Performance of Various Feature Selection Techniques for Text Clustering

Navanith Rayavarapu
G. Sri Krishna Karthik

Abstract—**Feature selection techniques are frequently used in supervised machine learning models. But in unsupervised techniques they are seldom used because of unavailability of class labels. Feature selection methods can be used to identify and remove unneeded, irrelevant and redundant attributes from data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model. Feature selection is different from dimensionality reduction. Both methods seek to reduce the number of attributes in the dataset, but a dimensionality reduction method do so by creating new combinations of attributes, where as feature selection methods include and exclude attributes present in the data without changing them. Therefore this paper aims to conduct performance on various feature selection techniques, both supervised and unsupervised based on text clustering.**

Index Terms- **Feature Selection, K-means clustering, Information Gain, chi-squared, Random Forest Importance, Texture**

I. Introduction

CLUSTERING is the task of organizing unlabelled objects in a way that objects in the same group are similar to each other and dissimilar to those in other groups. In other words, clustering is like unsupervised classification where the algorithm models the similarities instead of the boundaries [1].

Text clustering is one of the central problems in text mining and information retrieval area. The task of text clustering is to group similar documents together. In general, there are two common algorithms. The first one is the hierarchical based algorithm, which includes single link, complete linkage, group average and Ward's method. By aggregating or dividing, documents can be clustered into hierarchical structure, which is suitable for browsing. However, such an algorithm usually suffers from efficiency problems. The other algorithm is developed using the K-means algorithm and its variants. Generally hierarchical algorithms produce more in-depth information for detailed analyses, while algorithms based around variants of the K-means algorithm are more efficient and provide sufficient information for most purposes. It had been applied to several applications, including improving retrieval efficiency of information retrieval systems, organizing the results returned by a search engine in response to user's query, browsing large document collections etc. In addition, the automated information retrieval for many document collections helped in reading, understanding, indexing and tracking a large amount of data. For this cause, researchers in fields of document retrieval, computational linguistics, and textual data mining are working hard on development new methods to process these data [2].

Text is represented in the form of document-term matrix where documents represents rows and terms represents columns. This results in large number of attributes. This representation suffers from two major challenges the problem of feature selection, and the problem of high dimensionality. In the set-of-words model, every word in the document can be selected as a feature, and the dimension of the feature space is equivalent to the number of different words in all of the documents.

This is where the feature selection techniques come into picture. It leads to the reduction of dimensionality of the original data set. The selection term set should contain enough or more reliable information about the original data set. In general, feature selection methods are usually divided into three categories: embedded, wrapper, and filter methods. [3]. Embedded methods learn which features best contribute to the accuracy of the model while the model is being created. The most common type of embedded feature selection methods are regularization methods. Regularization methods are also called penalization methods that introduce additional constraints into the optimization of a predictive algorithm (such as a regression algorithm) that bias the model toward lower complexity (fewer coefficients). Examples of regularization algorithms are the LASSO, Elastic Net and Ridge Regression. Wrapper methods consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. A predictive model us used to evaluate a combination of features and assign a score based on model accuracy. The search process may be methodical such as a best-first search, it may stochastic such as a random hill-climbing algorithm, or it may use heuristics, like forward and backward passes to add and remove features. An example if a wrapper method is the recursive feature elimination algorithm. Filter feature selection methods apply a statistical measure to assign a scoring to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset. The methods are often univariate and consider the feature independently, or with regard to the dependent variable. Some examples of some filter methods include the Chi squared test, information gain and correlation coefficient scores [4].

Since both embedded and wrapper based feature selection methods interact with the classifier, they can only select the optimal subset for a particular classifier. So the features selected by them may be worse for other classifiers. Moreover, another disadvantage of the two methods is that they are more time consuming than filter method. Therefore, filter method is more fit for dealing with data that has large amounts of features since it has a good generalization ability [5].

In this paper we have chosen k-means clustering algorithm for text clustering. Accuracy of the performance measure of this clustering algorithm with all features is compared with accuracy of different feature selection techniques (filter-based) like information gain, chi-squared, random-forest-importance, Correlation-based, Document Frequency.

