UNSUPERVISED FEATURE SELECTION FOR TEXT CLUSTERING USING DIFFERENTIAL INVERSE DOCUMENT FREQUENCY

Sivaram Prasad Nalluri

Research Scholar, Computer Science and Engineering,
Acharya Nagarjuna University, AP 522510, India
becithod@gmail.com

Rajasekhara Rao Kurra

Professor, Computer Science & Engineering, Usha Rama College of Engg. & Tech.,
Telaprolu, AP 521109, India
krr_it@yahoo.co.in

Abstract

Text clustering is gaining importance among researchers because of rapid increase in the availability of online text collections without class labels. It helps to organize, summarize and retrieve useful information from corpora. High dimensionality of text datasets leads to poor performance of clustering algorithms. Dimensionality can be reduced using feature extraction or feature selection methods. Feature selection methods scale well and are easy to interpret. An unsupervised univariate filter feature selection method was proposed for dimensionality reduction. The proposed method outperformed nine other filter methods reported in the literature, by identifying most relevant features that lead to good clustering performance on eight popular text datasets.

***Keywords*: Feature Selection; Unsupervised; Filter Method; Text Clustering; Differential Inverse Document Frequency.**

1. Introduction

Due to tremendous growth in the usage of internet [Warf (2021)], the volume of online text without class labels; in the form web documents, judicial records, medical health records, operation manuals, research articles, news articles and textbooks increase [Zhai and Massung (2016)]. Text clustering is essential to organize, summarize and to retrieve relevant information automatically from the online text corpora [Garcia-Dias *et al.* (2020)]. Curse of dimensionality is the major challenge faced by Text clustering [Li *et al.* (2017)]. Due to the presence of irrelevant features the performance of clustering algorithm is poor and consumes more computational resources [Wang *et al.* (2021)]. Hence, dimensionality reduction techniques help in identifying significant features. As the class labels are not available in clustering, unsupervised dimensionality reduction techniques should be used. Feature extraction and Feature selection are the two popular unsupervised dimensionality reduction methods [Ray *et al.* (2021)]. Feature extraction techniques map the given feature space to a virtual feature space, where the virtual features are difficult to interpret by human beings [Guyon *et al.* (2008)]. Feature selection methods selects a subset of quality features from the original feature set. Unsupervised Feature selection methods are classified into filters, wrappers and embedded methods [Solorio-Fernández *et al.* (2020a)]. Filter methods assign score to each feature, independent of the subsequent learning algorithm that makes use of the features. Wrappers assign score to features based on the feature's contribution to the performance of the subsequent application that makes use of the features. Embedded methods simultaneously select features while building the learning model. Filter methods scale well for large datasets when compared with Wrapper methods [Solorio-Fernández *et al.* (2020b)].

The relevant work done in the field of filter methods for unsupervised feature selection to efficiently cluster text documents is reported in Sec. II. The authors have proposed an unsupervised filter feature selection method which is discussed in Sec. III. Eight popular text datasets with different dimensionality, number of clusters and cluster distribution as described in Sec. IV, are considered to compare filter methods. The proposed filter method outperformed all other methods considered, as discussed in Sec. V, by identifying most relevant features for effective text clustering.

