ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 1, January 2015

Study on Feature Selection Methods for Text Mining

Divya P[1], G.S. Nanda Kumar[2]

M.E, Department of CSE, Kumaraguru College of Technology, Coimbatore, India[1]

Associate Professor, Department of CSE, Kumaraguru College of Technology, Coimbatore, India[2]

Abstract: Text mining has been employed in a wide range of applications such as text summarisation, text categorization, named entity extraction, and opinion and sentimental analysis. Text classification is the task of assigning predefined categories to free-text documents. That is, it is a supervised learning technique. While in text clustering (sometimes called document clustering) the possible categories are unknown and need to be identified by grouping the texts. Clustering of documents is used to group documents into relevant topics. Each of such group is known as clusters. It is an unsupervised learning technique. The major difficulty in document clustering is its high dimension. It requires efficient algorithms which can solve this high dimensional clustering. The high dimensionality of data is a great challenge for effective text categorization. Each document in a document corpus contains much irrelevant and noisy information which eventually reduces the efficiency of text categorization. Most text categorization techniques reduce this large number of features by eliminating stopwords, or stemming. This is effective to a certain extent but the remaining number of features is still huge. It is important to use feature selection methods to handle the high dimensionality of data for effective text categorization. Feature selection in text classification focuses on identifying relevant information without affecting the accuracy of the classifier. This paper gives a literature survey on the feature selection methods. The survey mainly emphasizes on major feature selection approaches for text classification and clustering.

Keywords: Text mining, Text classification, Text clustering, and Feature selection.

I. INTRODUCTION

Text mining has been employed in a wide range of applications such as text summarisation, text categorization, named entity extraction, and opinion and sentimental analysis. Text mining requires a great deal of pre-processing in which the text must be decomposed into smaller syntactical units (e.g., terms and phrases). Sometimes, the text data may also need to be transformed into other types. For example, in some text mining applications, terms extracted from the documents in the entire corpus are treated as features and documents are treated as records. Thus, each document can be represented as a Boolean vector in which a true (or false) value for a feature indicates the presence (or absence) of the corresponding term in the document. During mining stage, depending on the requirements of the specific applications, various data mining methods such as association mining, classification and clustering are frequently used in text categorization applications.

Generally, text-mining refers to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text classification is the task of assigning predefined categories to free-text documents. That is, it is a supervised learning technique. While in text clustering the possible categories are unknown and need to be identified by grouping the texts. Text clustering is also called document clustering. Clustering of documents is used to group documents into relevant topics. Each of such group is known as clusters. It is an unsupervised learning technique. The major difficulty in document clustering is its high dimension. It requires efficient algorithms which can solve this high dimensional clustering. There are several algorithms for text clustering which includes agglomerative clustering algorithm, partitioning clustering algorithm, Hierarchical clustering algorithm, Density-based clustering algorithm, Grid-based clustering algorithm, Model-based clustering algorithm, Frequent pattern-based clustering, and Constraint-based clustering.

The high dimensionality of data is a great challenge for effective text categorization. Each document in a document corpus contains much noisy and irrelevant information which may reduces the efficiency of text categorization. Most text categorization techniques reduce

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 1, January 2015

this large number of features by eliminating stemming or stopwords. It is effective to a certain extent but the remaining number of features is still huge. It is important to use feature selection methods to handle the high dimensionality of data for effective text categorization. Feature selection in text classification focuses on identifying relevant information without affecting the accuracy of the classifier. More sophisticated feature selection techniques have been reported in the literature [14, 15, 16].

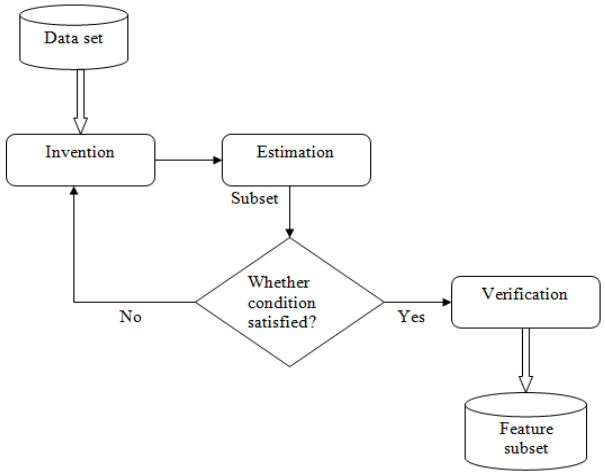

Fig. 1.1 Steps in feature selection

Since filters are independent from the classifier, it is inexpensive to use. Examples of filters include weight by correlation, chi square, information gain [14], mutual information [14], Gini index [14] and many more. Wrappers [14], by comparison, evaluate features by training a classifier; this means that the classifier's accuracy is calculated over several subsets of features determined by greedy algorithms. Wrappers yield better results but they are expensive and may suffer from overfitting.