II. Feature Selection Methods

In this section, we give a brief introduction on feature selection methods IG, CHI, Random Forest importance and then how we implemented exhaustive search and forward search on these techniques. In the following, D represents document set, M the dimension of the features, c the class, and N the number of documents in the dataset.

A. Information Gain:

Information gain of a term measures the number of bits of information obtained for category prediction by the presence or absence of the term in a document.Features that perfectly partition should give maximal information. Unrelated features should give no information. It measures the reduction in entropy.

This technique allows adjusting the boundaries of clusters. It is shown that the information gain ratio (IGR) grows monotonically and simultaneously with degree of connectivity between two variables. This approach has some preferences if compared, for example, with correlation analysis due to relatively smaller sensitivity to shape of functional dependencies.

Let m be the number of classes. The Information gain of a term is defined as

$$\begin{aligned} IG(t) = & -\Sigma_{i=1}^{m} p(c_i) log p(c_i) + \\ & p(t)\Sigma_{i=1}^{m} p(c_i|t) log p(c_i|t) + \\ & p(\bar{t})\Sigma_{i=1}^{m} p(c_i|\bar{t}) log p(c_i|\bar{t}) \end{aligned} \quad (1)$$

B. CHI square statistic:

A chi square (χ^2) statistic is used to investigate whether distributions of categorical variables differ from one another. Basically categorical variable yield data in the categories and numerical variables yield data in numerical form. The chi-square test is a statistical test of independence to determine the dependency between term and the category. It shares similarities with coefficient of determination, R^2. However, chi-square test is only applicable to categorical or nominal data while R^2 is only applicable to numeric data.

From the definition, of chi-square we can easily deduce the application of chi-square technique in feature selection. Suppose you have a target variable (i.e., the class label) and some other features (feature variables) that describes each sample of the data. Now, we calculate chi-square statistics between every feature variable and the target variable and observe the existence of a relationship between the variables and the target. If the target variable is independent of the feature variable, we can discard that feature variable. If they are dependent, the feature variable is very important.

$$\chi^2(t,c) = \frac{N * (p(t,c) * p(\bar{t},\bar{c}) - p(t,\bar{c}) * p(\bar{t},c))^2}{p(t) * p(\bar{t}) * p(c) * p(\bar{c})} \quad (2)$$

$$\chi^2(t) = avg_{i=1}^{m}\{\chi^2(t,c_i)\} \quad (3)$$

C. Random Forest Importance:

Random Forest is a supervised learning algorithm and are often used for feature selection in a data science workflow. The reason is because the tree-based strategies used by random forests naturally ranks by how well they improve the purity of the node. This mean decrease in impurity over all trees (called gini impurity). Nodes with the greatest decrease in impurity happen at the start of the trees, while nodes with the least decrease in impurity occur at the end of trees. Thus, by pruning trees below a particular node, we can create a subset of the most important features.

When constructing a tree, the RF method searches for only a random subset of input variables at each splitting node and the tree grows fully without pruning. The RF method is recognized as a specific instance of bagging. Random selection of variables at each node decreases the correlation among trees in a forest, thus forest error rate decreases. The random subspace selection method has been demonstrated to perform better than bagging when there are many redundant variables contributing to discrimination among classes. The computational load of the RF method is comparatively light. The computation time is on the order of ntree$\sqrt{mtry}$ nlogn, where ntree is the number of trees, mtry is the number of variables used in each split, and n is the number of training samples [6].

D. Correlation-based Feature Selection(CFS)

Correlation-based feature selection (CFS) ranks attributes according to a heuristic evaluation function based on correlations. The function evaluates subsets made of attribute vectors, which are highly correlated with the class but have low inter-correlation. Relevance of group of features grows with the correlation between features and class, and decreases with growing inter-correlation. The CFS method assumes that irrelevant features show a low correlation with the class and therefore should be ignored by the algorithm. CFS is used to determine the best feature subset and is usually combined with search strategies such as forward selection, backward elimination, bi-directional search, best first search and genetic search [7]. On the other hand, excess features should be examined, as they are usually strongly correlated with one or more of the other attributes. The criterion used to assess a subset of l features can be expressed as follows:

$$M_s = \frac{l * \bar{t_{cf}}}{\sqrt{l + l(l-1)\bar{t_{ff}}}} \quad (4)$$

where M_S is the evaluation of a subset of S consisting of l features, $\bar{t_c}$ is the average correlation value between features and class labels, and is the average correlation value between two features.