DOI : 10.21817/indjcse/2021/v12i4/2112040014

2. Related Work

Unsupervised univariate filter feature selection methods, which are proposed for efficient text clustering, are discussed in this section. Length Feature Weight (LFW) is proposed by [Agarwal *et al.* (2020)] to efficiently cluster web services description documents. Infinite Feature Selection (INFS) is proposed by [Roffo *et al.* (2020)]. A graph is built from a dataset in which each node represents a feature, and an edge represents a relationship between features. The significance of a feature is given by the value of the paths that contain the feature. Relevance of a term for document clustering is evaluated in terms of Detailed Document Frequency (DDF) of a term by [Abualigah (2019)]. DDF of a term depends on the number of times a term occurs in each document, average number of times a term occurs per each document, document length, document frequency a term and maximum term frequency in a document. The number of documents in which the term occurs at least once is called its document frequency. The total number of key terms in a document is called its length. Key terms are those terms which are not common words in English language. Semantic Association of Term (SAT) is proposed by [Nalluri and Kurra (2015a)]. A term is semantically related to other term if they co-occur in documents. The magnitude of the semantic relationship between a pair of terms, at document level, depends on the product of both the term frequencies in a document. This document level semantic relationship between a pair of terms, when aggregated across all documents gives collection level relationship between the pair of terms. The sum of the semantic relationship of a term with all other terms in a collection is called semantic association of a term. The larger the semantic association of a term is the greater the discriminative power it has in distinguishing documents that belong to different clusters. The Term Frequency Term Document Frequency (TFTDF) is proposed by [Nalluri and Kurra (2015b)]. The significance of a term is proportional to the product of document level score of a term aggregated across all documents and collection level score of a term. The document level score of a term is the ratio of number of times the term occurs in a document to length of the document. The collection level score of a term is the ratio of total number key terms in the collection to the document frequency of a term. Collection Frequency Inverse Document Frequency (CFIDF) is proposed by [Nalluri and Kurra (2014)]. The score assigned to a term is the product of the collection frequency of a term and logarithm of document frequency of a term to the base ten. The score is then discounted by the document frequency of a term. The number of times a term occurs in all the documents in a collection is called the collection frequency of a term. Laplacian score (LS) [He *et al.* (2005)] of a term reflects the ability of the term to preserve the graph constructed over the Laplacian matrix. A term that mostly preserves the locality of objects in the dataset is given highest score and vice versa. Term Variance (TV) is proposed by [Liu *et al.* (2005)], in which a term's score is proportional to variance of the frequency distribution of the term across documents in the collection. A term with highest variance in frequency distribution across documents is considered to have most discriminative power to distinguish documents that belong to different clusters. Term Contribution (TC) is proposed by [Liu *et al.* (2003)]. A term is ranked based on its contribution to similarity between all possible pairs of documents in a collection. A terms contribution to similarity between a pair of documents is calculated as the product of the weights of the term in the document pair.

3. Differential Inverse Document Frequency (DIDF)

The significance of a term in a text dataset depends on its document frequency. If its document frequency is high, then it has least significance because it is not useful in differentiating one document from the other. If a term occurs in very few documents, then it can be used effectively to differentiate these documents from others. Hence, a term's significance should decrease with the increase of its document frequency as in Eq. (1)

$$\mathrm{W_g(t, D)} = \log\left(\left(\frac{\mathrm{n}}{\mathrm{df(t, D)}}\right)\text{-}1\right). \qquad (1)$$

Where, D represents the text document collection, $\mathrm{W_g(t, D)}$ is the global weight of a term in D, n is the number of documents in the collection and $\mathrm{df(t, D)}$ is the document frequency of term t in the collection D, which is defined as the number of documents in the collection that contains the term t.

According to Eq. (1), if a term occurs in say 50% of n, then $\mathrm{df(t, D)} = (1/2) \cdot \mathrm{n}$ and the weight assigned to the term becomes zero. If $\mathrm{df(t, D)} = (3/4) \cdot \mathrm{n}$ then weight of the term is $\log(1/3)$ which is -0.477. Thus, Eq. (1) penalizes terms that occur in fifty or more percentage of n. The document level weight of a term $\mathrm{W_{l1}(t, d)}$, based on document's length $\mathrm{len(d)}$, is given by Eq. (2),

$$\mathrm{W_{l1}(t, d)} = \frac{\mathrm{tf(t,d)}}{\mathrm{len(d)}}. \qquad (2)$$

The cumulative value of the local weight $\mathrm{W_{l1}(t, d)}$ for all the documents in the corpus is given by Eq. (3),

$$\mathrm{CV_{lw1}(t, D)} = \sum_{i=1}^{n} \mathrm{W_{l1}^{2}(t, d_i)}. \qquad (3)$$

The document level weight of a term, based on the number of occurrences of the most frequent term in the document $\mathrm{W_{l2}(t, d)}$, can be calculated using Eq. (4),

$$\mathrm{W_{l2}(t, d)} = \frac{\mathrm{tf(t, d)}}{\mathrm{mtf(d)}}. \qquad (4)$$

Where $\mathrm{tf(t, d)}$ is the frequency of term t in document d and $\mathrm{mtf(d)}$ is the number of occurrences of the most frequent term in document d. The cumulative value of the local weight $\mathrm{W_{l2}(t, d)}$ for all the documents in the corpus is given by Eq. (5),

$$\mathrm{CV_{lw2}(t, D)} = \sum_{i=1}^{n} \mathrm{W_{l2}^{2}(t, d_i)}. \qquad (5)$$

