Indian Journal of Science and Technology, *Vol 10(6), DOI: 10.17485/ijst/2017/v10i6/104768, February 2017*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Optimal Feature for Text Similarity based Hybrid Clustering Technique with Aid of MGWO

Surya Narayana Goddumarri* and Vasumathi Devara

JNTUH, Hyderabad - 500085, Telangana, India; surya.aits@gmail.com, rochan44@gmail.com

Abstract

Objectives: The way toward gathering high dimensional information into groups is not exact and maybe not up to the level of desire when the dimensions of the dataset is high. It is presently centering gigantic consideration towards innovative work. **Methods/Analysis:** Initially the input high dimensional data is fed to similarity measure for text processing for feature selection, in which similarity between the categorical data is evaluated. Then we have planned to utilize optimal feature selection method. Feature determination is a vital subject in data mining, particularly for high dimensional datasets. In our proposed technique, Modified Grey Wolf Optimization technique is used for optimal feature selection. Next the selected features are grouped with the help of clustering technique. Here we are hybrid two clustering techniques for grouping the optimal features. **Findings:** The performance of the proposed technique is evaluated by means of clustering accuracy, Jaccard coefficient and Dice's coefficient. The proposed technique is compared with existing clustering algorithms. **Novelty/Improvements:** The primary intension of this research is to achieving promising results in text similarity based clustering technique. Here we are hybridizing k means and fuzzy c means clustering algorithm for grouping the optimal features.

Keywords: Fuzzy C Means Clustering, Grey Wolf Optimization, Jaccard Coefficient and Dice's Coefficient, K Means, Similarity Measure for Text Processing

1. Introduction

The measure of information is expanding step by step so there is a requirement for proficient data processing applications. Information can be in various structures it can be either in text, picture, and spatial structure and so on. Among this the most widely recognized type of information that we are taking care of colossally is the text information. The news stories we are perusing, postings and messages on online networking all are for the most part in text structure. A summary is a text that is delivered from one or more documents, which consists of a basic bit of the data from the first document, and that is no more than even 50% of the first text. So now a day there is awesome criticalness for the text mining process[1,2]. After mining of data in order to reduce the text clustering is performed. Because of the immense information size and many-sided quality, information dimensionality lessening has additionally been an essential concern. Extraordinary levels of endeavors have been placed in this heading, so that the significant issue of condemnation of dimensionality can be diminished. Text documents clusterization has been given careful consideration[3]. Creating strategies to sort out a lot of unstructured text documents into a littler number of important clusters would be exceptionally useful as text clustering is fundamental to such errands as ordering, separating, computerized metadata era, and word sense disambiguation. So clustering of documents is a programmed gathering of text documents into clusters in such a way that documents inside a cluster have high likeness in contrast with each other, yet are not quite the same as documents in different clusters[4].

The standard meaning of clustering in view of similarity is to organize information objects into discrete clusters with the end goal that the intra-cluster similarity and in addition the inter-cluster dissimilarity is amplified since

**Author for correspondence*

the similarity measure assumes an essential part in text classification and clustering algorithm[5,6]. Discovering similarity between words is a key portion of text similarity which is then utilized as an essential stage for sentence, section and document similarities. Words can be similar in two ways lexically and semantically. Corpus-Based similarity is a semantic similarity measure that decides the similarity between words as per data picked up from vast corpora[7]. In agglomerative strategies, for example, single connection and complete connection, similarity between individual items is adequate, yet in partitioned clustering, for example, k-means and k medoids cluster delegate is additionally required to quantify object-to-cluster similarity[8]. Subsequent to distinguishing similarity it is vital to perform clustering. Clustering all in common is a vital and helpful procedure that consequently arranges a gathering with a considerable number of information items into a much littler number of reasonable gatherings. The goal of clustering is to discover inherent structures in information, and compose them into significant subgroups for further study and examination. There have been numerous clustering algorithms distributed each year[9].

Since most clustering algorithms need a vectorial representation of the items, the clear cut components are generally characterized as a double vector, with each position characterizing whether the object has a given element or not[10]. In addition to that commonly clustering utilizing distance capacities called distance based clustering is an extremely prominent procedure to cluster the items and has given great results[11]. On the basis of this clustering technique characterizes gigantic information as indicated by their qualities, and it is used for different applications, for example, text clustering, sentence clustering, network clustering and handling information in sensor systems[12]. In numerous text processing exercises, text clustering assumes an imperative part. Case in point, different creators have contended that joining text clustering into extractive multi document outline maintains a strategic distance from issues of substance cover, prompting better scope[13]. Without clustering in these frameworks, it would not be conceivable to sort out documents into accumulations; such association encourages key tasks at the levels of capacity recuperation and union[14]. This could be made much clear utilizing the following phenomena. We intend to cluster documents on the basis of the similarity of one's sub charts in the document diagrams and consequently this strategy helps in sparing the time[15].