Feature reduction transforms the original set of features into new features by applying some transformation function. This new feature set contains far fewer features or dimensions than the original set. Common feature reduction techniques include latent semantic indexing, term clustering [15], and principal component analysis [16]. They yield good results over a reduced dimension feature space. But they are expensive to use.

TABLE I
FEATURE SELECTION CATEGORIES

	Single feature evaluation	**Subset selection**
Filter	Mutual information Chi square statistic Entropy Information gain Gini index	Category distance
Wrapper	Ranking accuracy using single feature	For LR (SFO, Grafting)

This paper is organized as follows. Section I has introduced this work. Section II, by comparison, explains the necessary background and related work. Section III, on the other hand, explains the literature review. Finally, Section IV, draws the conclusions of this survey work.

II. Related Work

There are several text classification and clustering methods. Most text categorization techniques reduce this large number of features by eliminating stopwords, or stemming. This is effective to a certain extent but the remaining number of features is still huge. It is important to use feature selection methods to handle the high dimensionality of data for effective text categorization. Feature selection in text classification focuses on identifying relevant information without affecting the accuracy of the classifier. There are many feature selection methods. Mainly they are classified as filter and wrapper feature selection methods. One of the feature selection method can be choose to identify relevant information from documents based on their content. Selection of feature selection is an important and challenging step in text mining.

III. Literature Review

The Drawbacks of the traditional clustering algorithms are mentioned [6]. The algorithms generally favor clusters with spherical shapes and similar sizes, or are very fragile in the presence of outliers. Sudipto Guha, Rajeev Rastoji, Kyuseok Shim proposed a new algorithm CURE [6].

CURE is a hierarchical clustering algorithm, which employs the features of both the centroid based algorithms and the all point algorithms. CURE improves upon BIRCH by ability to discover clusters of arbitrary shapes. CURE is also more robust with respect to outliers. These benefits are

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 1, January 2015

achieved by using several representative objects for a cluster. CURE [6] obtains a data sample from the given database. The algorithm divides the data sample into groups and identifies some representative points from each group of the data sample. In the first phase, the algorithm considers a set of widely spaced points from the given datasets. In the next phase of the algorithm the selected dispersed points are moved towards the centre of the cluster by a specified value of a factor α. As a result of this process, some randomly shaped clusters are obtained from the datasets. In the process it identifies and eliminates outliers. CURE is scalable to large datasets with a complexity of O(N), since CURE requires one scan of the dataset.

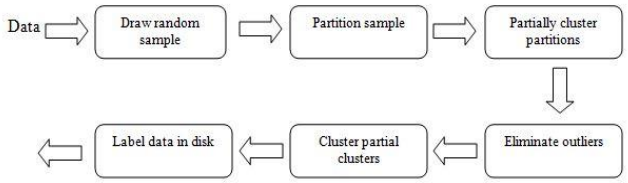

Fig. 1.2 Overview of CURE

In the next phase of the algorithm, the representative points of the clusters are checked for proximity with a threshold value and the clusters that are next to each other are grouped together to form the next set of clusters. In this hierarchical algorithm, the value of the factor α may vary between 0 and 1. The utilization of the shrinking factor alpha by the CURE overcomes the limitations of the centroid based, all-points approaches. As the representative points are moved through the clustering space, the ill effects of outliers are reduced by a greater extent. Thus the feasibility of CURE is enhanced by the shrinking factor α. The worst case time complexity of CURE is determined to be $O(n^2 \log n)$. A drawback is the user-specified parameter values, the number of clusters and the shrinking factor. A random sample of data objects is drawn from the given datasets. Partial clusters are obtained by partitioning the sample dataset and outliers are identified and removed in this stage. Final refined clusters are formed from the partial cluster set.