The combined cumulative value of local weights of a term is calculated as in Eq. (6),

$$\mathrm{CV_{lws}(t, D) = CV_{lw1}(t, D) + CV_{lw2}(t, D)}. \qquad (6)$$

The overall weight of a term considering combined cumulative local weights given by Eq. (6) and global weight given by Eq. (1) is calculated as in Eq. (7),

$$\mathrm{W(t) = W_g(t, D)}\sqrt{\mathrm{CV_{lws}(t, D)}}. \qquad (7)$$

The document frequency distribution of terms varies with document collection. Majority of terms in a collection have document frequencies that fall in the limits $(0,\ 50\%]$ of n. However, some collections have considerable number of terms with document frequencies that fall in the limits $(50\%,\ 100\%]$ of n. Hence, using fixed threshold $(50\%\ of\ n)$ for document frequency to penalize the terms as shown in Eq. (1) is not an optimal feature selection technique for all document collections. An adaptive threshold to penalize terms based on their document frequency is proposed by the authors and is given by Eq. (8),

$$\mathrm{th_a(df)} = \frac{2\sum_{\mathrm{df}=1}^{n}\left(\mathrm{df \cdot bc(df)}\right)}{\mathrm{m \cdot n}} \cdot 100. \qquad (8)$$

Where $\mathrm{th_a(df)}$ is the adaptive threshold for document frequency of a term measured as percentage of n, $\mathrm{bc(df)}$ is the bin count for the document frequency value “df” in the histogram for document frequency distribution of terms in the corpus and m is the number of terms in the term vocabulary of the corpus. A term occurs at least in one document or at most in all documents in the collection. So, the limits for df are [1, n].

The global weight of a term as given by Eq. (1) is modified making use of adaptive threshold given by Eq. (8) as in Eq. (9),

$$\mathrm{W_g^{'}(t, D)} = \log\left(\left(\frac{\mathrm{n}}{\mathrm{df(t, D)}}\right) - \frac{1}{\mathrm{th_a(df)}} + 1\right). \qquad (9)$$

If $\mathrm{th_a(df)}$ is say 75% of n, then calculation of $\mathrm{W_g^{'}(t, D)}$ as per Eq. (9) is as follows.

$$\mathrm{W_g^{'}(t, D)} = \log\left(\left(\frac{\mathrm{n}}{\mathrm{df(t, D)}}\right) - \frac{1}{0.75} + 1\right).$$

According to Eq. (9) if a term's document frequency satisfies the equation $\mathrm{df(t, D) = th_a(df) \cdot n}$ then its global weight $\mathrm{W_g^{'}(t, D)}$ is zero. Any term whose document frequency satisfies the inequality $\mathrm{df(t, D) > th_a(df) \cdot n}$ has negative global weight. The overall weight of a term as given by Eq. (7) is modified, making use of adaptive global weight given by Eq. (9), as in Eq. (10),

$$\mathrm{W^{'}(t) = W_g^{'}(t, D)} \cdot \sqrt{\mathrm{CV_{lws}(t, D)}}. \qquad (10)$$

4. Research Method

Experiments are conducted on eight text document collections to compare the 10 filter methods. Vector Space Model (VSM) [Liu *et al.* (2020)] representation, in the form of Document Term Matrix (DTM) is used. Each element $w_{i\,j}$ of DTM represents weight assigned to j-th term in i-th document. No special term weighting measures are used. Hence, $w_{i\,j}$ is the count of j-th term in i-th document. Terms that do not occur in any of the documents are removed. Let "utc" be the collection of unique terms count in documents. If a document's unique terms count is less than 1 percentile of "utc” it is excluded from the text corpus. Also, if a document's total terms count is less than 110% of its unique terms count, it is removed from the documents collection. K-Means clustering algorithm is used to cluster documents represented with only top p % of terms chosen by filter methods. K-Means clustering algorithm is more suitable for datasets with large variation in class distribution [Zhou and Yang (2020)].

Feature selection methods are evaluated based on the clustering performance obtained using the top p % of terms. The clustering performance achieved using 100 % of features in a dataset is taken as base line performance AF. To assess the impact of number of features (p %) used for document representation, on clustering performance, p is varied from 1 % to 25 % in steps of 0.5 %. If the number of clusters obtained differs from actual number of classes in a dataset (k), the clustering performance is shown using symbol "X".

To assess the quality of clusters obtained, normalized Van Donzen (NVD) [Wu *et al.* (2009)] and BCubed F measure (BCF) [Bagga and Baldwin (1998)] metrics are used. BCF is the harmonic mean of BCubed precision and BCubed recall which according to [Amigó *et al.* (2009)] satisfies all the constraints that a good clustering evaluation metric should have. The range of BCF is [0, 1]. If BCF value is 1, it means the clustering solution perfectly matches with class distribution of the dataset. According to [Wu *et al.* (2009)] NVD is the best metric to evaluate the clustering solution of K-means algorithm. The range of NVD metric falls within the limits [0, 1]. Lower the NVD value means better the clustering solution.

