The Journal of Supercomputing
https://doi.org/10.1007/s11227-018-2526-z

High-performance link-based cluster ensemble approach for categorical data clustering

N. Yuvaraj[1] · **C. Suresh Ghana Dhas**[2]

© Springer Science+Business Media, LLC, part of Springer Nature 2018

Abstract

In recent years, the clustering ensembles emerged as a problem solver for extracting the data points into clusters in an efficient way. However, still clustering poses a serious issue due to the presence of imperfect information while partitioning the data into clusters. This creates a serious issue in creating an efficient cluster with cluster ensembles. In this paper, we propose a solution to solve the degradation in clustering during data partitioning. The initial clusters are generated using firefly algorithm. A linked cluster ensemble approach uses similarity measurement using multi-viewpoint and weighted triple quality using entropy measurements that ensembles the data points into clusters. This avoids the problem of local optimum and avoids the issues arise from high-dimensional datasets and improve the quality of clustering. Here, the data partitioning is done with bipartite spectral algorithm and similarity measurement. Finally, the artificial neural network is used to generate classified results from the optimized clustered datasets. The experimental results are carried out over UCI repository datasets and the results show that the proposed method attains an effective ensemble clustering with higher clustering accuracy than the conventional ones.

Keywords Multi-viewpoint similarity measure · Bipartite spectral algorithm · Hybrid firefly algorithm · Entropy weighted triple quality · Artificial neural network

✉ N. Yuvaraj
yraj1989@gmail.com

C. Suresh Ghana Dhas
sureshc.me@gmail.com

[1] Department of Computer Science and Engineering, St. Peter's Institute of Higher Education and Research, St. Peter's University, Avadi, Tamil Nadu, India

[2] Department of Computer Science and Engineering, Vivekanandha College of Engineering for Women, Tiruchengode, Tamil Nadu, India

Published online: 27 September 2018

1 Introduction

Data clustering is a fundamental tool to know the dataset structure. It plays a crucial role in data mining, machine learning, pattern recognition and retrieval. Clustering groups the datasets into clusters, where the data in cluster are similar than the other clusters. Clustering is widely applied in image processing, pattern recognition, market analysis, etc., and this works well with any datasets. Many clustering algorithm to cluster the categorical data is available in the literature that includes protein interaction data [8] and it uses Gower's similarity coefficient [3]. The *k*-mode [10] has been extended to centroid update rule using dissimilarity matching technique and frequency update method. Single pass algorithm uses pre-assigned similarity threshold value to find the clusters (either old or new) for examining the assigned data points. Information bottleneck framework is used by a clustering algorithm with good scalability, and it is hierarchical to identify the distance of categorical data points, i.e., LIMBO [1]. Other algorithms like genetic and evolutionary computing has already been used for categorical datasets. Further, graph models like ROCK [9], STIRR [7] and CLICK [24] are also used. Additionally, density algorithms are used like graph models, i.e., CACTUS [6]. Other techniques for categorical datasets include No Free Lunch theorem [21], where not a single algorithm is efficient in being best for the datasets [14] and it identifies the shapes and structures of the related dataset [22]. These algorithms in clustering the categorical datasets have its own strength and weaknesses; however, based on its parametric setting, it provides a distinctive solution.

Clustering categorical information is comparatively demanding than clustering numeric information. In numeric information, the inherent geometric possessions can be castoff in significant distance functions among data facts. In case of categorical facts, a distance or distinction purpose cannot be well-defined unswervingly. Owing to the superior possessions of definite attributes, the grouping of categorical information appears more complex than that of numerical information. An insufficient procedure has been anticipated in current years for grouping categorical information. It is adequate to plug out that all the attribute standards are treated uniformly in the resemblance calculation of those accessible algorithms. That is, unusual attribute value counterpart has the similar donation to parallel computation like those of recurrent ones.

However, these limitations are avoided using cluster ensemble technique, which is seen as an effective solution for clustering algorithms, which uses categorical datasets. It improves well the robustness and clustering quality. Besides its advantages, it possesses certain disadvantage which provide incorrect clusters during final data partitioning. The information or binary clustering ensemble matrix brings the relations between the data points, where many null entries are left unknown and this reduces the data partition quality. Such degradation in data partition quality, a linked approach is presented that uses refined cluster association matrix. It forms a relation between the cluster and that improves the quality of data partition and finds the unknown entries by measuring the cluster ensemble similarity. This ensemble technique is used for combining the multiple partitions over different clusters and forms a single clustering solution to attain better accuracy, increased robustness and improved clustering quality.