Advantages:

- More robust to outliers.
- High quality of clusters produced.
- Time complexity is $O(n^2 \log n)$

Disadvantage:

- Errors occur if clusters are not well separated.

RObust Clustering using links algorithm (ROCK) [7] is targeted to both Boolean data and categorical data. It uses the Jaccard coefficient as the measure similarity. The input is a set S of n sampled points to be clustered (that are drawn randomly from the original data set), and the number of desired clusters k. It samples the data set in the same manner as CURE. This agglomerative algorithm assumes a similarity measure between objects and defines a 'link' between two objects whose similarity exceeds a threshold.

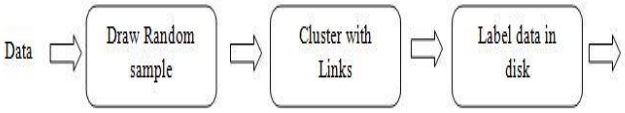

Fig. 1.3 ROCK Algorithm

Initially, each object is assigned to a separate cluster. Then, clusters are merged repeatedly according to their closeness: the sum of the number of 'links' between all pairs of objects between two clusters. ROCK has cubic complexity in N, and is unsuitable for large datasets. A pair of items are said to be neighbors if their similarity exceeds some threshold. The number of links between two items is defined as the number of common neighbors they have. The objective of the clustering algorithm is to group together points that have more links. A recent development in genetics used this algorithm called GE-ROCK[7]. GE-ROCK is an improved ROCK algorithm which combines the techniques of clustering and genetic optimization. In GE-ROCK, similarity function is used throughout the iterative clustering process, while in the "conventional" ROCK algorithm similarity function is only to be used for the initial calculation. The analysis showed that GE-ROCK leads to the superior performance in not only better clustering quality but also shorter computing time when compared to the ROCK algorithm commonly used in the literature.

Advantages:

- The computation of links is more efficient.
- Performs well on real categorical data, and respectably on time series data.

Disadvantage:

- The Jaccard coefficient fails to capture the natural clustering of well separated data sets with categorical attributes.

Many clustering techniques have been applied to clustering documents. Documents can be easily browsed when they are clustered. But it has two problems. First one is

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 1, January 2015

cluster is too slow for large corpora. Second one is cluster does not appreciably improve retrieval. Scatter/Gather [8] is a document browsing method which uses document clustering as its primitive operation. To implement this, fast document clustering is necessary.

Scatter/Gather Session:

- User is presented with short summaries of a small number of document groups.
- User selects one or more groups for further study.
- Continue this process until the individual document level.

Cutting et al. adapted various partition-based clustering algorithms to clustering documents. Two of the techniques are Buckshot and Fractionation. Buckshot selects a small sample of documents to pre-cluster them using a standard clustering algorithm and assigns the rest of the documents to the clusters formed. Fractionation splits the N documents into m buckets where each bucket contains N/m documents. Fractionation takes an input parameter ρ, which indicates the reduction factor for each bucket. The standard clustering algorithm is applied so that if there are n documents in each bucket, they are clustered into n/ρ clusters. Now each of these clusters is treated as if they were individual documents and the whole process is repeated until there are only K clusters.

Advantage:

- Clustering can be done quickly by working in a local manner on small groups of documents rather than trying to deal with the entire corpus globally.

Disadvantage:

- Accuracy of the Buckshot and Fractionation algorithms is affected by the quality of the clustering provided by the slow cluster subroutine.

The main idea of topic-driven clustering [12] is to organize a document collection according to a given set of topics. The traditional classification algorithms cannot solve the topic-driven clustering problem because of the insufficient information about each class (topic). Ying Zhao [12] proposes an effective and efficient of topic-driven clustering method that emphasizes the relationship between documents and topics and relationship among documents themselves simultaneously. Topic-driven clustering algorithm is based on the idea that it is possible to use explicitly available domain knowledge to constrain or guide the clustering process. It takes the advantages of both Supervised and Unsupervised components.

Advantage:

- It is efficient and effective approach.

