Foundations of Science (2020) 25:1077–1094
https://doi.org/10.1007/s10699-019-09592-w

A Similarity Function for Feature Pattern Clustering and High Dimensional Text Document Classification

Vinay Kumar Kotte[1,2] · **Srinivasan Rajavelu**[3] · **Elijah Blessing Rajsingh**[3]

Published online: 9 March 2019

Abstract
Text document classification and clustering is an important learning task which fits to both data mining and machine learning areas. The learning task throws several challenges when it is required to process high dimensional text documents. Word distribution in text documents plays a very key role in learning process. Research related to high dimensional text document classification and clustering is usually limited to application of traditional distance functions and most of the research contributions in the existing literature did not consider the word distribution in documents. In this research, we propose a novel similarity function for feature pattern clustering and high dimensional text classification. The similarity function proposed is used to carry supervised learning based dimensionality reduction. The important feature of this work is that the word distribution before and after dimensionality reduction is the same. Experiment results prove the proposed approach achieves dimensionality reduction, retains the word distribution and obtained better classification accuracies compared to other measures.

Keywords Classification · Clustering · Dimensionality · Feature selection · Feature reduction

1 Introduction

Many applications have evolved in literature that proposed several dimensionality reduction techniques. Dimensionality reduction involves converting higher dimensional data to lower dimensional data. The objective behind dimensionality reduction is used to address both

✉ Vinay Kumar Kotte
kotte.vinaykumar@gmail.com; vinaykumar.kitswgl@gmail.com

Srinivasan Rajavelu
srini0402@gmail.com

Elijah Blessing Rajsingh
elijahblessing@karunya.edu

[1] Department of CSE, Karunya Institute of Technology and Sciences (Deemed to be university), Coimbatore, Tamilnadu, India

[2] Department of CSE, Kakatiya Institute of Technology and Science, Warangal, Telangana, India

[3] Karunya Institute of Technology and Sciences (Deemed to be university), Coimbatore, Tamilnadu, India

computation time and space. The traditional way to perform the dimensionality reduction is principal component analysis (PCA). PCA is computationally expensive for high dimensional data. However, there are many other traditional methods to convert the high dimensional data into data with lower dimensions. Machine learning approaches like clustering, classification are used in text mining related applications recently to convert the high dimensional data into smaller subsets to increase the computational efficiency (Bingham and Mannila 2001).

Usually text documents are comprised with irrelevant and noisy features which make learning algorithms fail to produce better accuracies. To remove the unwanted data, different data mining techniques can be applied. Feature selection and Feature extraction are the two different techniques used to classify the data (Abualigah et al. 2017). Applying techniques like text clustering for feature selection and classification of the text data is one of the mostly applied strategies recently. Feature selection is used to eliminate the unwanted text features so as to perform the text clustering and classification in efficient way. Early approach of research has focused more on converting high dimensional data to lower dimensional data by using existing distance functions. Dimensionality reduction reduces the computation time and increases the classification efficiency. Text retrieval and information retrieval are used in identifying the meaning and synonym from the document (Berka and Vajteršic 2013). Many approaches were proposed by various researchers in order to perform clustering and classification tasks. Clustering was performed based on unsupervised methods with various class label information (Bharti and Singh 2014).

Feature selection and feature extraction are the two different techniques that are widely used in dimensionality reduction. Feature selection is used to eliminate the unused features and identify the attributes in the original feature space (He et al. 2008), whereas feature extraction is used to map the data from high dimension to low dimension. The most widely used technique for performing dimensionality reduction is PCA. Feature selection is used to find the representative features in the original space (He et al. 2008). Latent semantic indexing (LSI) and latent semantic analysis (LSA) are two different components used to detect the high dimension data and perform the dimensionality reduction techniques. LSA is used in detecting the low rank optimization for the document term matrix. Detection of LSA is implemented using singular value decomposition which is also equivalent with PCA (He et al. 2008). Locally linear embedding algorithm is another method used to compute the high dimensional data with great efficiency.

LLE is used to protect the local configurations in order to protect the low dimensional space. Though the space complexity matters, the LLE is widely used to convert high dimensional data to lower dimensions (He et al. 2008). Most commonly used LLE algorithm uses Euclidean distance measure. The main focus in classification of data is towards the conversion of higher dimensions into lower dimensions by using either the PCA or Linear Discriminant analysis (LDA) which is based on supervised learning algorithms. LDA is linearly available only when there is a possibility of data/text present in the Gaussian distributed space.

Clustering is an un-supervised learning method or technique for placing similar entities at one place together. Clustering process is a challenging task from the need to overcome several challenges to achieve accuracy and efficiency. Accuracy is challenging because there is no rule of thumb that exists to help us decide the correct number of clusters. Also, the challenge for achieving efficiency and cluster quality brings the necessity and requirement for designing new and better similarity measures.

In the existing literature, several similarity functions are proposed which are used for computing similarity between two entities (Aggarwal 2007). But most of these similarity measures are suitable for computing similarity in low dimensional data space. There is an

immediate need for coming up with new and accurate similarity measures that are applicable for applications related to the high dimensional data space. Now the important point is

- What is the minimal high dimensional data space? (VinayKumar et al. 2015)
- Why similarity measures actually suitable for low dimensional data space turns unsuitable when moving for high dimensional data spaces (VinayKumar et al. 2015).
- How to handle noise in high dimensional data space
- Evaluating suitability of similarity measure to find similarity between objects defined over high dimensional data space.

The present research contribution addresses the problem of dimensionality reduction designing an appropriate similarity measure for classification and clustering text data, text stream data.

Let 'D' indicate a data object defined over a finite set of representative attributes. Now, any randomly chosen data object defined by less than 10 attributes is treated to be low dimensional and the one which is defined by more than 10 attributes is treated to be high dimensional data object (Han et al. 2012b; VinayKumar et al. 2015). Clustering high dimensional data may be thought and viewed as a search problem of finding the clusters and the spatial dimensional over which these clusters may be generated, so as to be reliable (Aggarwal 2007; Han et al. 2012a). In Jiang et al. (2011c), the authors propose an approach for reducing the dimensionality of document-word matrix using feature clustering approach and then try to classify the test document using the reduced dimension matrix. The authors in Lin et al. (2013), introduce a new similarity measure (SMTP) for clustering text documents and document sets. The information and discussion on different algorithms, data stream models available in the literature is explained in Aggarwal (2007), Babcock et al. (2002), Tatbul and Zdonik (2006), Gaber et al. (2004). Chang and Lee (2005) discussed a sliding window based approach for finding frequent patterns in data streams. Various methods for clustering data streams are contributed in Charikar et al. (2003), Aggarwal et al. (2003), Gaber et al. (2005), Phridviraj et al. (2014). A more detailed discussion on research issues in clustering process is discussed in Sect. 2.

1.1 Need for Dimensionality Reduction

Dimensionality reduction is a key process which reduces the laborious computation process involved during clustering process. For instance, if we have 1000 dimensions, then the distance function that is applied requires computation on these 1000 dimensions. Each time during learning process, one has to consider all these 1000 dimensions for every computation. Another problem is w.r.t memory required. For example, if there are 10,000 documents and each document vector is frequency vector defined over 1000 features then the space required is equal to that required for 100,000,000 element values. Assuming that each element value requires 4 byte space, then the space required is equal to 4×10^8 byte, i.e. 381.46973 MB.

