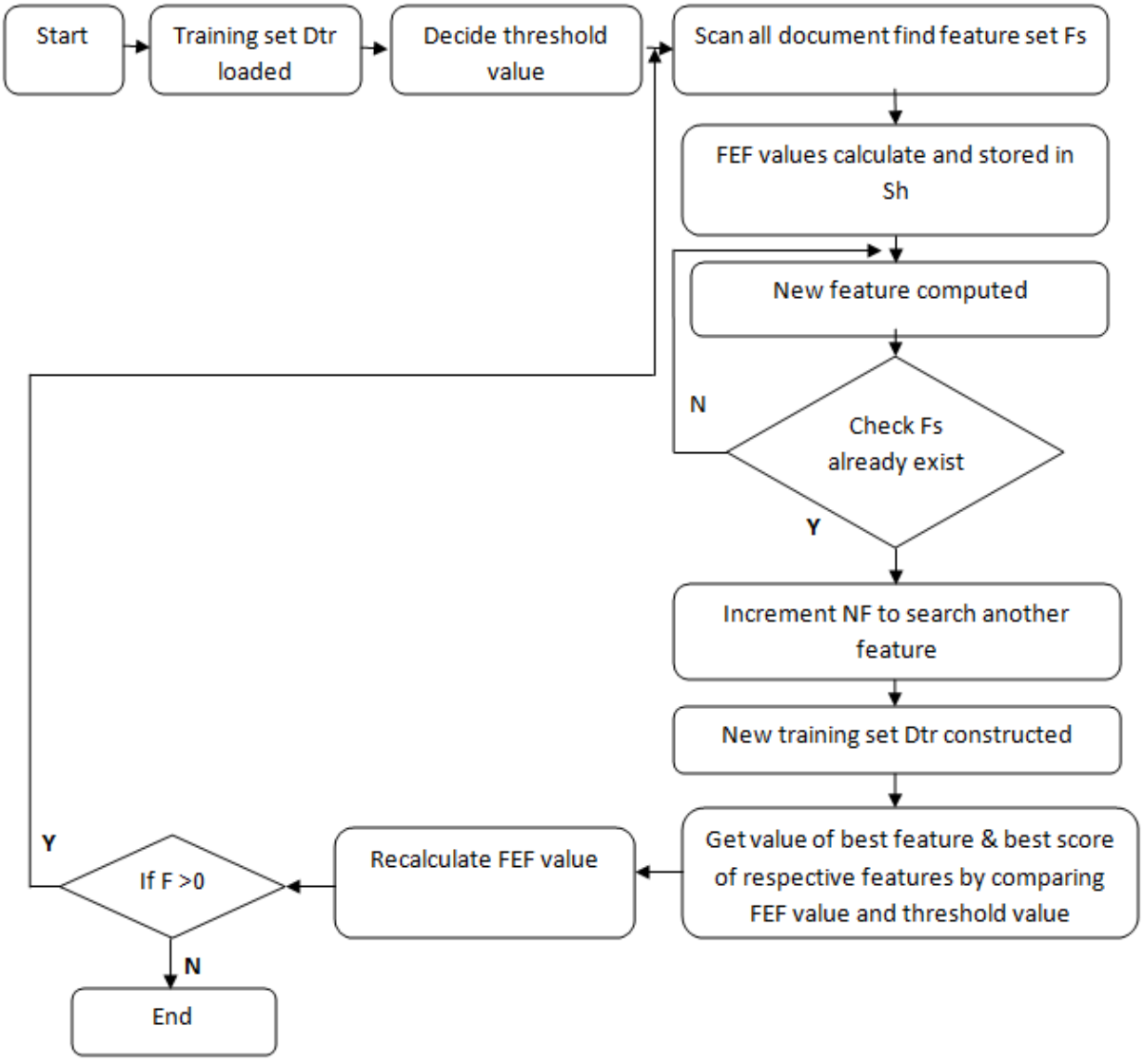

Fig .1 Flow chart of proposed algorithm

VI RELATED WORK

In this paper proposed two filtering methods for feature selection in text categorization, namely: Maximum f Features per Document (MFD), and Maximum f Features per Document – Reduced (MFDR). Both algorithms determine the number of selected features f in a data-driven way using a global-ranking Feature Evaluation Function (FEF). The MFD method analyzes all documents to ensure that each document in the training set is represented in the final feature vector. Whereas MFDR analyzes only the documents with high FEF valued features to select less features therefore avoiding unnecessary ones.[3]

This paper proposed A new feature ranking metric termed as relative discrimination criterion (RDC), which takes document frequencies for each term count of a term into account while estimating the usefulness of a term. RDC includes the information of term counts, which is ignored by other feature ranking metrics, while determining the rank of a term. In which The performance of RDC is compared with four well known feature ranking metrics, information gain (IG), CHI squared (CHI), odds ratio (OR) and distinguishing feature selector (DFS) using support vector machines (SVM) and multinomial naive Bayes (MNB) classifiers.[4]