Disadvantages:

- It is time consuming
- Costly

Online approach for clustering massive text and categorical data streams. C. C. Aggarwal and P. S. Yu [10] proposed a framework, which is provided in order to execute the clustering process continuously in online fashion, and it is also important to provide end users with the ability to analyze the clusters in an offline fashion. For achieving greater accuracy in the clustering process, it is necessary to maintain a high level of granularity in the underlying data structures. In order to achieve this goal, use a process in which condensed clusters of data points are maintained. Such groups are called cluster droplets. The cluster droplet maintenance algorithm starts with an empty set of clusters. Then unit clusters are created. Once k clusters have been created, begin the process of online cluster maintenance. This framework can be used for both text and categorical data streams and it provides higher cluster purity.

Advantages:

- Cluster purity is high.
- Algorithm is effective in quickly adapting to temporal variations in the data stream.

Disadvantage:

- Cluster purity may reduce for large number of classes running in parallel.

C. C. Aggarwal, S. C. Gates, and P. S. Yu [11] proposed a text categorization method on using partial supervision. Clustering is used to create categories and it can be used for document classification. The documents are clustered based on supervised clustering algorithm then the clustered documents are categorized based on categorization algorithm. The supervised clustering algorithm is based on the seed-based technique. This algorithm starts with a set of seeds. Some of the words are projected out in order to represent the seeds. In each iteration, the number of words gradually reduced thus clusters get more refined. Smaller number of words is required to isolate the subjects of the documents in that cluster. The following four steps are applied iteratively to the final set of clusters.

- Document Assignment

In the first step, assign the documents to their closest seed.

- Project

In this step, project out the words with least weight. This ensures that only the terms which are frequently occurring

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 1, January 2015

within a cluster of documents are used for the assignment process.

- Merge

After getting the words with least weight merge all clusters using a simple single linkage method.

- Kill

In last step discard all the seeds from S such that the number of documents is fewer than the predefined number. These documents either get redistributed to other clusters or get classified as outliers in later iterations.

The categorization algorithm works as follows.

- Find the k closest cluster seeds to the test documents.
- The similarity of each cluster to the test document is calculated by using the cosine measure.
- The k categories are the candidates for the best match and may contain a set of closely related subjects. This ranking process is designed to re-rank these categories more appropriately.

Advantages:

- Much more easy to classify documents.
- Higher quality of categorization.

Disadvantage:

- Hard to provide a direct comparison of proposed technique to any other classification method by using a measure such as classification accuracy.

Angelova and Siersdorfer [12] proposed an approach to linked document clustering by means of iterative relaxation of cluster assignments on a linked graph. Relaxation labeling is effective in Web page classification [12]. Based on the framework for modeling link distribution through link statistics, a combined logistic classifier is used based on content and link information. This approach not only showed improvement over a textual classifier, but also outperformed a single flat classifier based on both content and link features. An alternative to this approach is not all neighbors are considered. Instead, only neighbors that are similar enough in content are used.

In this approach the statistical knowledge about the cluster assignments of the nodes in the formed neighborhoods in a graph G. Furthermore, the assignment of the edge weights, and thus the type of graphs used by the above approaches, are based on node-node similarity. To improve the robustness of the algorithm here proposes two beneficial extensions. Aim of this approach is to ignore the unnecessary and most probably noisy information behind all irrelevant links in a neighborhood by assigning to each edge *e* a weight here equal to the cosine similarity between the feature vectors of the documents connected by the edge. Also explore further the hypothesis that neighboring documents should receive similar cluster assignments. Here introduce a metric over the set of clusters where thematically close clusters are separated by a shorter distance and thus it reduces the cost for assigning neighbors to similar clusters. The performance of the graph-based clustering is better if the content combination technique is used as initialization step for the graph-based methods.

Advantages:

- It improves accuracy.
- Since it ignore the unnecessary and most probably noisy information behind all irrelevant links in a neighborhood this approach provides more robustness.

Disadvantage:

- Running time is high.

There are several effective feature selection methods, including Information Gain (IG), chi square statistic (CHI), Document Frequency (DF), Term Strength (TS), Term Contribution (TC) etc.. IG and CHI are two supervised methods, and DF, TS, En and TC are four unsupervised methods. All these methods assign a score to each individual feature and then select features which are greater than a pre-defined threshold. Then Tao Liu [14] proposes a new feature selection method called Term Contribution (TC). It can perform a comparative study on a variety of feature selection methods for text clustering.