The present work concentrates in categorical datasets and grouping them into clusters and each category has its own value. This technique uses similarity measure based

on multi-viewpoint and WTQ using entropy measurement, which measures the similarity between the cluster points and improves the quality of clustering. Initially, the base clusters are generated using firefly algorithm (FFA). Strehl and Ghosh [19] said that multi-viewpoint similarity measure is suitable for cluster points using its similarity measurement. The relative similarities between the clustering data points is measured using the proposed method, and the results are provided with different viewpoints. Once the similarity is measured, the partition between the data occur. The problems in partitioning the data points are resolved with spectral co-clustering or bipartite algorithm [5]. This results in good bipartition. Ensemble method implemented over ANN [13] provides high-performance classification systems, once the data partition is done. The proposed method aims at selecting the best subset formed with error-independent sets. Nanda and Panda [15], proposes a stochastic global optimization that clusters the categorical data sets. The present technique with above methods well identifies the real number available in clusters over the categorical dataset.

The outline of the paper is presented as follows: Sect. 2 provides the related works. Section 3 discusses the proposed modules. Section 4 provides the experimental results and discussions. Finally, Sect. 5 concludes the paper with future work.

2 Related work

Iam-On et al. [12] analyzed the problem of degradation associated with clustering quality and presented a link-based algorithm that improves the matrix computation for finding the unknown elements in the dataset using cluster ensemble similarity. The final clustering quality is obtained using graph partitioning over weighted bipartite graph from the refined matrix. The results over real-time datasets proved that the link-based algorithm outperforms well than conventional clustering over categorical datasets and other cluster ensemble algorithms.

Bai and Liang [2] approach a new term to reflect the separation exist in cluster. An optimization function is developed using k-mode and other objective function. This optimization framework updates well and provides convergence over several iterations. The effectiveness of clustering is improved with k-mode algorithm, and the results over large categorical dataset proved better scalability and better linear time complexity over data objects and clusters.

Iam-on and Garrett [11] used link cluster ensemble technique to improve the clustering quality using internal and external criteria. The results proved better clustering quality over real and synthetic datasets.

Pavithra and Chandrakala [17] used C-Rank link algorithm for improving the quality of clusters and improved ranking of clusters over the weighted networks. This algorithm has 3 phases: (1) identifying the clusters, (2) ranking the data points using integrated cohesion and (3) eliminating the reduced clusters. Finally, graph partitioning is applied over weighted bipartite graph to attain the refined matrix.

Chen and Liu [4] used entropy property for finding the set of candidates in categorical dataset. This helps to obtain best K candidates, and it uses hierarchical clustering algorithm.

3 Proposed methodology

The present method initially clusters the datasets using firefly algorithm to form base cluster. A multi-viewpoint-based similarity measure obtains optimal partitioning by finding the specific distance function between the clustered datasets. It then uses cluster ensemble method to create clustered categorical datasets. The data partitioning is generated using ensemble even in the presence of incomplete information. The refined matrix or cluster association is used to find the value of 0 or unknown associations over 1 or known associations from the binary matrix. The graph partitioning problem in clusters is resolved using bipartite graph model. The entropy-based weighted link quality is used to acquire the important knowledge in the network. This outputs the finally improved clustered results even in the presence of incomplete or unknown data points in the network. This bridges the gap between link analysis and clustering and enhances the clustering over categorical datasets. The architecture of clustering ensemble method is shown in Fig. 1.

3.1 FFA data clustering

3.1.1 Clustering problem

Clustering is used to find the inter-collections and intra-relations of data samples. Clustering is used to reduce the data volume and dimensions of data for preprocessing