4.1. *Datasets*

The properties of eight text corpora, considered for evaluating feature selection methods, are given Table 1. Multi label documents are excluded from datasets. The coefficient of variation (CV) is a measure of variation in class distribution around mean, for each dataset. The CV value is used to compare the dispersion of class distributions of different datasets. The adjusted Fisher-Pearson standardized moment coefficient [David (2018)], which is a measure of coefficient of skewness (CSK) is given for each dataset, to compare the skewness in class distributions of different datasets. All classes of Ohsumed dataset have exactly same number of documents (400), so its coefficient of skewness is not defined (ND). Datasets D1 and D3 are downloaded from the following URL http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html. Rest of the corpora are downloaded from the following URL http://sites.labic.icmc.usp.br/text_collections/.

Dataset	Acronym	Number of			Sparsity	CV	CSK
		documents	features	categories			
Reuters	D1	8213	18933	41	99.75%	3.23	4.81
Ohsumed	D2	9200	13510	23	99.59%	0	ND
20Newsgroups	D3	18846	26214	20	99.66%	0.1	-2.39
Ohscal	D4	11162	11465	10	99.47%	0.27	0.24
Sports	D5	8580	14870	7	99.14%	1.02	1.01
Webkb	D6	8282	22890	7	99.61%	1.06	1.74
Hitech	D7	2301	12940	6	98.90%	0.5	-0.62
Reviews	D8	4069	22925	5	99.20%	0.64	-0.43

Table 1. Important properties of datasets.

5. Results and Discussion

The efficacy of the filter methods in identifying most relevant features for better clustering solution is judged using three criteria.

5.1. *Criterion 1*

Ability to achieve clustering performance that is either better than or equal to base line performance (AF) using minimum percentage of features, chosen as per scores assigned by feature selection method. The clustering performance achieved using 100% of features in a dataset is considered as the base line performance (AF).

The clustering performance of feature selection methods using NVD metric is shown in Table 2. For a given dataset, the clustering performance obtained by a feature selection method using least number of features, that is either equal to or better than the base line performance (AF) is underlined. If there is a tie among feature selection methods, regarding the minimum percentage of features required to achieve at least AF, the value corresponding to the method with best clustering performance is shown in boldface. The proposed method DIDF, was able to achieve clustering performance better than base line performance for five datasets (D2, D3, D5, D6 and D7) with minimum percentage of features. No other feature selection method was able to perform best for at least two datasets.

Dataset	D1	D2	D3	D4	D5	D6	D7	D8
VAR	3.0 %	16.5 %	4.0 %	56.5 %	11.5 %	33.0 %	8.5 %	12.5 %
LAP	42.0 %	74.0 %	55.0 %	93.5 %	9.5 %	99.5 %	14.5 %	9.5 %
TC	3.5 %	12.0 %	4.0 %	39.0 %	52.0 %	17.5 %	5.5 %	2.5 %
INFS	4.5 %	19.5 %	3.5 %	56.5 %	6.5 %	44.0 %	3.5 %	2.5 %
DDF	2.5 %	12.0 %	4.5 %	43.0 %	8.0 %	36.5 %	9.5 %	**1.5 %**
LFW	3.5 %	28.5 %	5.0 %	37.0 %	21.0 %	13.5 %	9.5 %	2.5 %
DIDF	2.5 %	**11.5 %**	**2.0 %**	42.5 %	**2.0 %**	**3.0 %**	**1.5 %**	2.5 %
SAT	**1.0 %**	12.5 %	3.0 %	37.5 %	33.0 %	14.5 %	8.0 %	5.5 %
CFIDF	1.5 %	24.0 %	2.5 %	**34.5 %**	21.0 %	37.0 %	2.5 %	12.5 %
TFTDF	1.5 %	15.0 %	3.5 %	36.5 %	23.0 %	17.5 %	5.0 %	13.0 %

Table 2. Minimum percentage of features necessary to obtain at least base line clustering performance (AF) with NVD measure.

The performance of filter methods using BCF metric is shown in Table 3. DIDF method was able to achieve clustering performance better than base line performance for three datasets (D3, D5 and D7) with minimum percentage of features. No other feature selection method was able to perform best for at least two datasets without tie among filter methods.