1.2 Motivation

In Jiang et al. (2011a), the basic Gaussian function is used for defining membership function. The basic Gaussian function is used for similarity computation. The membership function is product based membership function. Motivated from Jiang et al. (2011b, c), Lin et al.

(2014), Radhakrishna et al. (2016a, 2017a, c, e), Aljawarneh et al. (2017b), the proposed membership function is defined. The difference between the membership function defined in (Jiang et al. 2011a) and the proposed membership function is that the former one is product-based function whereas the present membership function is a summation-based function.

1.3 Research Issues

1.3.1 Distance Function

Distance functions are always important in clustering process. The implicit operation in clustering is finding distance between two cluster elements. Membership functions are helpful to know the similarity degree between cluster elements. Existing distance functions have problems such as sparseness problem, high dimensionality problem and sensitivity to larger values and do not take into account distribution behavior (Jiang et al. 2011a, c; Radhakrishna et al. 2015, 2016b, c, d, e, 2017a, b, e, g, h, i, 2018; Aljawarneh et al. 2016, 2017a; Chen et al. 2015; Sammulal et al. 2017; Usha Rani et al. 2018; Usha Rani and Sammulal 2017; SureshReddy et al. 2014; VinayKumar et al. 2015).

1.3.2 Cluster Quality

Cluster quality is another important parameter that is to be considered. Quality clusters show properties of high cohesion w.r.t intra cluster elements and low coupling w.r.t inter cluster elements. Approaches such as silhouette and TSS (wss + bss) may be used to evaluate cluster quality.

1.3.3 Dimensionality and Sparseness

Dimensionality is a must to be addressed problem for performing clustering (Gama 2013). Dimensionality introduces noise and outliers and several methods to eliminate noise and outlier data have been reviewed in literature. Feature selection method and feature reduction method are two techniques that serve the requirement. Data sparseness is another key issue that hinders clustering. Sparseness increases complexity problem. Sparseness is situations where more number of zeroes exist in matrix instead of non-zero values. Naturally, computational process can't discard this sparseness situation and must have to inevitably consider the data.

1.3.4 Feature Distribution

Retaining feature distribution is important to clustering (Tsai et al. 2009; Aljawarneh and Vangipuram 2018) and other approaches of learning (Neagoe and Neghina 2016; Hanneke 2016; Adeli et al. 2016). Data representation methods that assure the distribution detainment are important and helps to achieve good cluster quality of underlying clusters. Approaches for evaluating cluster quality are discussed and available in the literature.

Section 2 carries detailed literature review; Sect. 3 introduces proposed similarity function for feature clustering and dimensionality reduction which is inspired from Lin et al. (2014)

and Radhakrishna et al. (2017e); Sect. 4 performs analysis of similarity values; Sect. 5 gives working example; Sect. 6 outlines the results and discussions and Sect. 7 concludes the work.

2 Literature Review

Bingham and Mannila (2001) used a random projection in dimensionality reduction for text and image data. Information retrieval for noisy and noiseless data is used in applications in order to process the data. Though the random projection method is used for dimensionality reduction, Euclidean distance measure is scaled for the generated vectors with a reduced space (Bingham and Mannila 2001). PCA was used in an optimal way for data projection in a mean square sense. One of the other methods used in text document is discrete cosine transform (Bingham and Mannila 2001). However many researchers have performed dimensionality reduction using different techniques but, computation time and the efficiency is also to be considered. He et al. (2008) used a LLE algorithm to secure the local configuration to identify the nearest neighbors. By using the LLE algorithm, it is claimed that errors are minimized with the fixed reconstruction weights (He et al. 2008). One of the main drawbacks using LLE algorithm is mapping of data points are closer and also sample data is used to identify the Euclidean attribute.

Though many approaches are used, LLE algorithm has different views in performing dimensionality reduction. The views are linear versus non-linear, local versus global, supervised versus unsupervised (He et al. 2008). Mallick and Bhattacharyya (2012) describes the maximum margin criteria to perform the dimensionality reduction on term-document matrix. Cosine similarity measure is used between the document matrices to calculate the distance between the two sample documents (Mallick and Bhattacharyya 2012). Maximum margin criterion (MMC) is calculated using the optimal projection discriminant matrix which in turn results to uncorrelated local MMC. In Mallick and Bhattacharyya (2012) SVD and PCA are not used and found the method was computationally efficient and the promising result. The initial result in reducing the data was towards stop word removal and stemming word removal. The average recognition was found to be comparatively good with 98.3% for 6 classes.

One of the methods used by Pang et al. (2013) is class centroid based dimensionality reduction. The method used in Pang et al. (2013) showed the promising results for text classification. The method proposed in Pang et al. (2013) was centroid based represented by vector space model, term frequency and inverse document frequency to compute the similarity with the reduced document matrix. The centroid dimensionality reduction is used in two stages likely with class centroid generation and class centroid projection to identify the similar documents. Thus by performing the said approaches in Pang et al. (2013), the results were found good in fetching the lower dimensional data. As the research moved towards the dimensionality reduction and accuracy in detecting the results, one such approach used by Ganguly et al. (2015) was context driven dimensionality reduction. The approach used in Ganguly et al. (2015) is for clustering the text documents. The approach proposed is to effectively increase the efficiency in clustering the document of large dimensional dataset. Many evaluation metrics and most often the named entity recognition for dimensionality reduction is applied on the document matrix to fetch the desired outputs with great efficiency (Ganguly et al. 2015). K-means clustering and HAC are the traditional approaches applied to term-document to identify the similarity in the document.

Parallel rare term vector replacement algorithm for dimensionality reduction was proposed in Berka and Vajteršic (2013). The approach used was fast and effective in dimensionality reduction for the text documents and found promising results. The approach used

in Berka and Vajteršic (2013) is to convert high sparse corpus matrix document to dense sparse matrix to improve the efficiency in finding the similarity. Rare term vector (Berka and Vajteršic 2013) is used as vectors for different features to determine if the document contains particular feature. Initially the document is scanned and identified with rare elements and they are eliminated. Once the elimination is performed, the replacement of the term was applied which resulted (Berka and Vajteršic 2013) to give promising results. Truncation and unwanted feature elimination was performed for efficient computation. Parallelization was performed on a hybrid task with data partitioning to make sure the data is retrieved using parallel processing (Berka and Vajteršic 2013). Though the study in Johnson and Wichern (2007) says that there was independent component analysis (ICA) compared with SVD and PCA, the ICA has driven through the linear approximation approach of the original data which was again based on PCA. Multidimensional scaling is other approach (Hyvarinen et al. 2004) for the projection of the lower dimensional space for distance measure calculation. However though many approaches have come, most of the approaches are mainly based on SVD approach (Cox and Cox 2001). Thus the approach used in Cox and Cox (2001), Bartell et al. (1992), Berka and Vajteršic (2011), Paatero and Tapper (1994) represents the dimensionality reduction algorithm using self-organization approaches.