Two methods are used. They are:

- Supervised feature selection methods
- Unsupervised feature selection methods

1. Supervised feature selection method

- **Information Gain (IG)**

Information gain [14] 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. Let m be the number of classes. The information gain of a term t is defined as

$$IG(t) = -\sum_{i=1}^{m} p(c_i)\log p(c_i) + p(t)\sum_{i=1}^{m} p(c_i \mid t)\log p(c_i \mid t) + p(\bar{t})\sum_{i=1}^{m} p(c_i \mid \bar{t})\log p(c_i \mid \bar{t})$$

- **χ^2 statistic (CHI)**

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 1, January 2015

The χ^2 statistic (CHI) [14] measures the association between the term and the category. It is defined to be

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

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

2. Unsupervised feature selection method

- **Document Frequency (DF)**

Document frequency [14] is the number of documents in which a term occurs in a dataset. It is the simplest criterion for term selection and it can easily scales to a large dataset with linear computational complexity. It is simple but effective feature selection method for text categorization.

- **Term Strength (TS)**

Term strength [14] is originally proposed and evaluated for vocabulary reduction in text retrieval. It is computed based on the conditional probability that a term occurs in the second half of a pair of related documents given that it occurs in the first half:

$TS(t) = p(t \in dj \mid t \in d_i), d_i, dj \in D \cap sim(d_i, dj) > \beta$

where β is the parameter to determine the related pairs. Since need to calculate the similarity for each document pair, the time complexity of TS is quadratic to the number of documents. Because the class label information is not required, this method is also suitable for term reduction in text clustering.

- **Term Contribution (TC)**

Term Contribution [14] takes the term weight into account. Because the simple method like DF assumes that each term is of same importance in different documents, it is easily biased by those common terms which have high document frequency but uniform distribution over different classes. TC is proposed to deal with this problem; the result of text clustering is highly dependent on the documents similarity. So the contribution of a term can be viewed as its contribution to the documents' similarity. The similarity between documents *di* and *dj* is computed by dot product:

$$sim(d_i, d_j) = \sum_t f(t,d_i) \times f(t,d_j)$$

where $f(t,d)$ represents the tf*idf weight of term t in document d. So define the contribution of a term in a dataset as its overall contribution to the documents' similarities. The equation is

$$TC(t) = \sum_{i,j \cap i \neq j} f(t,d_i) \times f(t,d_j)$$

- **Entropy Based Ranking (EN)**

The term is measured by the entropy reduction when it is removed. The entropy [14] is defined as the equation

$$E(t) = -\sum_{i=1}^{N}\sum_{j=1}^{N}(S_{i,j} \times \log(S_{i,j}) + (1-S_{i,j}) \times \log(1-S_{i,j})),$$

where $Si\,j$, is the similarity value between the document *di* and *dj*. $Si\,j$, is defined as the equation

$$S_{i,j} = e^{-\alpha \times dist_{i,j}}, \alpha = -\frac{\ln(0.5)}{\overline{dist}}$$

where $dist\ i\ j$, is the distance between the document *di* and *dj* after the term *t* is removed, *dist* is the average distance among the documents after the term *t* is removed. The most serious problem of this method is its high computation complexity $O(MN^2)$. It is impractical when there is a large number of documents and terms, and therefore, sampling technique is used in real experiments. The summary is

- Supervised Feature Selection
  - IG and CHI feature selection methods are performed.
  - Feature selection makes little progress on datasets of Reuters and 20NG.
  - Achieves much improvement on Web directory dataset.
- Unsupervised Feature Selection
  - DF, TS, TC and En feature selection methods are performed.
  - While 90% of terms removed, entropy is reduced by 2% and precision is increased by 1%.
  - When more terms are removed, the performance of unsupervised methods is dropped quickly; however, the performance of supervised methods is still improved.

H. H. Hsu, C. W. Hsieh [15] proposed feature selection method via correlation coefficient clustering. There are several measures which are helpful in finding the

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 1, January 2015

redundant features: Mutual information, correlation coefficient, and chi square are used to find the dependency between two features. Correlation coefficient of two random variables is a quantity that measures the mutual dependency of the two variables. When two features are mutually dependent, it means that the occurrences and variation of the two features must be almost the same.