Dataset	D1	D2	D3	D4	D5	D6	D7	D8
VAR	3.5 %	9.0 %	6.0 %	10.5 %	11.5 %	7.0 %	8.5 %	2.0 %
LAP	42.0 %	21.0 %	16.0 %	28.0 %	9.5 %	19.0 %	8.0 %	9.5 %
TC	3.5 %	7.0 %	6.5 %	10.5 %	14.0 %	4.0 %	5.5 %	**1.0 %**
INFS	4.5 %	4.5 %	8.5 %	10.5 %	18.0 %	25.0 %	3.5 %	2.0 %
DDF	4.0 %	8.5 %	4.5 %	**9.5 %**	7.0 %	5.0 %	7.0 %	1.5 %
LFW	3.5 %	8.5 %	7.0 %	12.5 %	13.5 %	**3.0 %**	4.0 %	2.5 %
DIDF	2.5 %	4.0 %	**2.0 %**	10.0 %	**2.0 %**	5.0 %	**1.5 %**	2.5 %
SAT	2.5 %	**2.0 %**	5.0 %	11.0 %	33.0 %	4.0 %	6.0 %	1.0 %
CFIDF	3.0 %	5.0 %	6.5 %	9.5 %	5.5 %	5.5 %	8.0 %	1.5 %
TFTDF	**2.0 %**	3.0 %	6.0 %	11.5 %	21.5 %	4.0 %	4.0 %	4.5 %

Table 3. Minimum percentage of features necessary to obtain at least base line clustering performance (AF) with BCF measure.

5.2. *Criterion 2*

Average clustering performance, relative to base line (AF), achieved using top p% of features, where p varies from 1% to 5% at an increment of 0.5%.

The mean percentage change in clustering performance relative to base line performance (AF), of feature selection methods using NVD metric is shown in Table 4. If the number of clusters obtained using top 5% of features ranked by a feature selection method is less than number of classes in a dataset, “X” is shown in the tabular cell at the intersection of the dataset and the filter method. Best mean clustering performance, for a dataset is shown in boldface. The proposed method, DIDF was able to achieve best mean clustering performance better than base line performance for five datasets (D1, D3, D5, D6 and D7). No other feature selection method was able to achieve this for at least one dataset.

Dataset	D1	D2	D3	D4	D5	D6	D7	D8
VAR	-0.03 %	-1.97 %	-0.93 %	-4.99 %	-19.39 %	X	-1.37 %	-3.28 %
LAP	X	X	X	X	X	X	X	X
TC	-1.94 %	-2.37 %	-1.88 %	-5.10 %	-28.60 %	-0.50 %	-1.58 %	-1.68 %
INFS	-1.40 %	-1.99 %	-1.66 %	-5.00 %	-18.21 %	X	-1.58 %	-2.24 %
DDF	0.21 %	-2.01 %	-0.97 %	-5.02 %	-21.21 %	-0.38 %	-0.74 %	-0.88 %
LFW	-2.16 %	-2.97 %	-3.24 %	-5.99 %	-30.62 %	-0.60 %	-2.32 %	-2.88 %
DIDF	**4.02 %**	-0.95 %	**9.61 %**	-4.54 %	**4.52 %**	**0.64 %**	**1.09 %**	-4.26 %
SAT	-0.71 %	-1.59 %	-1.76 %	-4.10 %	-21.95 %	-0.58 %	-0.90 %	-0.61 %
CFIDF	-1.45 %	-1.68 %	-1.11 %	-5.38 %	-19.42 %	-0.44 %	-0.84 %	-2.45 %
TFTDF	-0.25 %	-2.01 %	-1.78 %	-4.10 %	-20.53 %	-0.75 %	-0.63 %	-2.56 %

Table 4. Mean percentage change in clustering performance (NVD measure) relative to AF with top p % of features, p = 1 % to 5 % in steps of 0.5 %.

The mean percentage change in clustering performance, relative to base line performance (AF), of feature selection methods using BCF metric is shown in Table 5. The proposed method DIDF, was able to achieve best mean clustering performance better than base line performance for five datasets (D1, D2, D3, D5 and D7). Only one more method DDF was able to achieve best mean performance slightly better than base line performance for two datasets (D6 and D8).