Stop word removal, stemming are the initial stages of any document classification procedure. The normalized data after the stemming is called as term matrix. After performing additional normalization, different dimensionality reduction techniques are applied to the document and the result found is usually better with reduced dimensionality. Feature extraction and feature selection are two main processes in classifying the text data. One important feature explained in Uguz (2012) was unsupervised dimension reduction methods for text clustering. Though the author Uguz (2012) used a hybrid method to create the informative reduced dimensional feature subspace, the author proposed a filter-wrapper and feature selection feature extraction methods to convert high dimensional data to low dimensional data. The approach used in Bharti and Singh (2014), Uguz (2012) has given better results by applying filter wrapper as dimensionality reduction process. Though many approaches were used in Uguz (2012), the result set has tried to generate only the specific feature subset. Unler et al. (2011), Bharti and Singh (2014) has proposed a new method using filter wrapper with mutual information available from filter model to weight the bit selection probabilities using SVM algorithm, the algorithm has failed to remove the noise from the datasets. Thus in the approach (Bharti and Singh 2014), a three stage unsupervised dimension reduction was used to give the best results and select the data accordingly.

As many approaches have evolved, one recent approach explained in Xu et al. (2018) was towards applying classical dimensional data reduction and sample selection methods for large scale data. Classic machine learning approaches has been applied to clustering and random forest algorithms to identify the best results. Deep learning and other approaches were used to perform the dimensionality reduction on the large datasets (Xu et al. 2018). Processing of large datasets has become a trivial task and how to get the efficiency is always trivial.

3 Similarity Function for Feature Clustering

The similarity function for feature clustering is described in this section. The proposed similarity function is given by Eq. (1)

$$Sim(\alpha_i, \alpha_j) = \frac{F(\alpha_i, \alpha_j) + \varphi}{\varphi + 1} \tag{1}$$

$$F(\alpha_i, \alpha_j) = \frac{\sum_{h=1}^{h=m} \mathcal{G}(\alpha_{ih}, \alpha_{jh})}{\sum_{h=1}^{h=m} \mathcal{H}(\alpha_{ih}, \alpha_{jh})} \tag{2}$$

where

$$\mathcal{G}(\alpha_{ih}, \alpha_{jh}) = \begin{cases} e^{-\left(\frac{\alpha_{ih}-\alpha_{jh}}{\sigma}\right)^2}; & \alpha_{ih} \neq 0 \, and \, \alpha_{jh} \neq 0 \\ -\varphi; & either \, \alpha_{ih} \, or \, \alpha_{jh} \, is \, 0 \\ 0; & both \, \alpha_{ih}, \, \alpha_{jh} \, are \, 0 \end{cases}$$

$$\mathcal{H}(\alpha_{ih}, \alpha_{jh}) = \begin{cases} 0; & both \, \alpha_{ih} \, and \, \alpha_{jh} \, are \, 0 \\ 1; & else \end{cases}$$

In the equation for similarity function, variables α_{ih}, α_{jh} are probability values. For instance, α_{ih} and α_{jh} denotes the probabilistic chance that word w_i and w_j may belong to a given class label, h.

The function $F(\alpha_i, \alpha_j)$ is the fraction of $\mathcal{G}(\alpha_{ih}, \alpha_{jh})$ and $\mathcal{H}(\alpha_{ih}, \alpha_{jh})$. The parameter φ is a constant and the value of φ that best fits is equal to 1. The highest possible value of $F(\alpha_i, \alpha_j)$ is equal to 1 and lowest possible value is $-\varphi$. As φ is set to 1, hence the lowest possible value is -1.

4 Analysis of Similarity Values

The similarity values attained using proposed function are analysed in the next three subsections. For analysis, three cases are considered. They are (a) similarity value in worst case, (b) similarity value in best case and (c) similarity value in average case.

4.1 Worst Case

In the worst case, each component of $\mathcal{G}(\alpha_{ih}, \alpha_{jh})$ is equal to $-\varphi$. The function $F(\alpha_i, \alpha_j)$ is computed as a fraction of $\mathcal{G}(\alpha_{ih}, \alpha_{jh})$ and $\mathcal{H}(\alpha_{ih}, \alpha_{jh})$.

$$\begin{aligned} F(\alpha_i, \alpha_j) &= \frac{\sum_{h=1}^{h=m} \mathcal{G}(\alpha_{ih}, \alpha_{jh})}{\sum_{h=1}^{h=m} \mathcal{H}(\alpha_{ih}, \alpha_{jh})} \\ &= \frac{-\varphi - \varphi - \varphi \cdots m \, times}{1 + 1 + 1 + \cdots m \, times} \\ &= -\varphi \end{aligned}$$

The resulting value of $F(\alpha_i, \alpha_j)$ is equal to $-\varphi$ in worst case.

So, the similarity value in worst case is obtained as 0.

$$Sim(\alpha_i, \alpha_j) = \frac{F(\alpha_i, \alpha_j) + \varphi}{\varphi + 1} = \frac{-\varphi + \varphi}{\varphi + 1} = 0$$

This proves that $Sim(\alpha_i, \alpha_j)$ has lower bound.

4.2 Best Case

In the best case, each component of $\mathcal{G}(\alpha_{ih}, \alpha_{jh})$ is equal to 1. The function $F(\alpha_i, \alpha_j)$ is computed as a fraction of $\mathcal{G}(\alpha_{ih}, \alpha_{jh})$ and $\mathcal{H}(\alpha_{ih}, \alpha_{jh})$.

$$\begin{aligned} F(\alpha_i, \alpha_j) &= \frac{\sum_{h=1}^{h=m} \mathcal{G}(\alpha_{ih}, \alpha_{jh})}{\sum_{h=1}^{h=m} \mathcal{H}(\alpha_{ih}, \alpha_{jh})} \\ &= \frac{1 + 1 + 1 + \cdots m\ times}{1 + 1 + 1 + \cdots m\ times} = 1 \end{aligned}$$

The resulting value of $F(\alpha_i, \alpha_j)$ is equal to 1 in best case.
So, the similarity value in best case is obtained as 1.

$$Sim(\alpha_i, \alpha_j) = \frac{F(\alpha_i, \alpha_j) + \varphi}{\varphi + 1} = \frac{1 + \varphi}{\varphi + 1} = 1$$

This proves that $Sim(\alpha_i, \alpha_j)$ has upper bound.