In text classification issues, the representation of a record strongly affects the execution of learning frameworks. The large dimensionality of the traditional organized representations can prompt difficult calculations because of the considerable size of genuine information. Subsequently, there is a requirement for decreasing the amount of took care of data to enhance the classification procedure. L. Borrajo, et al.[16] proposed a strategy to decrease the dimensionality of a traditional content representation in light of a clustering procedure to gathering records, and a formerly created Hidden Markov Model to signify them. They had connected tests with the k-NN and SVM classifiers on the OHSUMED and TREC benchmark content corpora utilizing the proposed dimensionality diminishment system. The test yields were exceptionally agreeable contrasted with regularly utilized systems like Info Gain and the factual tests performed showed the appropriateness of the proposed strategy for the preprocessing venture in a content classification errand.

With the developing significance for examination of the textual similarity and between the client substances, Kuldeep Singh, et al.[17] had highlighted textual similarity between different individuals in an informal organization. Words utilized as a part of social locales were utilized for finding textual similarity. On the premise of the regular words utilized as a part of interpersonal organizations, they had defined a metric. The information had been separated from interpersonal interaction destinations and after that that was handled for producing the measurements. They contrasted normal k-means and spectral k-means algorithms for finding textual similarity. They had utilized Word Net to gathering words together taking into account their implications.

In these days, numerous organizations institutionalize their operations through Business Process (BP), which are put away in archives and reutilized when new functionalities are needed. In any case, discovering particular procedures may turn into an unwieldy assignment because of the substantial size of these storehouses. Hugo Ordonez, et al.[18] presented Multi modal Group, a model for grouping and seeking business forms. The grouping system was based upon a clustering algorithm that utilized a similarity capacity taking into account of fuzzy logic; that grouping was performed utilizing the aftereffects of every client demand. By that's part, the pursuit depended on a multimodal representation that incorporated textual and structural data of BP. The appraisal of the proposed model was completed in two stages: 1. Inner quality evaluation of groups and 2. Outer evaluation of the made groups contrasted with a perfect arrangement of groups. The appraisal was performed utilizing a closed BP gathering planned cooperatively by 59 specialists. The trials brought in every stage were promising and confirmed the legitimacy of the proposed model.

Melissa Ailema, et al.[19] demonstrated how the particularity measure could serve as a valuable paradigm for co-clustering record term grids. They showed and explored the execution of CoClus, a novel, powerful block-diagonal co-clustering algorithm which specifically augmented that seclusion measure. The boost was performed utilizing an iterative substituting optimization system as opposed to algorithms that utilized spectral relaxations of the discrete optimization issues. Broad near investigations performed on different report term datasets exhibited that that methodology was exceptionally compelling, stable, and beats other block-diagonal co-clustering algorithms gave to the same errand. Another essential preferred standpoint of utilizing modularity in the co-clustering connection was that that gave a novel, straightforward method for deciding the proper number of co-clusters.

Text classification can help clients to adequately handle and adventure valuable data covered up in vast scale archives. Yet, the scarcity of information and the semantic affectability to setting regularly thwart the classification execution of short texts. Keeping in mind the end goal to defeat the shortcoming, Wang, et al.[20] introduced a brought together system to grow short texts taking into account word inserting Clustering and Convolution Neural Network (CNN). Experimentally, the semantically related words were normally near each other in implanting spaces. Accordingly, they first found semantic cliques through quick clustering. At that point, by utilizing added substance piece over word embeddings from connection with variable window width, the representations of multiscale semantic units in short texts were processed. In implanting spaces, the limited Nearest Word Embeddings (NWEs) of the semantic units were constituted extended lattices, where the semantic coteries were utilized as supervision data. At long last, for a short content, the anticipated lattice and extended frameworks were joined and bolstered into CNN in parallel. Exploratory results on two open benchmarks accepted the adequacy of the proposed technique.

As of late, regular scene content identification increases expanding consideration since it assumes an imperative part in numerous PC related systems. Zailiang Chen, et al.[21] conveyed a text detection technique comprising of two noteworthy strides: Connected Components (CCs) extraction and non-text separating. For CCs extraction, a multi-scale adaptive color clustering methodology was proposed, which could remove content from pictures in various shading complexities and was powerful to difference variety. For non-text separating, they joined Text Covariance Descriptor (TCD) with Histogram of Oriented Gradients (HOG) to develop highlight vectors and utilized them to recognize text from foundation at character and text line levels. Additionally, another text line era system joining both refined and foul CCs was connected, which could recover some miseliminated characters and created more incorporated text lines. Trials were directed on two freely accessible datasets, the ICDAR 2013 and the ICDAR 2011 datasets, the acquired F-measures on which were 0.76 and 0.75, separately. Relative results with text discovery algorithms exhibited that the proposed strategy accomplished focused execution on text detection.

Wireless sensor networks are occupied with different information gathering applications. Arunraja, et al.[22] indicated that the real bottleneck in wireless information gathering frameworks was the limited vitality of sensor hubs. By preserving the on board vitality, the life range of wireless sensor networks could be very much augmented. Information correspondence being the overwhelming vitality devouring movement of wireless sensor networks, information diminishment could serve better in saving the nodal vitality. Spatial and transient relationship among the sensor information was abused to lessen the information correspondences. Data similar cluster arrangement was a powerful approach to abuse spatial relationship among the neighboring sensors. By sending just a subset of information and evaluation the rest utilizing that subset was the contemporary method for misusing transient relationship. In distributed similarity based clustering and compressed forwarding for wireless sensor networks, they developed information comparative iso-clusters with negligible correspondence overhead. The intra-cluster correspondence was lessened utilizing adaptive normalized least mean squares based double expectation structure. The cluster head lessened the inter cluster information payload utilizing a lossless compressive sending strategy. The proposed work accomplished huge information lessening in both the intra-cluster and the inter cluster connections, with the ideal information precision of gathered information.