Dataset	D1	D2	D3	D4	D5	D6	D7	D8
VAR	-0.57 %	-6.50 %	-7.50 %	-4.63 %	-4.11 %	X	-1.19 %	-1.90 %
LAP	X	X	X	X	X	X	X	X
TC	-3.35 %	-7.75 %	-10.43 %	-5.13 %	-11.96 %	0.20 %	-1.16 %	-0.45 %
INFS	-1.78 %	-5.15 %	-8.89 %	-4.67 %	-4.10 %	X	-1.14 %	-0.67 %
DDF	-0.17 %	-5.55 %	-6.89 %	-4.82 %	-6.68 %	**0.21 %**	-0.70 %	**0.05 %**
LFW	-5.96 %	-9.57 %	-15.35 %	-6.11 %	-13.20 %	0.16 %	-1.36 %	-1.28 %
DIDF	**8.26 %**	**1.61 %**	**22.49 %**	-2.98 %	**9.47 %**	-2.64 %	**0.01 %**	-2.84 %
SAT	-1.29 %	-1.87 %	-7.81 %	-3.68 %	-5.44 %	-0.11 %	-0.63 %	-0.08 %
CFIDF	-2.66 %	-4.31 %	-5.94 %	-5.00 %	-5.03 %	-0.12 %	-0.83 %	-1.44 %
TFTDF	-0.71 %	-4.89 %	-7.91 %	-3.47 %	-5.81 %	-0.28 %	-0.24 %	-1.60 %

Table 5. Mean percentage change in clustering performance (BCF measure) relative to AF with top p % of features, p = 1 % to 5 % in steps of 0.5 %.

5.3. *Criterion 3*

Change in clustering performance as the number of quality features varies from 1% to 25% of features in a dataset, with an increment of 0.5%.

Figure 1 shows the influence of the number of quality features on the clustering performance, measured using NVD, for different feature selection methods on dataset D3. Feature ratio (fr) is the ratio of number of features used for clustering to total number of features in a dataset. The range of NVD metric is [0, 1]. Lower the NVD value means better the clustering solution. For convenience of interpretation of the graph, 1-NVD value is taken along Y-axis. So, higher the Y-axis value (1-NVD) means better the clustering solution. The clustering performance achieved using 100% of features in a dataset is considered as the base line performance (AF).

Figure 2 shows the influence of the number of quality features on the clustering performance, measured using BCF, for different feature selection methods on the dataset D3. As shown in Figs. 1 and 2, none of the feature selection method can achieve base line performance (AF), with feature ratio below 2 %.

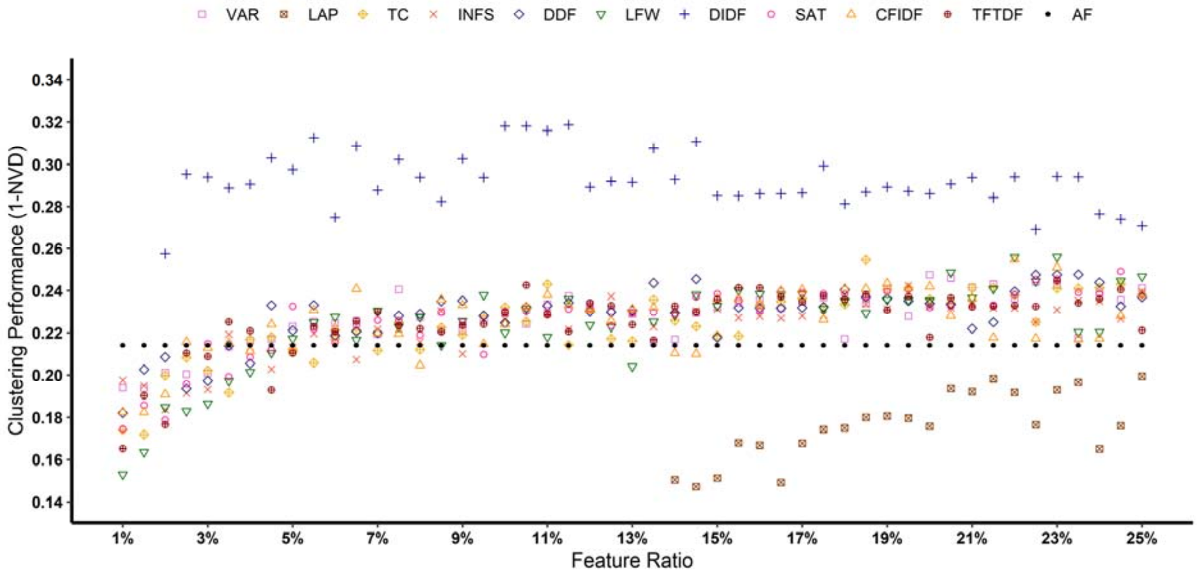

Fig. 1. Comparison of feature selection methods based on Clustering Performance measured using NVD on the dataset D3.