4.3 Average Case

In the average case, some component of $\mathcal{G}(\alpha_{ih}, \alpha_{jh})$ are defined by exponential function, some components are defined as $-\varphi$ and other remaining components can be zero. The function $F(\alpha_i, \alpha_j)$ is computed as a fraction of $\mathcal{G}(\alpha_{ih}, \alpha_{jh})$ and $\mathcal{H}(\alpha_{ih}, \alpha_{jh})$.

$$\begin{aligned} F(\alpha_i, \alpha_j) &= \frac{\sum_{h=1}^{h=m} \mathcal{G}(\alpha_{ih}, \alpha_{jh})}{\sum_{h=1}^{h=m} \mathcal{H}(\alpha_{ih}, \alpha_{jh})} \\ &= \frac{e^{-\left(\frac{\alpha_{ih} - \alpha_{jh}}{\sigma}\right)^2} + e^{-\left(\frac{\alpha_{ih} - \alpha_{jh}}{\sigma}\right)^2} + \cdots A\ times + -\varphi - \varphi \cdots B\ times}{1 + 1 + \cdots A\ times + 1 + 1 + \cdots B\ times} \end{aligned}$$

The above expression can be rewritten as given by

$$F(\alpha_i, \alpha_j) = \frac{A * e^{-\left(\frac{\alpha_{ih} - \alpha_{jh}}{\sigma}\right)^2} - \mathrm{B}^*\varphi}{(A + B)}$$

So, the expression for similarity value in average case is derived as

$$Sim(\alpha_i, \alpha_j) = \frac{\frac{A*e^{-\left(\frac{\alpha_{ih}-\alpha_{jh}}{\sigma}\right)^2} - B*\varphi}{(A+B)} + \varphi}{\varphi + 1}$$

$$\text{i.e.} \quad Sim(\alpha_i, \alpha_j) = \frac{A\left(e^{-\left(\frac{\alpha_{ih}-\alpha_{jh}}{\sigma}\right)^2} + \varphi\right)}{A+B}(1+\varphi)$$

The analysis proves that the similarity function has tight upper and lower bounds. The upper bound value is one and lower bound value is zero.

5 Text Clustering Process

Incremental method is another method used for clustering of data. In this method, we start with first word pattern chosen as mean and this word pattern acts element of the first cluster. Using this cluster mean, we now take each word pattern and find the similarity of the new word pattern with the first cluster's mean. If this similarity is greater than mentioned threshold value, then w.r.t given similarity threshold, we place the word pattern into the first cluster and find the resultant mean of cluster. If the similarity is less than the given similarity threshold, we place word pattern in the new cluster. For each word pattern, we find its similarity with the existing clusters and place it in the cluster to which it is more similar. Thus, we can generate finite clusters. This, method is better than k-means because if a new word pattern is to be added to the existing clusters, we can achieve it by just finding the similarities with all the existing cluster means.

The Step by step algorithm flow is as follows:

Algorithm: Incremental Clustering and Dimensionality reduction

Step-1: Generate word patterns using the procedure defined in Jiang et al. (2011a,b).
Step-2: Select the next word pattern and compare it with the first word pattern which is cluster-1. If similarity condition is satisfied, then the word pattern is added to the cluster. If the condition is not satisfied, create new cluster.
Step-3: If a word pattern is added to cluster, update cluster mean.
Step-4: Repeat procedure in step-2 for all word patterns.
Step-5: Output all the generated clusters and update their mean and deviation.
Step-6: After generation of clusters from word patterns, use these cluster information to represent data as a matrix such that rows of this matrix correspond to words and columns represent clusters.
Step-7: Use the matrix in step-6 to determine reduced document matrix. To obtain reduced document matrix multiply the original document matrix with the resultant matrix obtained in step-6.
Step-8: Perform classification and prediction using classifiers such as kNN, Decision tree based classifier, SVM etc.

5.1 Overview of Dimensionality Reduction Process

Table.1 : Matrix in Frequency Form

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
File -1	0	1	0	0	1	1	0	0	0	1
File-2	0	0	0	0	0	2	1	1	0	0
File -3	0	0	0	0	0	0	1	0	0	0
File-4	0	0	1	0	2	1	2	1	0	1
File-5	0	0	0	1	0	1	0	0	1	0
File-6	2	1	1	0	0	1	0	0	1	0
File-7	3	2	1	3	0	1	0	1	1	0
File -8	1	0	1	1	0	1	0	0	0	0
File -9	1	1	1	1	0	0	0	0	0	0

Table.2: Matrix in Binary Form

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
File -1	0	1	0	0	1	1	0	0	0	1
File-2	0	0	0	0	0	1	1	1	0	0
File -3	0	0	0	0	0	0	1	0	0	0
File-4	0	0	1	0	1	1	1	1	0	1
File-5	0	0	0	1	0	1	0	0	1	0
File-6	1	1	1	0	0	1	0	0	1	0
File-7	1	1	1	1	0	1	0	1	1	0
File -8	1	0	1	1	0	1	0	0	0	0
File -9	1	1	1	1	0	0	0	0	0	0

Word Patterns of $\mathbf{W}$

	cluster-1	cluster-2	cluster-3	cluster-4
x1	1	0	0	0
x2	0	1	0	0
x3	0	1	0	0
x4	1	0	0	0
x5	0	0	1	0
x6	0	1	0	0
x7	0	0	0	1
x8	0	1	0	0
x9	0	0	0	0
x10	0	0	0	1

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{x}_8$	$\mathbf{x}_9$	$\mathbf{x}_{10}$
0.00	0.20	0.20	0.00	1.00	0.50	1.00	0.67	0.00	1.00
1.00	0.80	0.80	1.00	0.00	0.50	0.00	0.33	1.00	0.00

	dimension-1	dimension-2	dimension-3	dimension-4
d1	0	2	1	1
d2	0	2	0	1
d3	0	0	0	1
d4	0	3	1	2
d5	1	1	0	0
d6	1	3	0	0
d7	2	4	0	0
d8	2	2	0	0
d9	2	2	0	0

The demonstration using pictorial representation shows that the word distribution before dimensionality reduction and after dimensionality reduction remained same. Also, it is to be noted that proposed approach does not lose any information and do not add any noisy data. Hence, the reduced low dimensional matrix is suitable form for performing clustering and classification tasks. The approach followed for clustering word patterns is motivated from (Jiang et al. 2011b). In our approach for feature clustering, we considered the similarity function introduced in this paper.

6 Results

This section gives the experiment results using proposed approach. Figure 1 plots the number of features before dimensionality reduction (DR). The number of documents considered 200, 300, 400, 500, 600 and 700. The number of features are 2581, 2836, 3200, 3020, 3456, and 3263 respectively.

Figure 2 plots the number of features after dimensionality reduction using information gain followed by SVD and proposed approach for dimensionality reduction.

The number of features using feature clustering based DR is less than dimensionality reduction using information gain, followed by SVD. Another important advantage of proposed approach is the word distribution in documents before DR and after DR are same. This means we have transformed documents in one form to another form such that distributions are same before and after DR.