In this paper proposed B&B feature selection in Reproducing Kernel Hilbert Space(B&B RKHS).This algorithm employs two existing criterion functions and one new criterion function however, all computed in RKHS. This enable B&B RKHS to conceive inherent nonlinear data structures. The algorithm was experimentally compared the popular wrapper approach that use an exhaustive to guarantee optimality. The classification accuracy achieved with both method was comparable however , runtime of B&B RKHS was superior Using two existing criterion function and even completely out of reach using the new criterion function.[5]

In this paper introduce an an improved global feature selection scheme (IGFSS) where the last step in a common feature selection scheme is modified in order to obtain a more representative feature set is proposed. Although feature set constructed by a common feature selection scheme successfully represents some of the classes, a number of classes may not be even represented. Consequently, IGFSS aims to improve the classification performance of global feature selection methods by creating a feature set representing all classes almost equally. For this purpose, a local feature selection method is used in IGFSS to label features according to their discriminative power on classes and these labels are used while producing the feature sets.[6]

This paper proposed a novel system for feature selection ,which integrates the 1-nearest neighbor technique determined by the LOOCV method and the binary improved GSA. The main aim this work is to optimize the feature subset selection effectively. The key to the success of the improved GSA is to utilize PWL for increas-ing its diversity of species, and to use SQP for accelerating local exploitation.[7]

VII CONCLUSION

Text mining is process to extract information from the unstructured data. A major difficulty of text classification is high dimensionality of feature space. Feature selection method used for dimension reduction. Feature selection is a process that chooses a subset from the original feature set according to some criterions. There are three general classes of feature selection algorithms: filter methods, wrapper methods and embedded methods .We use filter method for feature selection that use MFD algorithm .In MFD algorithm it is it necessary to assign value of F so it's work for only given features F we overcome this by proposed method in which F value select automatically by pre analyses data set extract unique feature and create feature set F so its work for all unique features and we will reduce maximum redundancy by imposing threshold value.

VIII REFERENCES

1. Vishal Gupta and Gurpreet S .Lehal,'A survey of text mining techniques and applications' ,Vol.1, Issue 1, Journal of emerging technology in web intelligence ,August 2009.

2. K.Nalini and Dr . L . Jaba Sheela ' Surevy on text classification ', Vol.1, Issue 6, International Journal of Computer Applications , July 2014.
3. Roberto H.W. Pinheiro, George D.C. Cavalcanti and Tsang Ing Ren,' Data-driven global-ranking local feature selection methods for text Categorization', October 2014
4. Abdur Rehman, Kashif Javed ,Haroon A. Babri and Mehreen Saeed,' Relative discrimination criterion – A novel feature ranking method for text data' December 2014
5. MatthiasRing and BjoernM.Eskofier ,'Optimal feature selection for non linear data using branch and bound in kernel space', August 2015
6. Alper Kursat Uysal,' An improved global feature selection scheme for text classification' 2015
7. Jie Xiangb, XiaoHong Hana, Fu Duanb, Yan Qiangb, XiaoYan Xionga, Yuan Lana, and Haishui Chaib ,'A novel hybrid syatem for feature selection based on an improved gravitational search algorithm and k NN method'2015
8. Nidhi and Vishal Gupta,'Recent trends in text classification techniques', Vol.35, Issue 6, International Journal of Computer Applications ,December 2011.
9. Anuradha and Patra, Divakar Singh,'A Surevy report on text classification with different term weighing methods and comparison between classification algorithem ', Vol.75, Issue 7, International Journal of Computer Applications , August 2013.
10. S.Niharika , V.sneha Latha and D. R . Lavanya 'A survey on categorization', Vol.3, Issue 1, International Journal of Computer Applications 2012.
11. Meenakshi and Swati Singla'Review paper on text categorization technique' , SSRG International Journal of Computer science and engineering(SSRG-IJCSE)-EFES , April 2015.
12. Vandana Korde and C Namrata Mahender ,'Text classification and classifiers : survey', Vol.3, Issue 2, International Journal of Computer Applications 2012. International Journal of Artificial Intelligence & Applications, March 2012.
13. Upendra Singh and Saqib Hasan ,' Survey Paper on Document Classification and Classifiers', Vol.3 Issue 2, International Journal of Computer Science Trends and Technology, Mar-Apr 2015.
14. Inoshika Dilrukshi and Kasun de Zoysa. 'A feature selection method of machine learning and computing', Vol.4, Issue 4, International Journal of Machine Learning and Computing, August 2014.
15. Basant Agarwal and Namita Mittal, 'Text classification using machine learning method- A survey',2014, Springer India 2014

BIOGRAPHY

Vala Mital is a PG student of Computer Engineering Department in B.H. Gardi college of engineering & Technology, Rajkot, Gujrat, India.

Prof. Jay Gandhi is an Assistant Professor Information Technology Engineering Department in B.H. Gardi college of engineering & Technology, Rajkot, Gujrat, India.