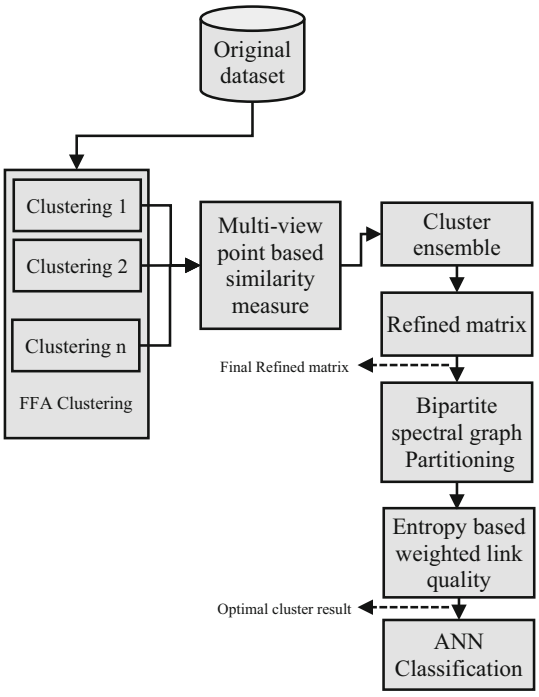


Fig. 1 Overall diagram for proposed methodology

and similarity or dissimilarity analysis. Many standards are considered to solve the problems in clustering. Among which, the Euclidean distance metric plays a major role in the process of clustering and it is used to find the distance between the cluster center and the data points. The Euclidean distance metric is used in clustering algorithm:

$$d(X, c_j) = \sqrt{\sum_{i=1}^{m} (x_i - c_j)^2} \tag{1}$$

where m—total number of data, X—data vector, and c_j—cluster center.

Consider a sample with points in N-dimensional space, and these points are assigned in each cluster using Euclidean distance standards. The main aim of the FFA is to obtain local optimal solution by finding the cluster center. However, it fails at solving the clustering problem due to its dependency with final output on an initial value of centers and new convergence and being in a local optimum.

3.1.2 Firefly algorithm for data clustering

The firefly algorithm is considered as a stochastic global optimization method, and the hybrid clustering is based on firefly HFA and that make use of distance metric. The objective function $f(x)$ uses centroid cluster difference between one and other data points, and finally the best data point is chosen [23]. The fireflies travel to a best location of a one data point within a cluster to other data point. The optimal data points are obtained using new location and the best centroid value is found in the similar data point group and classification arises without any degradation in quality of clusters.

The main problems in FFA are finding the light intensity difference and absorbency amount. We consider that luminosity with absorbency depends mainly on target function. Since absorbency in FFA is directly related to light intensity of adjacent firefly, the absorbency (β) is thus defined as:

$$\beta(r) = \beta_0 e^{-\gamma r^2}$$

where β_0 is regarded as the absorbency that shows total amount of absorption from the light source when $r = 0$. The parameter γ determines the convergence speed and firefly behavior, which is determined based on changes in absorbance within [0,∞) limits.

The distance between fireflies is given by Cartesian distance relation:

$$r_{ij} = \|x_i - x_j\| = \sqrt{\sum_{k=1}^{d} \left(x_{ik} - x_{jk}\right)^2}$$

where k—part of spatial coordinates of firefly in 2D state, which is given by:

$$r_{ij} = \sqrt{\left(x_i - x_j\right)^2 + \left(y_i - y_j\right)^2}$$

The movement of firefly and absorption is determined as:

$$x_i = x_i + \beta(r)\left(x_i - x_j\right) + \alpha(\text{rand} - 0.5)$$

From the above equation, the absorbency is given as $\beta(r)\left(x_i - x_j\right)$ and α is the chance maker parameter and rand $\in [0,1]$ is the random number obtained through uniform distribution. The chance maker is the third statement, which is given by $\alpha(\text{rand} - 0.5)$.

The fitness function to evaluate the problem of clustering using distance-based measurement is defined in the following equation.

$$f = \frac{1}{M}\sum_{i=1}^{M} \| D_i - C_i \|$$

where C_i—cluster center, D_i—total number of available data and M—maximum distances between the cluster centers.

Algorithm data clustering using HFFA	
1:	Initialize the objective function $f(x_i)$, $X = (X_1, X_2, X_3, \ldots, X_n)$
2:	Generate an initial population for suitability index with fireflies X_i $(i = 1, 2, 3, \ldots, n)$
3:	Set Max number of $iterations = Maxgenerations$, $Sett = 1$
4:	Devise light intensity I so that it is linked with $f(x_i)$,
5:	Define absorption coefficient λ and β
6:	While $(t < MaxGeneration)$
7:	For $i = 1$: n (for all n fireflies)
8:	For $j = 1$: n (n fireflies)
10:	If $(I_j > I_i)$
11:	Move firefly i toward j;
12:	$X_j^{t=1} = X_j^t + \beta \exp[-\gamma r_{ij}^2]$
13:	End if
14:	Vary attractiveness with distance r_{ij} via update the cluster centroids
15:	Evaluate new solutions and update light intensity for each firefly;
16:	End for j;
17:	End for i;
18:	Rank the fireflies and finding best cluster data points;
19:	End while;
20:	Post-processing and location visualization;
21:	End process.