E. Document Frequency:

A term's document frequency is the number of documents in which the term occurs in the whole collection. DF thresholding is computing the document frequency for each unique term in the training corpus and then removing the terms whose document frequency are less than some predetermined threshold. That is to say, only the terms that occur many times are retained. DF thresholding is the simplest technique for vocabulary reduction. It can easily scale to very large corpora with a computational complexity approximately linear in the number of training documents. At the same time, DF is based on a basic assumption that rare terms are noninformative for category prediction. So it is usually considered an empirical approach to improve efficiency. Obviously, the above assumption contradicts a principle of information retrieval (IR), where the terms with less document frequency are the most informative ones [8] [9].

III. Experiment

We first conducted an ideal case experiment to demonstrate how clustering works for text mining without any use of feature selection techniques. The dataset which is subset of RCV1 which is already used in author identification experiments. In the top 4 authors (with respect to total size of articles) were selected. 4 authors of text documents labeled with atleast one subtopic of the class CCAT(Corporate/Industrial) were selected. This way it is attempted to minimize the topic factor in distinguishing among the text documents. The training corpus consists of 200 text documents (50 per author) and it includes other 200 text documents non-overlapping with the training text documents. Each document is represented as a vector with each term as dimension. Since the preprocessing is required for any text mining experiment, the data has undergone some techniques like removing numbers,sparse terms, stop words and punctuation.