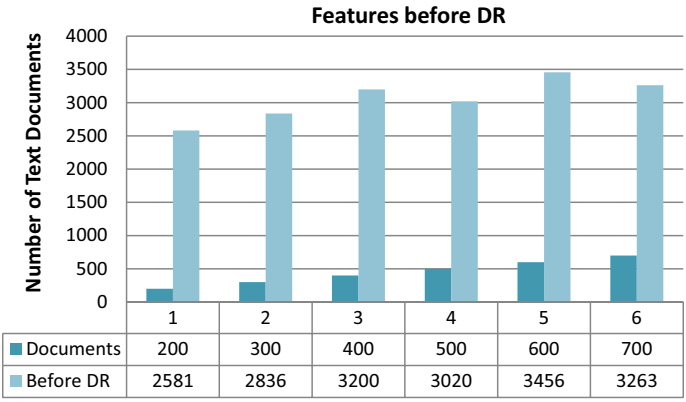


Fig. 1 Number of features after initial pre-processing

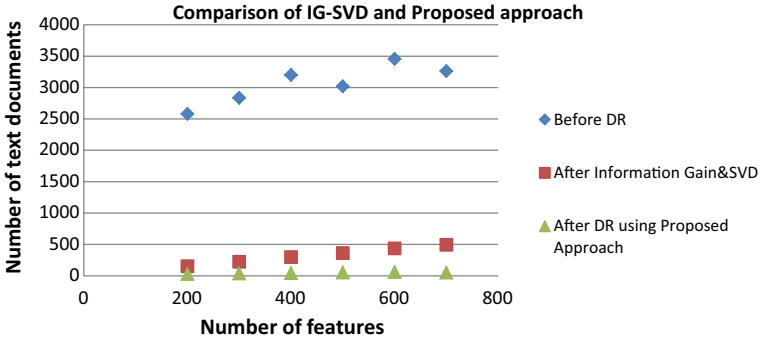


Fig. 2 Number of features after dimensionality reduction using two approaches

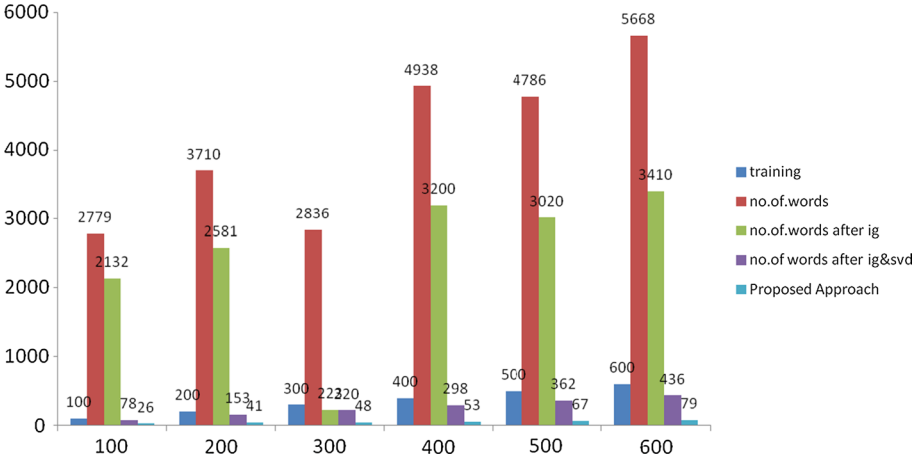


Fig. 3 Comparison of dimensionality reduction of various approaches

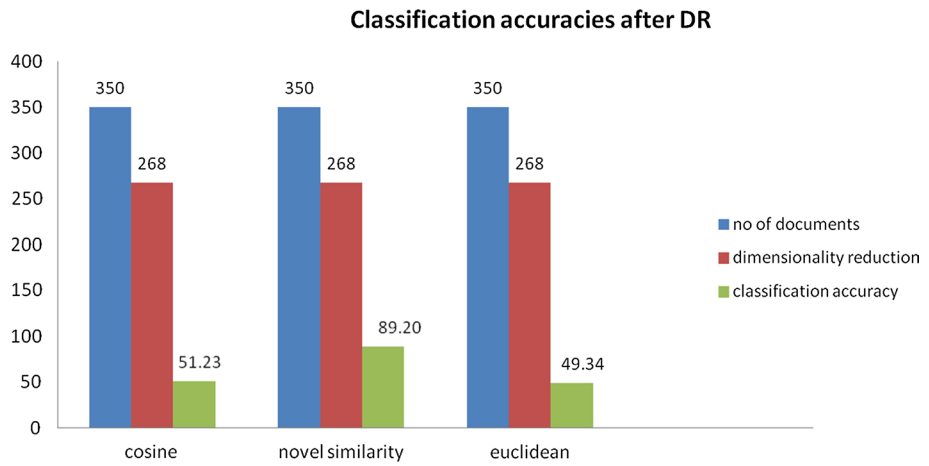


Fig. 4 Classifier accuracies

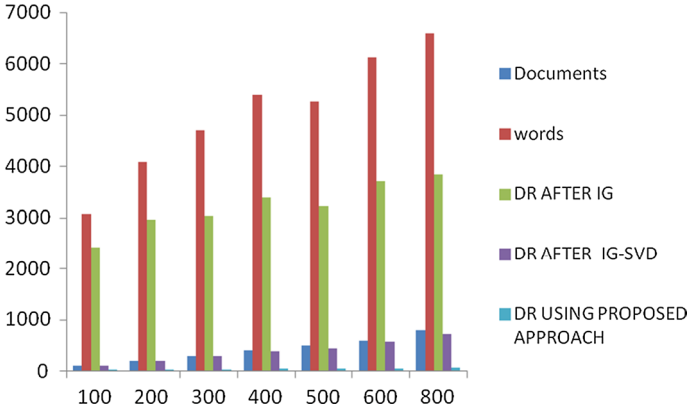


Fig. 5 Comparison of dimensionality reduction for random sample

Figure 3 shows the number of features using feature clustering based DR is less than dimensionality reduction using IG and information gain followed by SVD. The number of documents considered are a random sample of 100, 200, 300, 400, 500 and 600 documents from R8 dataset.

Classifier accuracies for 350 documents with 268 features after dimensionality reduction using proposed approach are computed for euclidean and cosine distance measures and these are compared to accuracies obtained using proposed measure. The accuracy using proposed measure is better to euclidean and cosine measures as depicted in Fig. 4.

Figure 5 shows the number of features using feature clustering based DR is less than dimensionality reduction using IG and information gain followed by SVD. The number of documents considered are a random sample of 100, 200, 300, 400, 500, 600 and 800 documents from R8 dataset.

Figure 6 shows the number of features using feature clustering based DR is less than dimensionality reduction using IG and information gain followed by SVD. The number

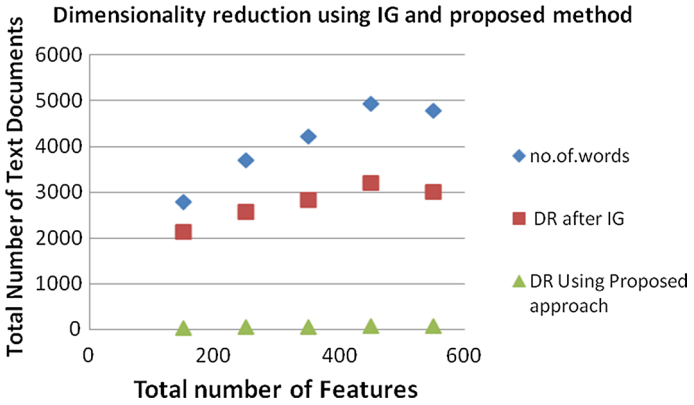


Fig. 6 Comparison of dimensionality reduction for random sample

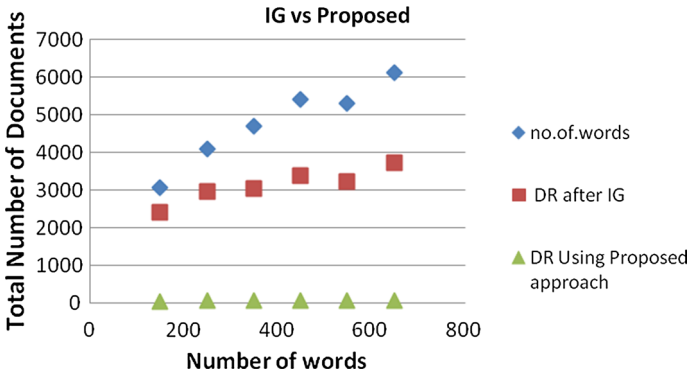


Fig. 7 IG versus proposed DR approach

of documents considered are a random sample of 150, 250, 350, 450, 550 documents from R58 dataset.