Correlation coefficient clustering for feature selection:

- To find related feature groups is not easy.
- The pair wise similarity measurements of the whole feature set are hard to be realized due to a large amount of huge calculations.
- In statistics, the correlation coefficient indicates the strength and direction of a relationship between of two random variables.
- Two variables have strong dependency when their correlation coefficient value is close to 1 or -1.
- If the value is zero then two variables are not related at all.

The key characteristic of the proposed method is to apply clustering analysis in grouping the collected features. Only one representative feature is needed from each feature group. This can greatly reduce the total number of features.

H. Liu and L. Yu. [16] integrated feature selection algorithms for classification and clustering. Feature selection is one of the important and frequently used techniques for data preprocessing for data mining. It reduces the number of features, removes irrelevant, redundant or noisy data and brings the immediate effects for applications: speeding up a data mining algorithm, improving mining performance such as predictive accuracy and result comprehensibility. There are three types of feature selection approaches. They are filter and wrapper approaches.

Filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm. The wrapper model requires one predetermined mining algorithm and uses its performance as the evaluation criterion. It searches for features better suited to the mining algorithm aiming to improve mining performance, but it also tends to be more computationally expensive than the filter model.

General procedure of feature selection:

- Subset generation
  - This is a process of heuristic search with each state in the search space specifying a candidate subset for evaluation.
- Subset evaluation
  - Each newly generated subset needs to be evaluated by an evaluation criterion. Two criteria are independent criterion and dependency criterion. Typically, an independent criterion is used in algorithms of the filter model. It tries to evaluate the goodness of a feature or feature subset by exploiting the intrinsic characteristics of the training data without involving any mining algorithm. Some popular independent criteria are distance measures, information measures, dependency measures, and consistency measures. A dependency criterion used in the wrapper model requires a predetermined mining algorithm in feature selection and uses the performance of the mining algorithm applied on the selected subset to determine which features are selected. It usually gives superior performance as it finds features better suited to the predetermined mining algorithm, but it also tends to be more computationally expensive, and may not be suitable for other mining algorithms.
- Stopping criteria
  - It determines when the feature selection process should stop.
- Result validation
  - Directly measure the result using prior knowledge about the data.

Some frequently used stopping criteria are:

(a) The search completes.

(b) Some given bound is reached, where a bound can be a specified number (minimum number of features or maximum number of iterations).

(c) Subsequent addition (or deletion) of any feature does not produce a better subset.

(d) A sufficiently good subset is selected (e.g., a subset may be sufficiently good if its classification error rate is less than the allowable error rate for a given task).

TABLE II
FILTER FEATURE SELECTION METHODS

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 1, January 2015

IV. CONCLUSION

Text document contains a large number of features.

Feature selection methods	Advantages	Disadvantages
Gini index	• Select features efficiently. • Measures the features' ability to discriminate between classes. • Widely used in building Classification Trees and determining more important splits.	• Select large number of features.
Pearson's Correlation Coefficient	• Both Supervised and unsupervised. • Works in univariate setting. • Very simple to interpret and implement.	• Works only on numeric attributes. • Can detect linear relationship.
Correlation-based Feature Selection (CFS)	• Supervised, Multivariat • Works with all type of data. • Simplicity of the theory. • Select fewer features with higher accuracy. • Quickly identify irrelevant, redundant features and noise.	• To obtain the optimal feature set, have to perform a search in the feature subspace which may not be required.
Mutual Information	• Supervised • Works with all type of data.	• It is said to be biased towards features with more value.
Symmetric uncertainty (SU)	• Supervised • Works with all type of data. • Eliminate biased nature. • Select fewer features with higher accuracy.	• Cannot apply to large datasets.
Crammer's V	• Supervised • Works with all type of data.	• There is a criticism when this is applied on high dimensional datasets. • Works only in supervised setting.