2. Proposed Methodology

The primary intension of this research is to achieving promising results in text similarity based clustering technique. Initially the input high dimensional data is fed To Similarity Measure for Text Processing (SMTP) for feature selection, in which similarity between two categorical data is evaluated. Then we have planned to utilize optimal feature extraction process. For high dimensional data, feature extraction process is the most important one. In our proposed technique, Modified Grey Wolf Optimization (MGWO) technique is used for optimal feature extraction. Next the selected features are grouped with the help of clustering technique. Here we are hybridized two clustering algorithm for grouping the optimal features. The implemented method employed k means and fuzzy c means clustering algorithm for grouping the optimal features. The semantic structure of the proposed technique is shown in Figure 1. It is shown in beneath,

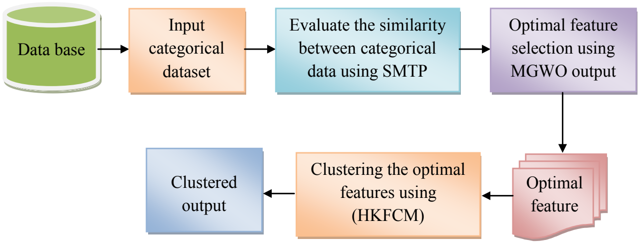
Data base
Input categorical dataset
Evaluate the similarity between categorical data using SMTP
Optimal feature selection using MGWO output
Optimal feature
Clustering the optimal features using (HKFCM)
Clustered output

Figure 1. The semantic diagram of propose technique.

The overall process of proposed technique has three main phases namely similarity measure phase, feature selection phase and clustering phase. Initially we have to find the similarity between the categorical data with the help of SMTP. At first the input data is fed to the SMTP and the resultant similarity features are given to the further process. The detail explanation is illustrated in beneath,

Phase 1: Similarity Measure for Text Processing (SMTP)

Based on the two features the similarity of categorical data is measured such as feature appears in any one of the data or both data or none of the categorical data. Based on the preferable properties mentioned above, we present a similarity measure, *SMTP* for cd_1 and cd_1 are,

$$SMTP(cd_1, cd_2) = \frac{F(cd_1, cd_2) + \gamma}{1 + \gamma} \tag{1}$$

Where, *F*is the function, which is defined for the two categorical data cd_1=< cd_{11}, cd_{12},...cd_{1n} and cd_{12}=<cd_{11},cd_{11} ,....cd_{1m} > and as follows:

$$F\left(cd_1, cd_2\right) = \frac{\sum_{k=1}^{l} N*\left(cd_{1k}, cd_{2k}\right)}{\sum_{k=1}^{l} N_{\cup}\left(cd_{1k}, cd_{2k}\right)} \tag{2}$$

Where,

$$N*\left(cd_{1k}, cd_{2k}\right) = \begin{cases} 0.5\left(1+\exp\left\{\left[\left(\frac{cd_{1k}-cd_{2k}}{\sigma_k}\right)^2\right]\right\}\right) & ,if\ cd_{1k}cd_{2k} > 0 \\ 0 & ,\ if\ cd_{1k} = 0\ \ and\ \ cd_{2k} = 0 \\ -\gamma & ,\ Otherwise \end{cases} \tag{3}$$

$$N_{\cup}(cd_{1k}, cd_{2k}) = \begin{cases} 0 & ,if\ cd_{1k} = 0\ and\ cd_{2k} = 0 \\ 1, & otherwise \end{cases} \tag{4}$$

After finding the similarity between the categorical data, the resultant similarity output is fed to the feature selection process. The clear explanation of the feature selection phase is described in below,

Phase 2: Feature Selection

In our proposed technique the optimal features are selected with the help of optimization technique. Here we are using Modified Grey Wolf optimization (MGWO) technique for optimal feature selection. The step by step procedure of MGWO is described in below section,

2.1 Modified Grey Wolf Optimization (MGWO)

The grey wolves sufficiently frame a Canidae's piece family and are respected as the apex predators presenting their position at the wherewithal's food chain. They routinely show an inclination to make due as a group. The heads constitute a male and a female, labeled as alpha, which are for the most part in charge of taking suitable choices viewing different factors; they are, basically, auxiliary wolves which adequately offer some assistance to the alpha in the choice making or comparable group functions. The selections made using the alpha are approved on to the set. The Beta expresses to the second rank in the striking order of the grey wolves. They are, essentially, supplementary wolves that sufficiently deal some aid to the alpha in the choice creating or equivalent set performances. The omega, which is at the least strata of the grey wolf pack, by and large functions as a substitute offering into the other leading wolves very nearly on each event and is permitted to have just the little scraps taking after a great blowout by the leader wolves. In GWO method the hunting (optimization) is guided by the α , β , δ and ω . For selecting the optimal features the proposed technique use modified grey wolf optimization algorithm. Here the traditional GWO technique is modified with the help of crossover and mutation process. In GWO, the solution will be updated by means of crossover and mutation operation. The pseudo code for the modified GWO algorithm is illustrated in beneath,