Figure 7 shows the number of features using feature clustering based DR is less than dimensionality reduction using IG and information gain followed by SVD. The number of documents considered are a random sample of 150, 250, 350, 450, 550 and 650 documents from R58 dataset.

Figure 8 gives the plot of number of features before and after dimensionality reductions carried by IG, SVD and proposed approach for a total of 500 text documents w.r.t trade class of Reuters dataset. The dimensionality using proposed approach is significantly less to other approaches and the distribution attained using proposed method is same as the distribution of words in documents before dimensionality reduction.

Figure 9 shows classifier accuracies of trade class of Reuter's dataset for 500 randomly chosen text documents. The classifier accuracies for kNN classifier for $k=3$ to $k=15$ are obtained for Euclidean, SMTP (Lin et al. 2014) and proposed measure. It can be concluded

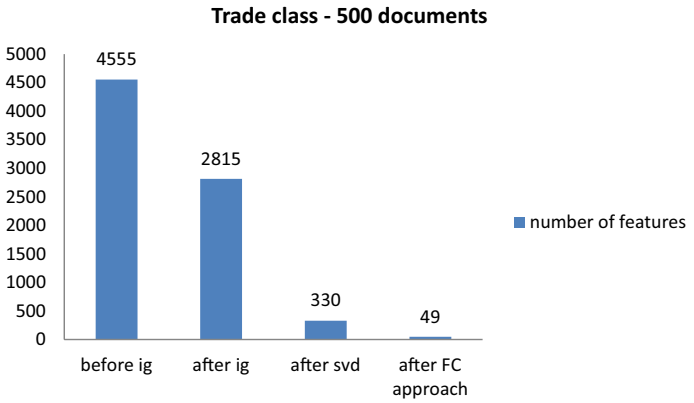
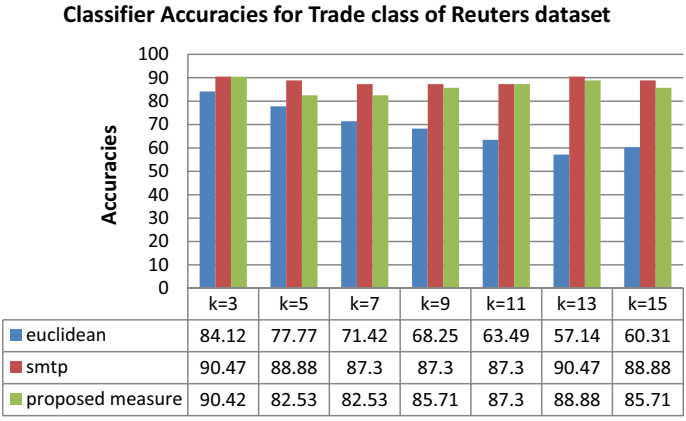


Fig. 8 IG versus proposed DR approach



	k=3	k=5	k=7	k=9	k=11	k=13	k=15
euclidean	84.12	77.77	71.42	68.25	63.49	57.14	60.31
smtp	90.47	88.88	87.3	87.3	87.3	90.47	88.88
proposed measure	90.42	82.53	82.53	85.71	87.3	88.88	85.71

Fig. 9 Classifier accuracies of trade class of Reuters dataset for randomly chosen 500 documents

from the experiments conducted that the proposed approach yields lesser dimensionality and at the same time retains distribution of words of documents.

7 Conclusions

Feature representation and dimensionality reduction techniques are two important tasks in text clustering and classification. In this paper, an approach for feature representation and dimensionality reduction of text documents is described. The feature representation and dimensionality reduction approaches introduced retains the distribution of features. Output of feature representation is a hard representation matrix. The hard matrix is used to obtain the low dimensionality document matrix. The input for clustering is the low dimensional matrix. The working of proposed approach is explained using a case study that supports the importance of the approach and advantage of dimensionality reduction. Experiment results

prove the importance of proposed approach for dimensionality reduction. A substantial amount of dimensionality reduction is achieved using proposed method and at the same time feature distribution after dimensionality reduction is retained w.r.t documents before dimensionality reduction. This makes the process important and allows obtaining better classification accuracies. As a future extension, there is a scope and possibility to devise new membership functions and apply them for clustering to obtain clusters with good cluster quality and subsequently obtain better classification accuracies.

References

Abualigah, L. M., Khader, A. T., Al-Betar, M. A., & Alomari, O. A. (2017). Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. *Expert Systems with Applications, 84,* 24–36.

Adeli, E., Shi, F., An, L., Wee, C. Y., Wu, G. R., Wu, T., et al. (2016). Joint feature-sample selection and robust diagnosis of Parkinson's disease from mri data. *Neuroimage, 141,* 206–219. https://doi.org/10.1016/j.neuroimage.2016.05.054.

Aggarwal, C. (2007). *Data streams models and algorithms*. Cham: Springer.

Aggarwal, C., Han, J., Wang, J., & Yu, P. (2003). A framework for clustering evolving data streams. In *VLDB conference*.

Aljawarneh, S., Radhakrishna, V., Kumar, P. V., & Janaki, V. (2016). A similarity measure for temporal pattern discovery in time series data generated by IoT. In *2016 international conference on engineering & MIS (ICEMIS), Agadir* (pp. 1–4).

Aljawarneh, S. A., Radhakrishna, V., & Cheruvu, A. (2017a). Extending the Gaussian membership function for finding similarity between temporal patterns. In *2017 international conference on engineering & MIS (ICEMIS)*, *Monastir* (pp. 1–6).

Aljawarneh, S. A., Radhakrishna, V., Kumar, P. V., & Janaki, V. (2017b). G-SPAMINE: An approach to discover temporal association patterns and trends in internet of things. *Future Generation Computer Systems, 74,* 430–443. https://doi.org/10.1016/j.future.2017.01.013.

Aljawarneh, S. A., & Vangipuram, R. (2018). GARUDA: Gaussian dissimilarity measure for feature representation and anomaly detection in Internet of things. *The Journal of Super Computing*. https://doi.org/10.1007/s11227-018-2397-3.

Babcock, B., Babu, S., Datar, M., Motwani, R., & Widom, J. (2002). Models and issues in data stream systems. In *Proceedings of PODS*.

Bartell, B. T., Cottrell, G. W., & Belew, R. K. (1992). Latent semantic indexing is an optimal special case of multidimensional scaling. In *Proceedings of SIGIR, ACM, USA* (pp. 161–167).