3.2 Multi-viewpoint-based similarity measure

The multi-viewpoint similarity measure is proposed to improve the clustering quality, and the problem in such clustering quality is improved with optimization process.

Here, an optimal partitioning method is obtained using estimation of specific distance function between the datasets. The efficiency of clustering depends mainly on similarity measurement between the datasets.

The new concept of similarity measurement is constructed using one or more data points as reference, and the similarity measurement between the datasets is called as multi-viewpoint similarity measurement, which is given as:

$$\mathrm{Sim}_{d_i,d_j\in S_r}(d_i,d_j)=\frac{1}{n-n_r}\sum_{d_h\in\frac{S}{S_r}}\mathrm{Sim}(d_i-d_h,d_j-d_h) \tag{2}$$

From Eq. (2), the similarity measurement between the documents, d_i and d_j, in same cluster is defined as the average similarity obtained from the different view of all documents that lies outside the cluster. The measuring of similarity between the documents should lie completely within a cluster, and the points for establishing such measures should lie outside the cluster. This method is called as MVS [16], which is applied in our method, and the MVS between the d_i and d_j is given by $\mathrm{MVS}(d_i,d_j|d_i,d_j\in S_r)$. The MVS final product depends mainly on individual similarity within the sum, and the similarity is given by the dot product of the document vectors,

$$\begin{aligned}\mathrm{MVS}((d_i,d_j|d_i,d_j)\in S_r)&=\sum_{d_h\in\frac{S}{S_r}}(d_i-d_h)^t(d_j-d_h)\\&=\frac{1}{n-n_r}\sum_{d_h}\cos(d_i-d_h,d_j-d_h)||d_i-d_h||||(d_j-d_h)||\end{aligned} \tag{3}$$

The similarity measurement between d_i and d_j within a cluster S_r over point d_h that lies outside the cluster is considered as a dot product of correlation distance over d_h and cosine angle between the documents from the point d_h.

3.3 Cluster ensemble

The database over given clustered categorical dataset is created, and the formation of base cluster is obtained using FFA algorithm. The classical ensemble problem is avoided with the use of consensus clustering, specifically for unsupervised learning in analogous with ensemble approach for supervised methods. This refers to the criteria that a number of varied input clusters are obtained over specific dataset and a single consensus clusters are found from the base clusters. The reconciliation of clustering is obtained from consensus clustering obtained from similar dataset of different sources or different iteration of algorithms. It uses three methods:

1. *Direct ensemble* is the categorical data points which are transformed directly to form ensemble without the use of base clusters. This is the first type of cluster ensemble, which is used to transform the categorical data clustering problem using cluster ensembles. It considers the categorical label as a cluster in an ensemble. With such formalism, the categorical data are transformed directly into cluster ensemble without base clusters. This direct ensemble method provides high-quality

clustering data, where the accuracy of single attribute partition is lesser than clustering the entire data attributes, which enables larger diversity inside a cluster ensemble.

2. *Full-space ensemble* is obtained from base clusters (obtaining clustering using a basic clustering algorithm over original categorical dataset). The base clustering is generated using FFA with cluster center initialization. The artificial instability to FFA is generated using two schemes, namely fixed and random modes that select the cluster from base clustering.
3. *Subspace ensemble* is the high diverse cluster ensemble over different data subsets. The cluster ensemble is generated from data subspace through base clustering results from FFA. The cluster ensemble is created using FFA from various subspace attributes set using fixed and random schemes for selecting the clusters.

When the dataset is raw, significance in finding the processed information is limited and only semantic relations provides the proper meaning. The proposed method aims at extracting the information from the dataset in an automated way.