2.2 Pseudo Code for MGWO

```
Step 1: Initialize the solution $SF_i=(i=1,....n)$
        Initialize $a, A,$ and $C$ are the coefficient vector
Step 2: Find the fitness of the initial solution $Fit_i$=max
    Accuracy
Step 3: Separate the solution based on the fitness
    $SF_\alpha$ = the first best search solution
    $SF_\beta$= the second best search solution
    $SF_\delta$= the third best search solution
Step 4: Update the position of the current search solution
    $F_p(t+1) = \frac{F_{p1} + F_{p2} + F_{p3}}{3}$
Step 5: Update the new search solution by crossover and
    mutation
Step 6: Calculate the fitness of the new search solution
Step 7: Store the best solution so far attained
Iteration=Iteration+1
Step 8: Stop after the optimal solution is attained.
```

The step by step process of gray wolf optimization algorithm is mentioned below,

***Step 1:* Initialization process**

Initialize the input features from the output of SMTP and *a, A, and C* as coefficient vectors.

***Step 2:* Fitness evaluation**

Evaluate the fitness performance on the basis of the equation (5) and following that pick the best result.

$$Fit_i = \max Accuracy \tag{5}$$

***Step 3:* Separate the solution based on the fitness**

Currently, we discover the distinct result on the basis of the fitness value. Let the first best fitness results be F_α, the second best fitness results F_β and the third best fitness solutions F_δ.

Step 4: Update the position

We assume that the alpha (best candidate solution), beta and delta have the improved knowledge about the potential location of the prey in order to reproduce mathematically the hunting behavior of the grey wolves. As a result, we hoard the first three best solutions attained so far and require the other search agents (including the omegas) to revise their positions according to the position of the best search agent. For repetition, the new solution F_p *(t+1)* is estimated by using the formulae mentioned below.

$$\vec{K} =| \vec{C}.F_p(t+1) - F_p(t) | \tag{6}$$

Where,

$$\vec{K}^\alpha =| \vec{C}_1.F_{p\alpha} - F_p |, \quad \vec{K}^\beta =| \vec{C}_2.F_{p\beta} - F_p |, \quad \vec{K}^\delta =| \vec{C}_3.F_{p\delta} - F_p |$$

$$F_p(t+1) = \frac{F_{p1} + F_{p2} + F_{p3}}{3} \tag{7}$$

Where,

$$F_{p1} = F_{p\alpha} - \vec{A}_1.(\vec{K}^\alpha), F_{p2} = F_{p\beta} - \vec{A}_2.(\vec{K}^\beta), \ F_{3p} = F_{p\delta} - \vec{A}_3.(\vec{K}^\delta)$$

$$\vec{A} = 2\vec{a}r_1 - \vec{a} \ \text{ And } \ \vec{C} = 2r_2 \tag{8}$$

Where, t represents the iteration number, *p(t)* represents the prey position, *A* and *C* represents the coefficient vector, $\vec{a}$ is linearly decreased from 2 to 0, r_1 *and* r_2 represents the random vector [0, 1].

It can be detected that the last situation would be in a haphazard place ordered a circle which is well-defined using the locations of alpha, beta, and delta in the search space. In other words alpha, beta, and delta assess the position of the prey and other wolves update their positions arbitrarily around the prey. Exploration and exploitation are definite by means of the adaptive values of a and A. The adaptive values of parameters a and A permit GWO to easily transition amongst exploration and exploitation. With diminishing A, half of the iterations are committed to the investigation (|*A*|<*1*) and the other half are dedicated to the usage. Enclosing the conduct, the subsequent equations are utilized keeping in mind the end goal to give numerical model.

Step 5: Cross over and Mutation

The crossover operation has many methods to produce the offspring. They are one point, two points, uniform, and arithmetic crossover, in these process single point crossover used. A crossover operates by randomly selecting a crossover point within a chromosome, then interchanging the two parent chromosomes between these points to produce two new offspring illustrated in Figure 2.

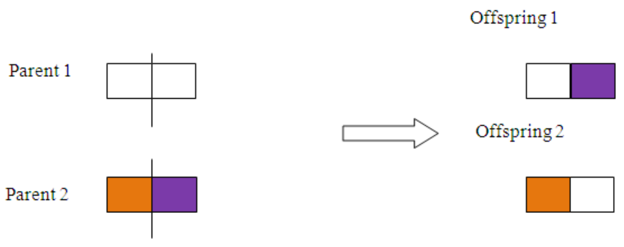

Figure 2. Single point crossover.

After that, the child chromosomes are altered for raising the effectiveness of the solution. Alteration is the process of creating new offspring from the single parent and preserves the variety of the every chromosome illustrated in Figure 3. There is a possibility to find the gene of a child to modify arbitrarily. This gene process is better than the old parents.

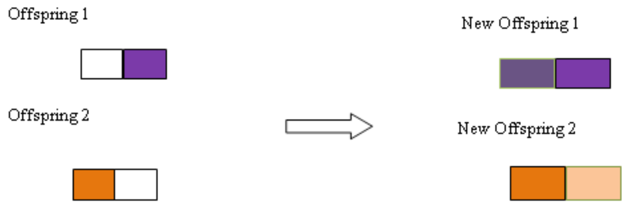

Figure 3. Mutation process.