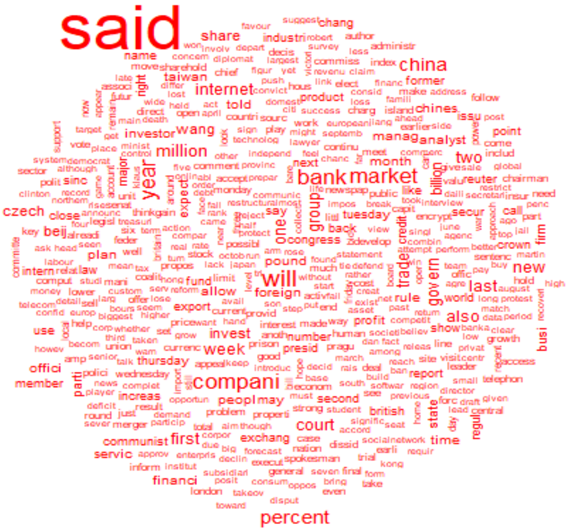


Fig. 1. Train Data Word Cloud

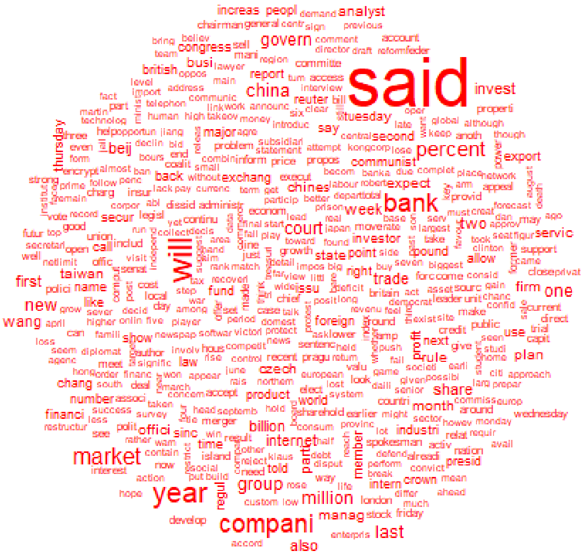


Fig. 2. Test Data Word Cloud

Frequency count of the top 30 terms is plotted below

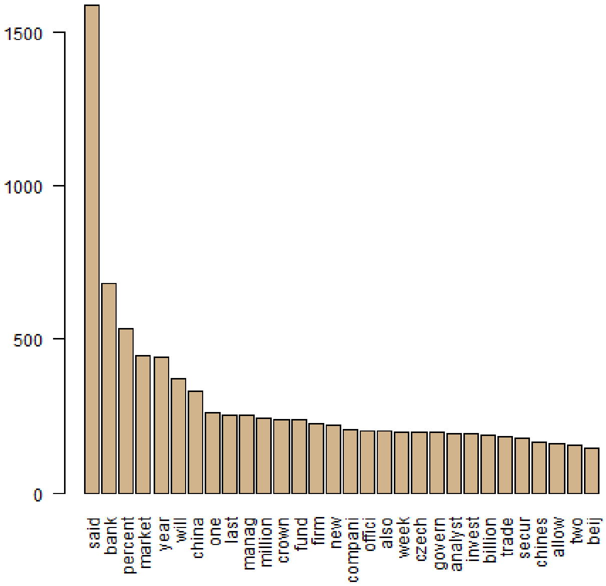


Fig. 3. Frequency for Test Data

These terms are converted into document-term matrix with frequency of each term. Thus the obtained data frame is taken as input for this experiment which consists of 904 features.

K-means clustering was chosen as our basic clustering algorithm. The data is trained using k-means clustering with 4 different clusters representing authors. The means obtained in the training phase are used as centroids for test data.Thus the test data is clustered. The performance metric accuracy is calculated with the available class label.

Since all the features are not so important for clustering, features selection techniques are applied to reduce the number of features and to reduce time complexity with more documents clustered correctly. The number of features taken are 904 for which these techniques are applied.

Information Gain, chi-squared statistic, Random Forest importance and Document Frequency separates the data into train and test sets. Each technique generates weights for each and every term according to their algorithms as described above. Here comes a challenge to select how many features should be present in subset for each algorithm. The accuracy for the test data in each of these techniques is calculated with number of features and plotted below. The peak values obtained in these plots respectively are the number of features to be selected. Thus the features are subsetted in each of these techniques and thereby maximizing the accuracy. This method of improving accuracy will work for any type of dataset. Correlation-based feature selection is a technique which evaluates subsets made of attribute vectors, which are highly correlated with the class but have low inter-correlation. Thus it doesn't generate weights for each term. It returns a subset of features from its algorithm.

A. Information Gain

This technique gave weights for all features. The cutoff for the number of features to be selected is estimated in the below graph against accuracy obtained with test data. From this estimation, feature subset with length 16 is selected for further processing with best accuracy.

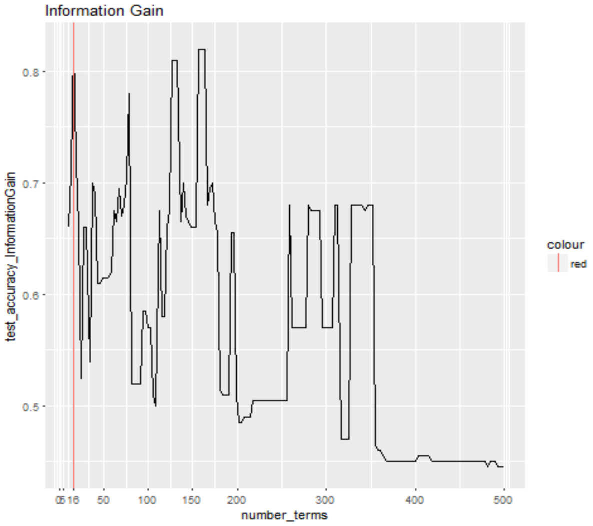


Fig. 4. Information Gain

B. Chi-Squared

The cutoff for the number of features to be selected is estimated in the below graph against accuracy obtained with test data. From this estimation, feature subset with length 20 is selected for further processing with best accuracy.