3.3.1 Creation of the ensemble E^*

The proposed method creates candidate ensemble during each iteration with the help of one net over each M formed cluster. The data point size is seen as varying ensemble during the process of clustering due to formation of different clusters during each iteration. A maximum average distance is exhibited over other chosen clusters. Here, the distance between the dataset and a cluster is calculated as:

$$\forall E_i, E_j i \neq jd\left(E_i, E_j\right) = \max_{n_{s\in E_i}, n_t \in E_j} \{d(n_s, n_t)\} \tag{4}$$

The datasets (M) are combined with individual data point ensemble E^* using grouping and improved classification accuracy; hence, it is calculated using dataset validation. The ensemble performance is thus created using HFFA and provides better performance. So, even in worst-case probability, the ensemble data point N is created during clustering and the proposed method hence exhibits a very limited computational complexity.

3.4 Generating refined matrix

The refined matrix or cluster association is used to find the value of 0 or unknown associations over 1 or known associations. Here, the degree of association is preserved using refined matrix. Further, the similarity using binary cluster association matrix between the clusters belonging to [0,1] signifies that the similarity exists between two different clusters.

With such insight, the refined matrix is defined as enhanced variations from original binary matrix. It aims at approximating the unknown association value ("0") from known value ("1"). The preservation of association degree is within the refined matrix limits, i.e., $\mathrm{BM}(x_i, cl) = 1 \rightarrow \mathrm{RM}(x_i, cl) = 1$. For each cluster $\pi_t, t = 1, 2, \ldots, M$,

the corresponding cluster is given by $C_1^t, C_2^t, \ldots, C_{k_t}^t$ and its association degree with total clusters k_t is given in terms of refined matrix $\mathrm{RM}(x_i, cl) \in [0, 1]$ that consists of data points $x_i \in X$ with $cl \in \left\{C_1^t, C_2^t, \ldots, C_{k_t}^t\right\}$ is given as,

$$\mathrm{RM}(x_i, cl) = \begin{cases} 1 & \text{if } cl = C_*^t(x_i) \\ \mathrm{Sim}(cl, C_*^t(x_i)) & \text{otherwise} \end{cases}$$

where $C_*^t(x_i)$ is considered as cluster label over cluster t to which x_i belongs. Additionally, the similarity between the two clusters C_x and C_y is given by $\mathrm{Sim}(C_x.C_y) \in [0, 1]$, which is obtained using weighted triple quality link-based ensemble.

3.5 Resolving graph partitioning problem in clustering using bipartite spectral graph

Here, the datasets are applied over various clustering technique and it provides different output, which is chosen as base clusters. This base cluster is linked with link similarity technique, which produces final clustering with spectral graph partition. The final clusters are generated based on complete information ensemble. Hence, the categorical datasets are clustered with better quality using the above procedure and entities are further included in the finalized cluster. Higher accuracy is attained using link-based spectral graph partition in an inexpensive manner. This link approach has the ability to find the null or unknown values. To further improve the accuracy, bipartite graph partition is used with link similarity.

Finding the suitable partition set Π of X is a difficult problem. The total number of different partition sets of the form $\Pi = \{C_1, C_2, \ldots, C_k\}$ may be expressed by the function $S(N, k)$ associated with the Stirling number of the second kind, which is given by:

$$S(N, k) = \frac{1}{k!} \sum_{i=0}^{k} (-1)^{k-i} \binom{k}{i} i^N$$

where $N = |X|$. For instance, for $N = 50$ and $k = 2$, $S(N, k) \approx 5.63 \times 1014$. This number illustrates the complexity of the problem that we need to solve. Therefore, it is necessary to resort to a method that allows us to explore efficiently a large solution space. In the following section, we briefly describe some meta-heuristic searches.

3.5.1 Bipartite graph model

A graph $G = (V, E)$ with V vertex set and E edge weight set and E_{ij} edge weight between the vertex $\{i, j\}$.

Definition 1 The $n \times n$ weighted adjacency matrix of an undirected graph is defined as the matrix (m_{ij}) $i, j = 1, \ldots, n$. If $m_{ij} = 0$, it implies that vertices v_i and v_j are not connected by an edge. If $m_i, j \neq 0$, it implies that the vertices $\{i, j\}$ are connected and $m_{i,j}$ is the corresponding edge weight. Since the graph is undirected, $m_{ij} = m_{ji}$ necessarily.