Step 6: Fitness calculation

Calculate the fitness of the new search solution using the equation (5). And then store the best solution.

Step 7: Stopping criteria

Repeat step 3 to 6, until a better fitness or maximum number of iterations are met. Based on above mention process attain the optimal features. Thenthe optimal features are used for the further process.

Phase 3: Clustering process

The selected optimal features are grouped with the help of clustering technique. Clustering is the most prominent data mining technique used for grouping the data into clusters based on distance measures. Here we are hybridized two clustering algorithm for grouping the optimal features. The implemented method employed k means and fuzzy c means clustering algorithm for grouping the optimal features. Initially the optimal features are given to the input for k means algorithm, in which the centroid value is selected. After the selection of centroid value, it will be given to the fuzzy c means clustering algorithm for grouping the input categorical data. The detail explanation of hybrid clustering process is shown in beneath,

2.3 K-Means Clustering Algorithm

One of the most widely used clustering algorithms is K-Means clustering. This minimizes the mean squared Euclidean distance from each input feature to its nearest centre. Here we have a good control upon the number of clusters produced. The underlying process employs an easy and effortless method to categorize the optimal features into a specific number of clusters. Initially we fix the number of cluster K and then we choose the centroids value arbitrarily. The k means clustering algorithm has three main steps, it is specified beneath.

Step1: Randomly pick K points as centroids of k clusters.
Step2: For each point assign the point to the nearest cluster. And then the cluster centroids are recomputed.
Step3: Repeat Step2 (until there is no change in clusters between consecutive iterations).

After the centroid selection, the selected centroids are given to the input for fuzzy c means clustering algorithm for grouping the available optimal features based on their similarity.

2.4 Fuzzy C Means Clustering Algorithm (FCM)

Fuzzy C-Means (FCM) is a data clustering technique in which each and every data in that group will comes under one cluster based on the membership function. In high dimensional search space, it will group all the data in to specific number of clusters. The degrees of the cluster are defined by the membership function in terms of [0, 1]. Which gives the flexibility that the data point can belong to more than one cluster? The proposed method use of FCM for clustering the input data. The objective function of proposed algorithm is effectively explained as follows.

$$Objec\ Fun = \sum_{i=1}^{n} \sum_{j=1}^{c} U_{ij}^{m} \parallel O_{fi} - c_j \parallel^2 \tag{9}$$

Where,
"U_{ij}" is the membership of j^{th}data in the i^{th}cluster c_j
"c" is the cluster center
"O_f" is the optimal feature

"*m*" is the any real number greater than one

"||*||" is the similarity between any measured optimal feature and the center

Now the cluster centre calculation is done by equation (2),

$$c_j = \frac{\sum_{i=1}^{n} U_{ij}^{m} O_{fi}}{\sum_{i=1}^{n} U_{ij}} \tag{9}$$

Membership updation is done by equation (3),

$$U_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| O_{fi} - c_j \right\|}{\left\| O_{fi} - c_k \right\|} \right)^{\frac{2}{m-1}}} \tag{10}$$

If $\left\| U^{(K+1)} - U^{(K)} \right\| < \in$ then stop, Where, "$\in$" is a termination criterion between 0 and 1.

Based on the above procedure of FCM the input optimal feature is clustered. After the FCM process, we obtain the number of cluster set such as C_1, C_2, C_3,.... C_n. The efficiency of the proposed hybrid k means and FCM clustering algorithm is evaluated by means of clustering accuracy. The performance of the proposed technique is evaluated in section.4. It is shown in below,

3. Results and Discussion

In this section we discuss the result obtained from the proposed technique. The proposed methodology is developed in MATLAB.

3.1 Dataset Description

The experimental results are analyzed with the help of mushroom dataset. The datasets are taken from UCI machine learning repository. There are 22 attributes and 8124 instance and the number of attribute missing values are 2480. The dataset is available at https://archive.ics.uci.edu/ml/datasets/Mushroom.

3.2 Evaluation Metrics

To evaluate the clustering performance, the proposed method uses different types of measures such as clustering accuracy, Jaccard coefficient and Dice's coefficient.

3.2.1 Jaccard Coefficient

The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets.

$$J(G_1,G_2) = \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|} = \frac{|G_1 \cap G_2|}{|G_1| + |G_2| - |G_1 \cap G_2|} \tag{12}$$

If G_1 and G_2 are both empty $J(G_1,G_2)=1$

Otherwise $0 \le J(G_1,G_2) \le 1$

3.2.2 Dice's Coefficient

$$D(G_1,G_2) = \frac{2|G_1 \cap G_2|}{|G_1| + |G_2|} \tag{13}$$

3.2.3 Accuracy

$$Accuracy = \frac{correctly\ cluster}{Total\ cluster\ size} * 100 \tag{14}$$

3.3 Performance Analysis

The main objective of this paper is to efficiently cluster the categorical data using optimal feature selection and hybrid clustering algorithm. Initially the input dataset is fed to SMTP then the similarity features are optimally selected by means of modified grey wolf optimization algorithm. To improve the clustering efficiency, the proposed method, use hybrid k means and fuzzy c means clustering algorithm. The parameters utilized in proposed system are specified in Table 1. Here we are varying the number of cluster size for our proposed technique.