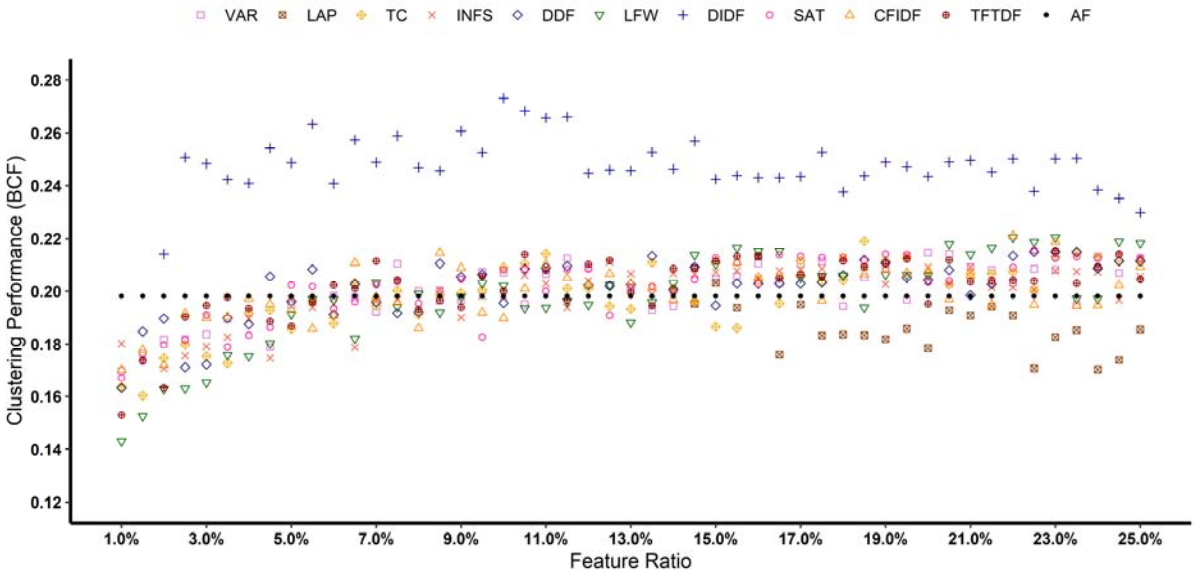

Fig. 2. Comparison of feature selection methods based on Clustering Performance measured using BCF on the dataset D3.

In both the test cases, starting from a feature ratio of 2 %, the proposed method DIDF was able to identify most discriminative features that resulted in a clustering performance better than those of other feature selection methods and base line performance. With the increase in the number of features, proportion of irrelevant features increase, which deteriorates clustering performance.

6. Conclusion

An unsupervised univariate filter feature selection method DIDF was proposed to reduce the ill effects of curse of dimensionality on text clustering performance. DIDF was compared with nine related feature selection methods reported in the literature. Eight datasets, with varied characteristics, were considered for testing the effectiveness of feature selection methods. DIDF was proved to be the most promising method in identifying relevant features for better clustering performance, for sparse datasets with varying degrees of skewness and dispersion in class distribution.

References

[1] Abualigah, L. M. Q. (2019). *Feature selection and enhanced krill herd algorithm for text document clustering,* Springer, Berlin.

[2] Amigó, E.; Gonzalo, J.; Artiles, J.; Verdejo, F. (2009): A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information retrieval, **12**(4), pp. 461–486.

[3] Agarwal, N.; Sikka, G.; Awasthi, L. K. (2020): Enhancing web service clustering using Length Feature Weight Method for service description document vector space representation. Expert Systems with Applications, **161**, pp. 113682.

[4] Bagga, A.; Baldwin, B. (1998): Entity-Based Cross-Document Co referencing Using the Vector Space Model. In 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, **1**, pp. 79–85.

[5] David, W. (2018): Basic Guidelines for Common Business Statistics Metrics. Business Education Innovation, **32**(2), pp. 21–26.

[6] Garcia-Dias, R.; Vieira, S.; Pinaya, W. H. L.; Mechelli, A. (2020). Clustering analysis, *In Machine Learning,* Academic Press, London.

[7] Guyon, I.; Gunn, S.; Nikravesh, M.; Zadeh, L. A. (2008). *Feature extraction: foundations and applications,* Springer, Berlin.