Definition 2 Given the weighted adjacency matrix of a graph and a partition of the vertex set V into two disjoint subsets V_1 and V_2, the cut between these two subsets is defined as:

$$\text{cut}(V_1, V_2) = \sum_{i \in V_i, j \in V_2} M_{ij}. \tag{5}$$

The objects in the dataset expressed in the form of graphs using bipartite graphs. Here, the undirected bipartite graph used is a triplet

$$G = (D, W, E).$$

where

$$W = (w_1, w_2, \ldots, w_m)$$
$$D = (d_1, d_2, \ldots, d_n)$$

This represents the sets of vertices with E as the edge set $\{\{w_i,\ d_j\} :\ w_i \in W, d_j \in D\}$.

Here, D is the set of datasets and W is the category set in dataset. The edge (w_i, d_j) occurs if w_i lies in dataset d_j. These edges are considered undirected, and the bipartite model has nil edges between the datasets and the categories. The problem in graph partition is NP complete problem, and this is solved with local minima using heuristic method, where the search occurs in local vicinity during initial partitioning and the solution stuck at local minima.

3.5.2 Spectral graph bipartitioning

The graph is formulated using spectral graph partitioning, which differ based on best partition search using spectral partitioning heuristic. If $G = (V, E)$ has m edges and n vertices, then the incidence matrix for G is I_G, which is $n \times m$ matrix with one row and column for a vertex and an edge, respectively. The value of column w.r.t. edge $\{i, j\}$ of I_G is 0 other than ith and jth entry, and they are $\sqrt{E_{ij}}$ and $-\sqrt{E_{ij}}$, respectively. A spectral graph bipartitioning is trained in within measured datasets, and the classifier is evaluated over different test problems that show considerable developments for better clustering over the datasets.

3.6 Weighted triple quality link clustering ensemble

Using cluster ensemble, the weighted graph is constructed $g = (v, w)$ where v—vertices set representing each cluster and w—weighted edge set between the clusters. Then, weights over each edge, $w_{de} \in w$, which gets connected to clusters of c_d and $c_e \in$ v. This is estimated finally by the overlapping members proportion.

$$w_{de} = |L_d \cap L_e| / |L_d \cup L_e|$$

The value lies in [0,1]. The value is 0 if two data points are considered disjoint, and the value is 1 if two data points are equal. This is a common indicator for measuring the similarity between the two dataset clusters. If w_{xy} is nearer to 1, the clusters are dissimilar and if w_{de} is nearer to 0, the clusters are similar.

Hence, total duplicated datasets are removed from the clusters. This is calculated based on hash value and that uses Division Remainder Hashing (DRH) over two datasets L_d and L_e. In DRH, L_d key is divided by a number greater than the total L_d keys, and the remainder value is regarded as the index over the hash table, i.e.,

$$H(L_d) = L_d \bmod m$$

The number is chosen as prime number and this reduces the total collisions. The hash function over each cluster maps the keys and that lies between 0 and $m-1$. Here, the hash lies between 1 and m rather than 0 to $m-1$.

$$H(L_d) = L_d \bmod m + 1$$

H—hash function that maps x and y to distinct clusters. In each cluster the value of $h_{\min}(S)$ belongs to the value x, which has minimum $h(x)$ values. Then, $h_{\min}(L_d) = h_{\min}(L_e)$ when the minimum H of the $L_d \cup L_e$ lies in $L_d \cap L_e$. Similarity between the dataset is found using weighted quality matrix (WTQ).

The similarity is thus measured using:

$$s(a,b) = \sum_{\forall z \in Z} \log^{-1}[\text{frequency}(z)]$$

where z—feature set between a and b and frequency(z) defines the total appearance of z. The clustering quality is determined by weighted parity that connects the clusters. The WTQ measure $c_d, c_e \in v$ w.r.t. to $c_k \in v$, which is estimated as:

$$\text{WTQ}_k = 1/w_k$$

Here,

$$W_k = \sum (\forall t \in N_k \subset V) w_{tk},$$

where N_k is the clusters set linked with the cluster c_k, where $\forall c_t \in N_k$, $w_{tk} \in w$.

The WTQ score between the clusters datasets c_d and c_e is defined as

$$\text{WTQ}(de,k) = \sum_{k=1-q} \text{WTQ}(de,k)$$

Following that, the similarity between clusters c_d and c_e can be estimated by:

$$\text{Sim}(c_d, c_e) = (\text{WTQ}_{de})/(\text{WTQ}_{\max}) \times \text{DC}$$

where WTQ_{max}—maximum WTQ value, WTQ_{de}—WTQ between the clusters. Here, $c_d, c_e \in v$ and DC—constant decay factor = [0,1].