Berka, T., & Vajteršic, M. (2011). Dimensionality reduction for information retrieval using vector replacement of rare terms. In *Proceedings of TM*.

Berka, T., & Vajteršic, M. (2013). Parallel rare term vector replacement: Fast and effective dimensionality reduction for text. *Journal of Parallel and Distributed Computing, 73*(3), 341–351.

Bharti, K. K., & Singh, P. K. (2014). A three-stage unsupervised dimension reduction method for text clustering. *Journal of Computational Science, 5*(2), 156–169.

Bingham, E., & Mannila, H. (2001). Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'01)* (pp. 245–250). New York, NY: ACM. http://dx.doi.org/10.1145/502512.502546D.

Chang, J. H., & Lee, W. S. (2005). estWin: Online data stream mining of recent frequent item sets by sliding window method. *Journal of Information Science, 31*(2), 7690.

Charikar, M., Callaghan, L., & Panigrahy, R. (2003). Better streaming algorithms for clustering problems. In *Proceedings of 35th ACM symposium on theory of computing*.

Chen, Y. C., Peng, W. C., & Lee, S. Y. (2015). Mining temporal patterns in time interval-based data. *IEEE Transactions on Knowledge and Data Engineering, 27*(12), 3318–3331.

Cox, T. F., & Cox, M. A. A. (2001). *Multidimensional scaling*. Boca Raton: Chapman & Hall.

Gaber, M. M., Zaslavsky, A., & Krishnaswamy, S. (2004). Towards an adaptive approach for mining data streams in resource constrained environments. In *Proceedings of sixth international conference on data warehousing and knowledge discovery*. *Lecture Notes in Computer Science (LNCS)*. Cham: Springer.

Gaber, M. M., Zaslavsky, A., & Krishnaswamy, S. (2005). Mining data streams—A review. *SIGMODC Record, 34*(2), 18–26.

Gama, J. (2013). *Knowledge discovery from databases*. Boca Raton: CRC Press.

Ganguly, D., Leveling, J., & Jones, G. J. F. (2015). Context-driven dimensionality reduction for clustering text documents. In P. Majumder, M. Mitra, M. Agrawal, & P. Mehta (Eds.), *Proceedings of the 7th forum for information retrieval evaluation (FIRE'15)* (pp. 1–7). New York, NY: ACM.

Han, J., Kamber, M., & Pei, J. (Eds.). (2012a). Advanced cluster analysis. In *The morgan kaufmann series in data management systems, data mining* (3rd ed., pp. 497–541). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00011-3.

Han, J., Kamber, M., & Pei, J. (Eds.). (2012b). Classification: Basic concepts. In *The morgan kaufmann series in data management systems, data mining* (3rd ed., pp. 327–391). Morgan Kaufmann. https://doi.org/10.1016/B978-0-12-381479-1.00008-3.

Hanneke, S. (2016). The optimal sample complexity OF PAC learning. *Journal of Machine Learning Research, 17*(38), 1–15.

He, C., Dong, Z., Li, R., & Zhong, Y. (2008). Dimensionality reduction for text using LLE. In *2008 international conference on natural language processing and knowledge engineering, Beijing* (pp. 1–7).

Hyvarinen, A., Karhunen, J., & Oja, E. (2004). *Independent component analysis*. Hoboken: Wiley.

Jiang, J. Y., Cheng, W. H., Chiou, Y. S., & Lee, S. J. (2011a). A similarity measure for text processing. In *2011 international conference on machine learning and cybernetics*, *Guilin* (pp. 1460–1465). https://doi.org/10.1109/icmlc.2011.6016998.

Jiang, J. Y., Liou, R. J., & Lee, S. J. (2011b). A fuzzy self-constructing feature clustering algorithm for text classification. *IEEE Transactions on Knowledge and Data Engineering, 23*(3), 335–349. https://doi.org/10.1109/TKDE.2010.122.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th ed.). Upper Saddle River: Prentice Hall.

Lin, Y. S., Jiang, J. Y., & Lee, S. J. (2014). A similarity measure for text classification and clustering. *IEEE Transactions on Knowledge and Data Engineering, 26*(7), 1575–1590. https://doi.org/10.1109/TKDE.2013.19.

Lin, Y., et al. (2013). A similarity measure for text classification and clustering. *IEEE Transactions of Knowledge and Data Engineering, 26,* 1575–1590.

Mallick, K., & Bhattacharyya, S. (2012). Uncorrelated local maximum margin criterion: An efficient dimensionality reduction method for text classification. *Procedia Technology, 4,* 370–374.

Neagoe, V. E., & Neghina, E. C. (2016). Feature selection with ant colony optimization and its applications for pattern recognition in space imagery. In *IEEE ICC*.

Paatero, P., & Tapper, U. (1994). Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics, 5*(2), 111–126.

Pang, G., Jin, H., & Jiang, S. (2013). An effective class-centroid-based dimension reduction method for text classification. In *Proceedings of the 22nd international conference on World Wide Web (WWW'13 Companion)* (pp. 223–224). New York, NY: ACM.

Phridviraj, M. S. B., Srinivas, C., & GuruRao, C. V. (2014). Clustering text data streams. A tree based approach with ternary function and ternary feature vector. *Procedia Computer Science, 31,* 976–984.

Radhakrishna, V., Aljawarneh, S. A., Janaki, V., & Kumar, P. V. (2017b). Looking into the possibility for designing normal distribution based dissimilarity measure to discover time profiled association patterns. In *2017 international conference on engineering & MIS (ICEMIS), Monastir* (pp. 1–5).

Radhakrishna, V., Aljawarneh, S. A., Kumar, P. V, et al. (2017c). ASTRA—A novel interest measure for unearthing latent temporal associations and trends through extending basic Gaussian membership function. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-017-5280-y.

Radhakrishna, V., Aljawarneh, S. A., Kumar, P. V., & Choo, K.-K. R. (2016a). A novel fuzzy Gaussian-based dissimilarity measure for discovering similarity temporal association patterns. *Soft Computing, 1,* 1. https://doi.org/10.1007/s00500-016-2445-y.

Radhakrishna, V., Aljawarneh, S. A., Kumar, P. V., & Janaki, V. (2017a). A novel fuzzy similarity measure and prevalence estimation approach for similarity profiled temporal association pattern mining. *Future Generation Computer Systems, 1,* 1. https://doi.org/10.1016/j.future.2017.03.016.

Radhakrishna, V., Kumar, P. V., Aljawarneh, S. A., & Janaki, V. (2017e). Design and analysis of a novel temporal dissimilarity measure using Gaussian membership function. In *2017 international conference on engineering & MIS (ICEMIS)*, *Monastir* (pp. 1–5).

Radhakrishna, V., Kumar, P. V., & Janaki, V. (2015). An approach for mining similarity profiled temporal association patterns using Gaussian based dissimilarity measure. In *Proceedings of the international conference on engineering & MIS 2015 (ICEMIS'15)*.

Radhakrishna, V., Kumar, P. V., & Janaki, V. (2016b). A computationally optimal approach for extracting similar temporal patterns. In *2016 international conference on engineering & MIS (ICEMIS), Agadir* (pp. 1–6).