[8] He, X., Cai, D.; Niyogi, P. (2005): Laplacian score for feature selection. Advances in neural information processing systems, **18**, pp. 507–514.

[9] Li, J., Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R. P.; Tang, J.; Liu, H. (2017): Feature selection: A data perspective. ACM Computing Surveys, **50**(6), pp. 1–45.

[10] Liu, L.; Kang, J.; Yu, J.; Wang, Z. (2005): A comparative study on unsupervised feature selection methods for text clustering. IEEE International Conference on Natural Language Processing and Knowledge Engineering, pp. 597–601.

[11] Liu, Z.; Lin, Y.; Sun. M. (2020). Document Representation, *In Representation Learning for Natural Language Processing,* Springer, Singapore.

[12] Liu, T.; Liu, S.; Chen, Z.; Ma, W. Y. (2003): An evaluation on feature selection for text clustering. In Proceedings of the 20th international conference on machine learning (ICML-03), pp. 488–495.

[13] Nalluri, S. P.; Kurra, R. R. (2014): Subspace clustering of text documents using collection and document frequencies of terms. International Review on Computers and Software, **9**(10), pp. 1692–1699.

[14] Nalluri, S. P.; Kurra, R. R. (2015a): Effective Clustering of Text Documents in Low Dimension Space using Semantic Association among Terms. International Review on Computers and Software, **10**(5), pp. 467–474.

[15] Nalluri, S. P.; Kurra, R. R. (2015b): Feature Selection based on Term Frequency and Term Document Frequency for Text Clustering. International Journal of Applied Engineering Research, **10**(10), pp. 26175–26190.

[16] Roffo, G.; Castellani, U.; Vinciarelli, A.; Cristani, M. (2020): Infinite feature selection: a graph-based feature filtering approach. IEEE Transactions on Pattern Analysis and Machine Intelligence, [preprint].

[17] Ray, P.; Reddy, S. S.; Banerjee, T. (2021): Various dimension reduction techniques for high dimensional data analysis: a review. Artificial Intelligence Review, pp. 1–43.

[18] Solorio-Fernández, S.; Carrasco-Ochoa, J. A.; Martínez-Trinidad, J. F. (2020a): A review of unsupervised feature selection methods. Artificial Intelligence Review, **53**(2), pp. 907–948.

[19] Solorio-Fernández, S.; Carrasco-Ochoa, J. A.; Martínez-Trinidad, J. F. (2020b): A systematic evaluation of filter Unsupervised Feature Selection methods. Expert Systems with Applications, **162**, pp. 113745.

[20] Warf, B. (Ed.). (2021). *Geographies of the Internet,* Routledge, New York.

[21] Wu, J.; Xiong, H.; Chen, J. (2009): Adapting the right measures for k-means clustering. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 877–886.

[22] Wang, J.; Zhang, H.; Wang, J.; Pu, Y.; Pal, N. R. (2021): Feature selection using a neural network with group lasso regularization and controlled redundancy. IEEE Transactions on Neural Networks and Learning Systems, **32**(3), pp. 1110–1123.

[23] Zhai, C.; Massung, S. (2016). *Text data management and analysis: A practical introduction to information retrieval and text mining,* ACM; Morgan and Claypool, California.

[24] Zhou, K.; Yang, S. (2020): Effect of cluster size distribution on clustering: a comparative study of k-means and fuzzy c-means clustering. Pattern Analysis and Applications, **23**(1), pp. 455–466.

Authors Profile

Sivaram Prasad Nalluri, is a research scholar in the department of Computer Science and Engineering of Acharya Nagarjuna University, Andhra Pradesh, India. He is also working as a faculty member in the Information Technology department of Bapatla Engineering College, Bapatla, India. He obtained Post Graduate Degree in Computer Science and Engineering from Jawaharlal Nehru Technological University, Hyderabad (India) in 2002. His research interests include Machine Learning, Deep Learning and Cyber Security. He is affiliated with CSI and ISTE as life member.

Rajasekhara Rao Kurra is working as a professor in Computer Science and Engineering department of Usha Rama College of Engineering & Technology, Telaprolu, India. He obtained Ph.D in Computer Science & Engineering from Acharya Nagarjuna University, Guntur (India). His research interests include Machine Learning, Embedded Systems, Software Engineering and Knowledge Management. He is a Fellow of IETE. He is affiliated with IE, ISTE, ISCA, CSI as life member and with ACM as member. He is the Chairman of CSI Vijayawada Chapter, India.