3.6.1 Entropy-based weighted triple quality

Once the similarity measurement between the data points is found, the entropy is increased and possesses certain correction based on clustering similarity. Let c_j be the cluster with elements represented as $CL = \{x_1, x_2, \ldots, x_n\}$ with cluster center $CC(t) = x$, where $x \notin (x_1 + x_2 + \cdots + x_n)/m$. Hence, the centers is obtained as:

$$CC^{(t+r)} = \frac{\sum_{i=1}^{m} x_i}{(m+1)} \frac{1 - \left(\frac{1}{m+1}\right)^r}{1 - \left(\frac{1}{m+1}\right)} + \frac{x}{(m+1)^r}$$

The equations are in geometric progression, and hence, to obtain entropy, the above equation can be converged as:

$$\frac{CC^{(t+r)}}{n} = \frac{\sum_{i=1}^{m} x_i}{m^r}$$

The present work finds the value entropy from the WTQ and RM is used to generate a bipartite graph, where the construction of RM is constructed using similar data points. Here, the categorical label similarity helps to resolve the RM construction problem. Thus, weights are calculated with WTQ for graph partitioning that improves the quality of clustering.

3.7 ANN classification over optimized clustered datasets

Based on multi-viewpoint similarity, the clustered data points are classified using ANN and generates initial data points ensemble. Here, ensemble (E) of ANN (N) using classification is given as:

$$E = \{n_1, n_2, \ldots, n_N\}$$

The selection phase for the selection of subset of E^* is combined effectively. The optimal subset E^* is formed with assess over accuracy in classification by subset, E. Thus, the total subsets (or possible subset) are,

$$\sum_{i=1}^{N} \binom{N}{i}$$

Here, the size (N) of set (E) is considered large during the phase of classification, and hence, the brute force method is not possible. The selection phase depends on hypothesis of set E that allows the subset selection from the dataset and it is combined without the need of enumeration from the possible subsets. Here, the set E is developed by forming union of subset (M) with E_i:

Table 1 Datasets

Dataset	N	D	AV	K
Zoo	101	16	36	7
Lymphography	148	18	59	4
Primary tumor	339	17	42	22
Breast cancer	683	9	89	2
20 Newsgroup	1000	6084	12,168	2

$$E = \bigcup_{i=1}^{M} E_i$$

E_i meets: $\forall i,\ ji \neq jE_i \cap E_j = \emptyset$ and the dataset that forms the subset satisfies:

$$\forall E_i, E_j, n_l, n_m, n_n, \quad i \neq j, n_l, n_m \in E_i, n_n \in E_j \ \text{prob}(n_l\text{fails}, n_m\text{fails}) > \text{prob}(n_l\text{fails}, n_n\text{fails})$$

The term prob($n_l$fails, $n_m$fails) and prob($n_l$fails, $n_n$fails) is the probability of compound error from the datasets. From Eq. (8), the probability of compound error between the two different datasets of similar subset is found higher than the subset between the two nets of different subsets.

This condition provides a better guide for selecting the dataset ensemble that are combined and classified. The selection phase based on Eqs. (4–7) is given as:

- Subset Identification E_i;
- Nets extraction from the subsets for creating the ensemble $E^* = \{n_1^*, n_2^*, \ldots, n_M^*\}$ using error-independent datasets.

4 Results

The proposed hybrid firefly algorithm linked clustering ensemble framework or HFALCEF is evaluated over real datasets. The improved quality in data partition is formed using the proposed method, which is evaluated and assessed with other ensemble and categorical data clustering methods. The datasets considered for evaluations include: Zoo, lymphographic, primary tumor, breast cancer and 20 newsgroup [18], shown in Table 1. The parameters of the datasets include: data points (N), attributes (d), attribute values (AV) and classes (K). For example, 20 newsgroup datasets have 1000 documents obtained from 2 newsgroups with 6084 different terms. Specifically, the frequency ($f \in \{0, 1, \ldots, \infty\}$) at which the keyword appears over the documents is converted to a nominal value: "Yes" if $f > 0$, "No" otherwise.

The quality of the clustering ensemble is evaluated first, and then, it is compared empirically and this employs the following clustering ensemble. Here, five major ensemble factors have been taken for the study and that include: Factor I, Factor II (Fixed k), Factor II (Random k), Factor III (Fixed k), and Factor III (Random k).