Therefore appropriate and accurate feature selection techniques are generally essential to the performance of text classification systems. In a text document, each word can be a possible feature. These large numbers of possible features can typically be reduced through a variety of techniques. Eliminating stopwords, or stemming, feature selection, and clustering are methods for reducing the dimensionality of a test classification problem. Most text categorization techniques reduce this large number of features by eliminating stopwords, or stemming. This is effective to a certain extent but the remaining number of features is still huge. Feature selection determines the most relevant features, or words, in a document. Clustering is a method by which related features are considered together as a single feature. This paper has provided a survey on text classification and clustering methods and feature selection methods. It is possible to combine these methods to further reduce dimensionality, but one must be careful to avoid removing too many possible features as this can adversely affect performance. It usually works well and takes less time when combining filter feature selection method with single feature evaluation. While combining wrapper feature selection method with subset selection the accuracy is very high but consumes time. Thus filter feature selection methods are more powerful than wrapper feature selection methods.

REFERENCES

[1] Han, J. and Kamber, M., *Data Mining: Concepts and Techniques*, 2nd ed., Elsevier, Morgan Kaufmann, 2006.
[2] C. C. Aggarwal and C. X. Zhai, *A survey of text classification algorithms in Mining Text Data*, New York, NY, USA: Springer, 2012.
[3] Charu C. Aggarwal, Yuchen Zhao, and Philip S. Yu, *On the Use of Side Information for Mining Text Data*, IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 6, June 2014.
[4] C. C. Aggarwal and P. S. Yu, '*On text clustering with side information*,' in Proc. IEEE ICDE Conf., Washington, DC, USA,2012.
[5] I. Dhillon, *Co-clustering documents and words using bipartite spectral graph partitioning*, in Proc. ACM KDD Conf., New York, NY, USA, pp. 269–274, 2001.
[6] S. Guha, R. Rastogi, and K. Shim, "*CURE: An efficient clustering algorithm for large databases"*, in Proc. ACM SIGMOD Conf., New York, NY, USA, 1998, pp. 73–84.
[7] S. Guha, R. Rastogi, and K. Shim, *ROCK*: *A robust clustering algorithm for categorical attributes*, Information Systems., vol. 25, no. 5, pp. 345–366, 2000.
[8] D. Cutting, D. Karger, J. Pedersen, and J. Tukey, "*Scatter/Gather: A cluster-based approach to browsing large document collections"*, in Proc. ACM SIGIR Conf., New York, NY, USA, pp. 318–329, 1992.
[9] H. Schutze and C. Silverstein, "*Projections for efficient document clustering"*, in Proc. ACM SIGIR Conf., New York, NY, USA, 1997, pp. 74–81.

ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 2, Issue 1, January 2015

[10] C. C. Aggarwal and P. S. Yu, *A framework for clustering massive text and categorical data streams*, in Proc. SIAM Conf. Data Mining, pp. 477–481, 2006.
[11] C. C. Aggarwal, S. C. Gates, and P. S. Yu, *On using partial supervision for text categorization*, IEEE Trans. Knowl. Data Eng., vol. 16, no. 2, pp. 245–255, Feb. 2004.
[12] Y. Zhao and G. Karypis, "*Topic-Driven Clustering for Document Datasets,"* Proc. SIAM Int'l Conf. Data Mining, pp. 358-369, 2005.
[13] Angelova, R., Siersdorfer, S., *A neighborhood based approach for clustering of linked document collections*, In Proc. of the 15th ACM CIKM, pp. 778–779, 2006.
[14] T Liu, S Liu, Z Chen, WY Ma., *An Evaluation on Feature Selection for Text Clustering* In ICML, <aaai.org, 2003>.
[15] H. H. Hsu, C. W. Hsieh, *Feature Selection via Correlation Coefficient Clustering*, Journal of Software, vol. 5, no. 12, pp. 1371-1377, 2010.
[16] H. Liu and L. Yu., *Toward integrating feature selection algorithms for classification and clustering*, Knowledge and Data Engineering, IEEE Transactions on, 17(4):49-502, April 2005.

BIOGRAPHY

Ms. Divya P, have completed her B.Tech (Computer Science and Engineering) in the year of 2011 from, Calicut University, Kerala, India. Currently pursuing Master of Engineering in Computer Science and Engineering at Kumaraguru College of Technology, Coimbatore, Tamil Nadu, India. Her interests: Data mining and Database.

Mr. G.S. Nanda Kumar, currently working as Associate Professor in the department of Computer Science and Engineering at Kumaraguru College of Technology, Coimbatore, Tamil Nadu, India and has a teaching experience of 16 years. He is currently pursuing PhD under Anna University. His research interests include Knowledge Management, Data Mining and Software Engineering.