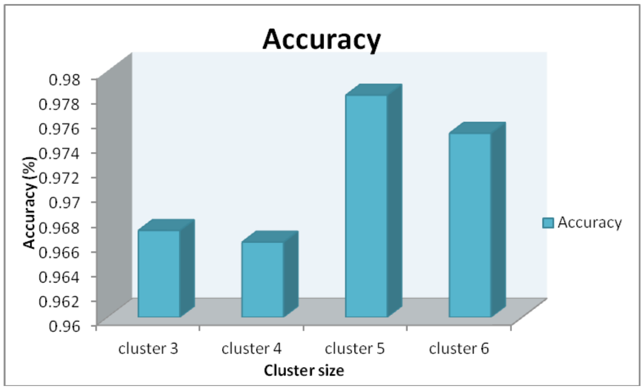

Figure 4. The performance of proposed clustering accuracy.

By varying the number of cluster size the proposed technique performance is evaluated. When we fix the cluster size is two the suggested technique attains the accuracy of 96.70%. If the cluster size is four, the accuracy value of the implemented technique is 96.61%. Likewise we are varying the cluster size is five and six, the proposed technique reach 97.80% and 97.49% of accuracy value. The graphical representation of the proposed clustering accuracy is shown in Figure 4.

Table 1. Performance of proposed method

Cluster Size	Accuracy (%)	JCC	DCC
Cluster 3	96.7049	0.721112	0.041749
Cluster 4	96.6101	0.607549	0.060632
Cluster 5	97.8032	0.725938	0.074608
Cluster 6	97.4938	0.727847	0.023523

The proposed technique also evaluates the Jaccard coefficient and Dice's coefficient by varying the cluster size. Initially we fix the cluster size is two, in which the proposed technique achieves 0.721112 for JCC and 0.041749 for DCC value. If we change the cluster size like four, five and six the suggested technique attains JCC value is 0.607549, 0.725938 and 0.727847. And DCC value is 0.060632, 0.074608 and 0.023523. The graphical representation of the Jaccard coefficient and Dice's coefficient is shown in Figure 5 and Figure 6. It is plotted in beneath,

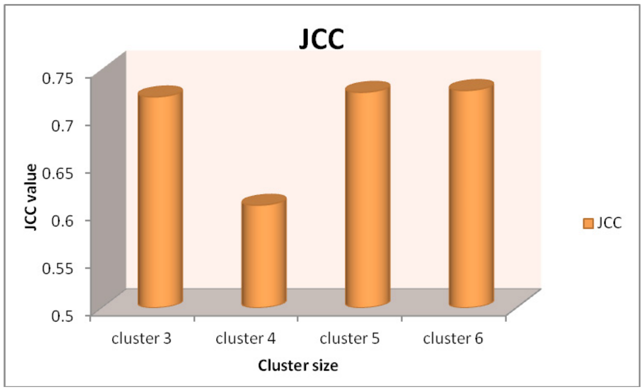

Figure 5. The Jaccard coefficient value for suggested technique.

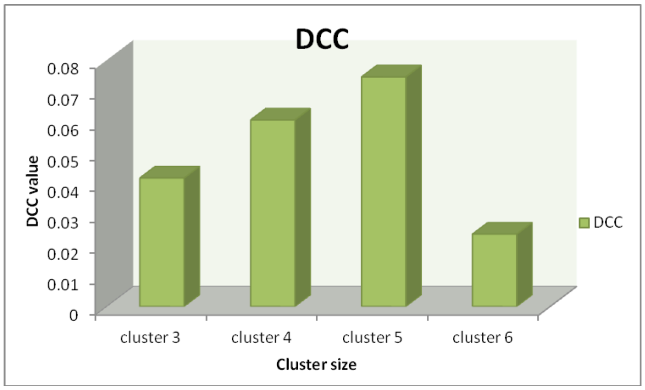

Figure 6. The Dice's coefficient value for implemented technique.

3.4 The Effectiveness of the Proposed Technique

In this section, we explain the efficiency of the text similarity based hybrid clustering technique. To prove the efficiency of our work, we compare the proposed work to existing work[23]. In existing method the categorical data is grouped with the help of PFCM and optimal MST. The comparison result for proposed technique with existing technique is tabulated in Table 2. It is shown in below,

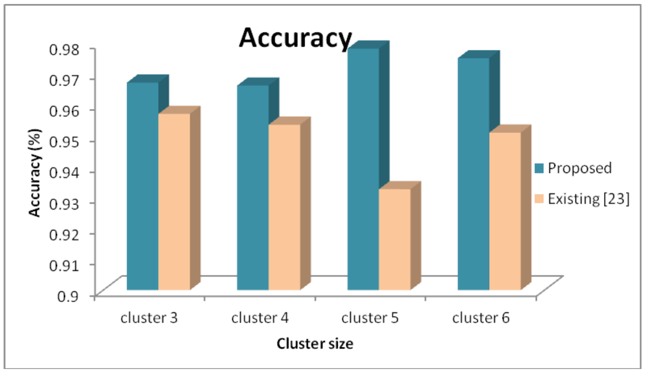

Figure 7. Clustering accuracy comparison by varying the cluster size.