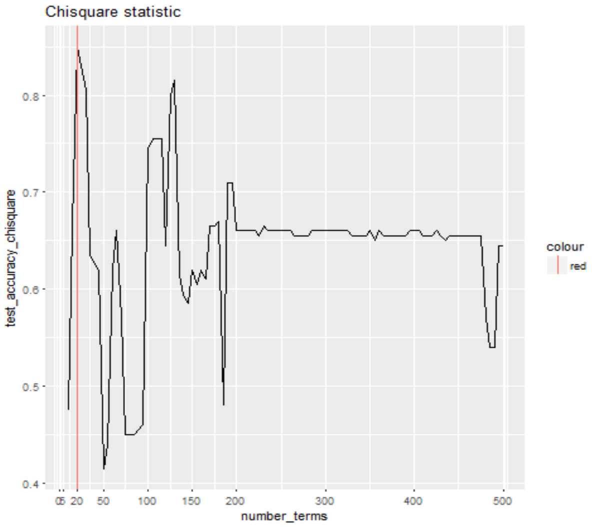


Fig. 5. Chisquare

C. Random Forest Importance

The cutoff for the number of features to be selected is estimated in the below graph against accuracy obtained with test data. From this estimation, feature subset with length 21 is selected for further processed with best accuracy.

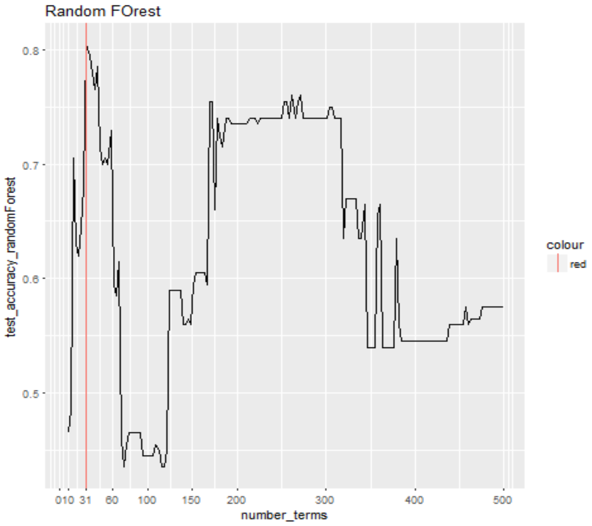


Fig. 6. Random Forest Importance

D. CFS

Correlation-based feature selection is a technique which evaluates subsets made of attribute vectors, which are highly correlated with the class but have low inter-correlation. Thus it doesn't generate weights for each term. It returns a subset of features from its algorithm. So there is no cutoff for this algorithm. The accuracy for test data is calculated with the returned subset of features. The accuracy obtained is 0.585.

E. Document Frequency

The cutoff for the number of features to be selected is estimated in the below graph against accuracy obtained with test data. From this estimation, feature subset with length 253 is selected for further processed with best accuracy.