Radhakrishna, V., Kumar, P. V., & Janaki, V. (2016c). Looking into the possibility of novel dissimilarity measure to discover similarity profiled temporal association patterns in IoT. In *2016 international conference on engineering & MIS (ICEMIS), Agadir* (pp. 1–6).

Radhakrishna, V., Kumar, P. V., & Janaki, V. (2017a). A computationally efficient approach for mining similar temporal patterns. In Matoušek R (Eds.), *Recent advances in soft computing. ICSC-MENDEL 2016. Advances in intelligent systems and computing* (Vol. 576). Springer, Cham.

Radhakrishna, V., Kumar, P. V., & Janaki, V. (2017e). SRIHASS—A similarity measure for discovery of hidden time profiled temporal associations. *Multimedia Tools and Applications.*. https://doi.org/10.1007/s11042-017-5185-9.

Radhakrishna, V., Kumar, P. V., & Janaki, V. (2017g). Design and analysis of similarity measure for discovering similarity profiled temporal association patterns. *IADIS International Journal on Computer Science and Information Systems, 12*(1), 45–60.

Radhakrishna, V., Kumar, P. V., & Janaki, V. (2017h). Normal distribution based similarity profiled temporal association pattern mining (N-SPAMINE). *Database Systems Journal, 7*(3), 22–33.

Radhakrishna, V., Kumar, P. V., & Janaki, V. (2018). Krishna Sudarsana: A Z-space similarity measure. In *Proceedings of the fourth international conference on engineering & MIS 2018 (ICEMIS'18)*. New York, NY: ACM, Article 44, 4 pp.

Radhakrishna, V., Kumar, P. V., Janaki, V., & Aljawarneh, S. (2016d). A similarity measure for outlier detection in timestamped temporal databases. In *2016 international conference on engineering & MIS (ICEMIS), Agadir* (pp. 1–5).

Radhakrishna, V., Kumar, P. V., Janaki, V., & Aljawarneh, S. (2016e). A computationally efficient approach for temporal pattern mining in IoT. In *2016 international conference on engineering & MIS (ICEMIS), Agadir* (pp. 1–4).

Radhakrishna, V., Kumar, P. V., Janaki, V., & Cheruvu, A. (2017i). A dissimilarity measure for mining similar temporal association patterns. *IADIS International Journal on Computer Science and Information Systems, 12*(1), 126–142.

Sammulal, P., Usha Rani, Y., & Yepuri, A. (2017). A class based clustering approach for imputation and mining of medical records (CBC-IM). *IADIS International Journal on Computer Science & Information Systems*, *12*(1), 61–74.

SureshReddy, G., Rajinikanth, T. V., & Ananda Rao, A. (2014). Design and analysis of novel similarity measure for clustering and classification of high dimensional text documents. In B. Rachev & A. Smrikarov (Eds.), *Proceedings of the 15th international conference on computer systems and technologies (CompSysTech'14)* (pp. 194–201). New York, NY: ACM. http://dx.doi.org/10.1145/2659532.2659615.

Tatbul, N., & Zdonik, S. (2006). A subset-based load shedding approach for aggregation queries over data streams. In *Proceedings of international conference on very large data bases (VLDB)*.

Tsai, S. C., Jiang, J. Y., Wu, C., & Lee, S. J. (2009). A fuzzy similarity-based approach for multi-label document classification. In *2009 second international workshop on computer science and engineering, Qingdao* (pp. 59–63). https://doi.org/10.1109/wcse.2009.766.

Uguz, H. (2012). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems, 24,* 1024–1032.

Unler, A., Murat, A., & Chinnam, R. (2011). mr2pso: A maximum relevance minimum redundancy feature selection method based on swarm intelligence for support vector machine classification. *Information Sciences, 181,* 4625–4641.

Usha Rani, Y., & Sammulal, P. (2017). An approach for imputation of medical records using novel similarity measure. In R. Matoušek (Eds.), *Recent advances in soft computing. ICSC-MENDEL 2016. Advances in Intelligent Systems and Computing* (Vol. 5). Cham: Springer.

Usha Rani, Y., Sammulal, P., & Golla, M. (2018). An efficient approach for imputation and classification of medical data values using class-based clustering of medical records. *Computers & Electrical Engineering, 66,* 487–504.

VinayKumar, K., Srinivasan, R., & Singh, E. B. (2015). A feature clustering approach for dimensionality reduction and classification. In R. Matoušek (Eds.), *Mendel 2015. ICSC-MENDEL 2016. Advances in intelligent systems and computing* (Vol. 378). Cham: Springer.

Xu, X., Liang, T., Zhu, J., Zheng, D., & Sun, T. (2018). Review of classical dimensionality reduction and sample selection methods for large-scale data processing. *Neurocomputing*. ISSN 0925-2312.

Vinay Kumar Kotte is currently with the Department of Computer Science and Engineering, Kakatiya Institute of Technology & Science, Warangal, Telangana, India. He is a also a research scholar at Karunya Institute of Technology and Sciences (Deemed to be university), Coimbatore, Tamilnadu, India and is working under the supervison of professors R. Srinivasan and R. Elijah Blessing. He has an academic teaching experience of 10 years and is actively involved in research for the past few years. Some of his research interests include data mining, machine learning, software engineering and algorithm design. He is also a member of ISTE and ACM. He has published papers in different international conferences within and outside to his credits. He contributed some of the papers presented in international conferences and also published some articles in peer reviewed journals.

Srinivasan Rajavelu is a highly motivated seasoned telecommunication solution sales consultant with more than 18 years of international experience in the technical areas of 5G Core, Service-Based architecture, Evolved Packet Core, PCRF, DPI, DRA/DEA, SBC, STP, SS7 Firewall, Diameter Firewall, HSS, Network Security, Data Mining, Internet Caching, CDN, IoT, NFV, SDN and Cloud Services. He received the Ph.D. degree in Faculty of Information and Communication Engineering from Anna University, India in 2007. His research areas include network security, mobile computing, wireless and adhoc networks and cloud computing. To his credit, he has good number of publications and has published peer reviewed articles. He contributed some of the papers presented in international conferences and also published some articles in peer reviewed journals.

Elijah Blessing Rajsingh is the Professor and Registrar of Karunya Institute of Technology and Sciences, India. He received his Master of Engineering with distinction from College of Engineering, Anna University, Chennai. He received the Ph.D. degree in Information and Communication Engineering from College of Engineering, Guindy, Anna University, India in 2005. He joined Karunya Institute of Technology and Sciences, Coimbatore in 1997. His research areas include network security, mobile computing, wireless and adhoc networks, medical image processing, parallel and distributed computing, grid computing and pervasive computing. To his credit, he has good number of publications and has published articles in Springer. He is an Associate Editor for International Journal of Computers & Applications, Acta Press, Canada and member of the editorial review board for many peer reviewed journals. He has been identified as an expert member of National Board of Accreditation (NBA), India. He is also being funded by Indian Council of Medical Research for his research project in Health Systems Research. He is a member of ISTE, CSI, and IEEE and has served as international advisory board member for various international conferences and workshops.