When analyzing the table 2, the proposed text similarity based hybrid clustering and optimal feature selection

Table 2. Comparative analysis of proposed against existing methods

Cluster Size	Accuracy (%)		JCC		DCC	
	Proposed	Existing[23]	Proposed	Existing[23]	Proposed	Existing[23]
Cluster 3	96.7049	95.6936	0.721112	0.625938	0.041749	0.051332
Cluster 4	96.6101	95.3457	0.607549	0.527847	0.060632	0.086235
Cluster 5	97.8032	93.2614	0.725938	0.583533	0.074608	0.086667
Cluster 6	97.4938	95.0894	0.727847	0.673031	0.023523	0.037235

technique is compared with existing[23]. The graphical representation of the comparison result for clustering accuracy is shown in Figure 7.

Inspecting this figure 6, we are able to identify that the proposed hybrid clustering approach possesses outperformed with higher accuracy value of 97.15% when compared to existing[23].By varying the number of cluster size the proposed clustering accuracy also changes. For cluster size three, the proposed hybrid clustering technique achieves 96.7049% accuracy value when the existing[23] achieves 95.6936% accuracy value. The overall accuracy value suggested technique is 96.6101% for cluster size four. But the existing[23] achieves 95.3457% accuracy value. For cluster size five, the existing[23] achieves the 93.2614% but our proposed method achieves the accuracy value of 97.8032%. For cluster size six, the proposed method achieves the accuracy value of 97.4938% which is maximum value when compared to the existing technique[23]. It is shown in Figure 8.

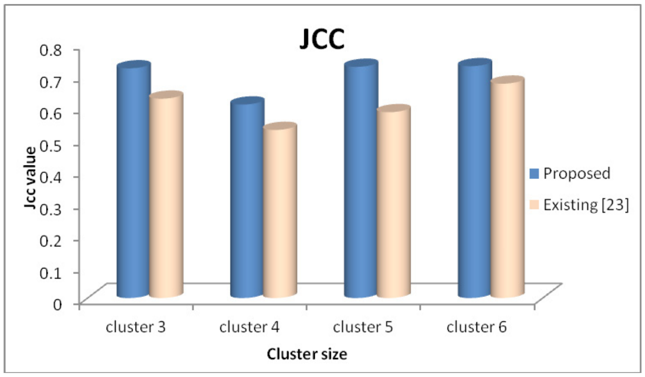

Figure 8. JCC comparison by varying the cluster size.

For cluster size three, the proposed hybrid clustering technique achieves 0.721112 JCC value when the existing[23] achieves 0.625938 JCC value. The JCC value of suggested technique is 0.607549 for cluster size four. But the existing[23] achieves 0.527847 JCC value. For cluster size five, the existing[23] achieves the JCC value of 0.583533 but our proposed method achieves the JCC value of 0.725938. For cluster size six, the proposed method achieves the JCC of 0.727847 which is maximum value when compared to the existing technique[23]. It is shown in Figure 9.

For cluster size three, the proposed hybrid clustering technique achieves 0.041749 DCC value when the existing[23] achieves 0.051332 DCC value. The DCC value of suggested technique is 0.060632 for cluster size four. But the existing[23] achieves 0.086235 DCC value. For cluster size five, the existing[23] achieves the DCC value of 0.086667 but our proposed method achieves the DCC value of 0.074608. For cluster size six, the proposed method achieves the DCC of 0.023523 which is minimum value when compared to the existing technique[23].From the result the proposed clustering method achieves the better result. The proposed clustering method achieves the maximum clustering accuracy value, maximum Jaccard coefficient value and minimum Dice's coefficient value when compared to the existing method[23].

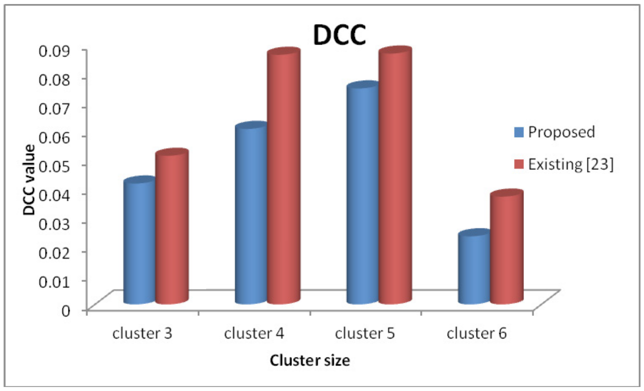

Figure 9. DCC comparison by varying the cluster size.

4. Conclusion

The effective clustering technique for categorical data is proposed in this paper. At first, Similarity Measure for Text Processing value is evaluated for each categorical data. Then the features are optimally selected by means of Modified Grey Wolf Optimization. To improve the clustering efficiency of the proposed method we employed hybrid k means and fuzzy c means clustering algorithm for grouping the input data. The performance of the proposed technique is evaluated by clustering accuracy, Jaccard coefficient and Dice's coefficient. The proposed method achieves the accuracy value of 97.15% and the existing technique attains 94.84% which is minimum value when compared to the suggested technique. The Jaccard coefficient of our proposed technique attains 0.6956 but the existing method attains only 0.60258. The suggested method reaches the minimum Dice's coefficient of 0.05012 but the existing technique reaches the Dice's coefficient of 0.065366. In future the researcher will have sufficient opportunities to perform efficient clustering technique and produce newer heights of excellence in performance.