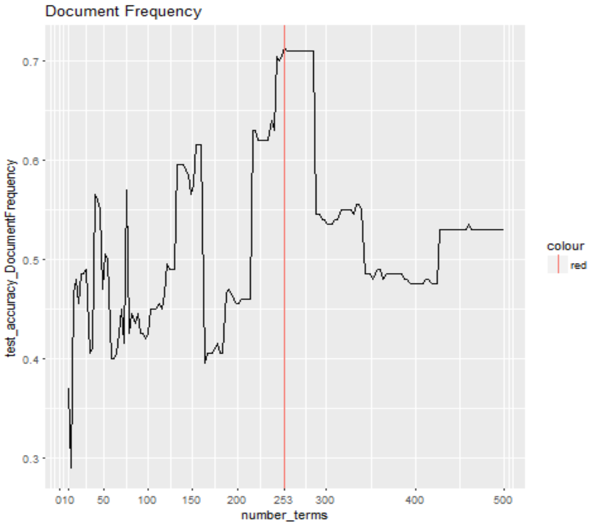


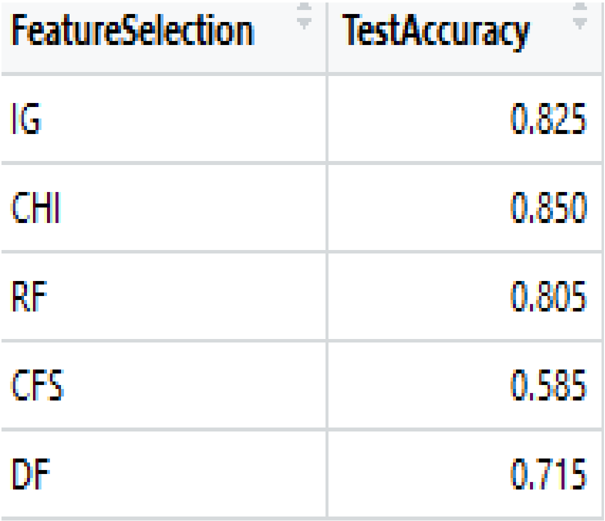
Fig. 7. Document Frequency

IV. Result and Analysis

First, we conducted an ideal case experiment to see whether good terms can help text clustering. That is, we applied supervised feature selection methods to choose the best terms based on the class label information. Then, we executed the text clustering task on these selected terms and compared the clustering results with the baseline system, which clustered the documents on full feature space.

From this analysis the accuracy of test data over each and every feature selection technique is calculated and shown below.

The accuracy of the test data is calculated by the ratio of correctly clustered documents in test data to the total number documents in test data.



FeatureSelection	TestAccuracy
IG	0.825
CHI	0.850
RF	0.805
CFS	0.585
DF	0.715

Fig. 8. Results

$$Accuracy = \frac{No.\,of\,documents\,clustered\,correctly\,in\,test\,data}{Total\,No.\,documents\,in\,test\,data} \tag{5}$$

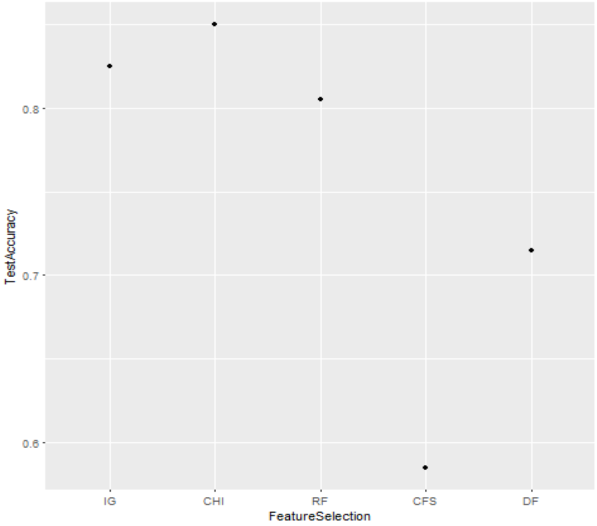


Fig. 9. Various Techniques

V. Conclusion

So from these results we can infer that applying feature selection techniques on the data, will increase algorithm's accuracy and efficiency. Among the techniques experimented chi-squared statistic gave best accuracy(85%). Thus we can infer that chi-squared statistic is the best option in feature selection.

References

[1] T. Svadas and J. Jha, "Document cluster mining on text documents," *International Journal of Computer Science and Mobile Computing, ISSN*, pp. 778–782, 2015.

[2] O. El Barbary and A. Salama, "Feature selection for document classification based on topology," *Egyptian Informatics Journal*, 2018.

[3] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[4] B. Harish and M. Revanasiddappa, "A comprehensive survey on various feature selection methods to categorize text documents," *International Journal of Computer Applications*, vol. 164, no. 8, pp. 1–7, 2017.

[5] H. Almuallim and T. G. Dietterich, "Learning with many irrelevant features.," in *AAAI*, vol. 91, pp. 547–552, Citeseer, 1991.

[6] H. Kawakubo and H. Yoshida, "Rapid feature selection based on random forests for high-dimensional data," *Expert Syst Appl*, vol. 40, pp. 6241–6252, 2012.

[7] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *Science and Information Conference (SAI), 2014*, pp. 372–378, IEEE, 2014.

[8] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[9] F. Song, S. Liu, and J. Yang, "A comparative study on text representation schemes in text categorization," *Pattern analysis and applications*, vol. 8, no. 1-2, pp. 199–209, 2005.