5. Reference

1. Thomas AM, Resmipriya MG. An efficient text classification scheme using clustering. International Conference on Emerging Trends in Engineering, Science and Technology. 2016; 24:1220-5. https://doi.org/10.1016/j.protcy.2016.05.095
2. Deshpande AR, Lobo LM. Text summarization using clustering technique. International Journal of Engineering Trends and Technology. 2013; 4:3348-51.
3. Puri S, Kaushik S. A technical study and analysis on fuzzy similarity based models for text classification. International Journal of Data Mining and Knowledge Management Process. 2012; 2:1-15. http://aircconline.com/ijdkp/V2N2/2212ijdkp01.pdf
4. Patil PI, Singh G. A comprehensive survey of the existing text clustering techniques. International Journal of Scientific Development and Research. 2016; 1:291-3.
5. Patil D, Dongre Y. A clustering technique for email content mining. IJCSIT. 2015; 7:73-81. https://doi.org/10.5121/ijcsit.2015.7306
6. Warad VC, Baron Sam B. Incremental MVS based clustering method for similarity measurement. International Journal of Computer Science and Information Technologies. 2014; 5:1486-91.
7. Gomaa WH, Fahmy AA. A survey of text similarity approaches. International Journal of Computer Applications. 2013; 68:13-8. https://doi.org/10.5120/11638-7118
8. Rezaei M, Franti P. Matching Similarity for Keyword-based Clustering. 2011. p. 1- 10. PMid:22038677
9. Chaudhari PJ, Dharmadhikari DD. Clustering with multi-viewpoint based similarity measure: An overview. International Journal of Engineering Inventions. 2012; 1:1-5.
10. De Franca FO. A hash-based co-clustering algorithm for categorical data. Elsevier on Expert Systems with Applications. 2016; 64:24-35. https://doi.org/10.1016/j.eswa.2016.07.024
11. Irani J, Pise N, Phatak M. Clustering techniques and the similarity measures used in clustering: A survey. International Journal of Computer Applications. 2016; 134:9-14. https://doi.org/10.5120/ijca2016907841
12. Yun U, Ryang H, Kwon OC. Monitoring vehicle outliers based on clustering technique. Applied Soft Computing Journal. 2016; 1-41. https://doi.org/10.1016/j.asoc.2016.09.003
13. Saranya J, Arunpriya C. Survey on clustering algorithms for sentence level text. International Journal of Computer Trends and Technology. 2014; 10:61-6. https://doi.org/10.14445/22312803/IJCTT-V10P111
14. Zamora J, Mendoza M, Allende H. Hashing-based clustering in high dimensional data. Expert Systems with Applications. 2016; 62:202-11. https://doi.org/10.1016/j.eswa.2016.06.008
15. Deepthi AL, Prasad J. Hierarchal clustering and similarity measures along with multi representation. International Journal of Research in Engineering and Technology. 2013; 2:76-9. https://doi.org/10.15623/ijret.2013.0208012
16. Vieira AS, Borrajo L, Iglesias EL. Improving the text classification using clustering and a novel HMM to reduce the dimensionality. Computer Methods and Programs in Biomedicine. 2016; 136:1-22.
17. Singh K, Shakya HK, Biswas B. Clustering of people in social network based on textual similarity. Perspectives in Science. 2016; 8:1-5. https://doi.org/10.1016/j.pisc.2016.06.023
18. Ordo-ez A, Ordo-ez H, Corrales JC, Cobos C, Wives LK, Thom LH. Grouping of business processes models based on an incremental clustering algorithm using fuzzy similarity and multimodal Search. Expert Systems with Applications. 2016; 67:1-21.
19. Ailem M, Role F, Nadif M. Graph modularity maximization as an effective method for co-clustering text data. Elsevier on Knowledge-based Systems. 2016; 109:1-48. https://doi.org/10.1016/j.knosys.2016.07.002
20. Wang P, Xu B, Xu J, Tian G, Liu CL, Hao H. Semantic expansion using word embedding clustering and convolution neural network for improving short text classification. Neurocomputing. 2016; 174:806-14. https://doi.org/10.1016/j.neucom.2015.09.096
21. Wu H, Zou B, Zhao YQ, Chen Z, Zhu C, Guo J. Natural scene text detection by multi-scale adaptive color clustering and non-text filtering. Neuro Computing. 2016; 1-15. https://doi.org/10.1016/j.neucom.2016.07.016
22. Arunraja M, Malathi V, Sakthivel E. Distributed similarity based clustering and compressed forwarding for wireless sensor networks. ISA Transactions. 2015; 59:180-92. https://doi.org/10.1016/j.isatra.2015.07.014 PMid:26343165
23. Narayana SG, Vasumathi D. Text Similarity based Clustering Technique for High Dimensional Categorical Data. 2016. p. 1-5.