The clustering algorithm of k-mode factor generates base clustering using HFALCEF, FALCEF, link-based cluster ensemble (LCE), similarity matrix (CO) with single linkage (CO + SL), similarity matrix (CO) with average linkage (CO + AL), cluster-based similarity partitioning algorithm (CSPA) and hyper-graph partitioning algorithm (HGPA). The classification accuracy between the different ensemble factors over the categorical dataset is shown in Table 2. The result shows that proposed method obtains higher classification accuracy than other ensemble methods for all datasets. The result further shows that III-Random k obtains higher classification accuracy than III-Fixed k, II-Random k, II-Fixed k and I.

The frequency of significant better (B) performance is shown in Table 3, and they are categorized as per the following indices: Normalized Mutual Information (NMI), Adjusted Rand (AR) and Classification Accuracy (CA). From the results, it is seen that the effectiveness of the proposed method has enhanced than the other clustering ensemble methods, where B and W represent the average number of time the technique attains the *better* and *worse* results, respectively. The Betterness (B) results is high in CA, NMI and AR for the proposed method than FALCEF, LCE, CO + SL, CO + AL, CSPA and HGPA. On the other hand, the Worseness (W) results for the proposed method is lesser than other ensemble methods like FALCEF, LCE, CO + SL, CO + AL, CSPA and HGPA.

The parametric analysis focuses mainly on presenting the means of employing HFALCEF at its greater potential. The performance of such technique depends mainly on the decay factor DC $\in$ [0, 1], which is deployed in order of ascertaining the cluster similarity.

The relationship thus exists between the DC $\in \{0.1, 0.2, , \ldots, 0.9\}$ and the clustering performance values in X- and Y-axis, respectively. The HFALCEF average score value of different ensemble factors depends on DC $\in \{0.1, 0.2, \ldots, 0.9\}$, as shown in Figs. 2, 3, 4, 5 and 6. The HFALCEF method attains higher quality consistency and results of DC value between 0.7 and 0.9, which is due to automated calculation of centroid value using HFA and this has improved the clustering results for DC $\in \{0.1, 0.2, \ldots, 0.9\}$.

Hence, the average score value of HFALCEF over different ensemble factors based on size $M \in (10, 40, \ldots, 100)$ is shown in Figs. 7, 8, 9 and 10. From the results, it is seen that the HFALCEF method attains a good average score results than LCE, CO-SL, CO-AL, CSPA and HGPA.

Cohesion relates the closeness of the data in the cluster, which are related to each other clusters [20]. The cohesion between each cluster is:

$$p_i = \frac{\sum_{m \neq n} C_{mn}}{K * \frac{(K-1)}{2}} \tag{6}$$

where m, n—cluster data point, K is cluster data point, C_{mn}—data points distances.

From Table 4, the cohesion parameter shows the datasets and the performance of parameter measurement between the datasets is shown in Fig. 11.

Table 2 Classification accuracy of different clustering methods

Dataset	Ensemble factor	HFALCEF	FALCEF	LCE	CO + SL	CO + AL	CSPA	HGPA
Zoo	I	0.9452	0.932	0.894	0.883	0.889	0.801	0.594
	II-Fixed *k*	0.9532	0.941	0.921	0.882	0.899	0.824	0.836
	II-Random *k*	0.9583	0.942	0.931	0.861	0.873	0.824	0.826
	III-Fixed *k*	0.9572	0.954	0.941	0.891	0.916	0.823	0.831
	III-Random *k*	0.9593	0.946	0.931	0.862	0.873	0.825	0.835
Lymphography	I	0.8063	0.7569	0.743	0.572	0.582	0.733	0.660
	II-Fixed *k*	0.8341	0.812	0.796	0.558	0.723	0.742	0.709
	II-Random *k*	0.8549	0.823	0.784	0.556	0.673	0.690	0.633
	III-Fixed *k*	0.8314	0.795	0.782	0.534	0.702	0.718	0.692
	III-Random *k*	0.8419	0.7925	0.777	0.553	0.662	0.673	0.633
Primary tumor	I	0.7253	0.689	0.445	0.273	0.381	0.445	0.293
	II-Fixed *k*	0.7328	0.645	0.478	0.308	0.447	0.450	0.438
	II-Random *k*	0.7418	0.694	0.496	0.298	0.428	0.445	0.428
	III-Fixed *k*	0.7354	0.6945	0.463	0.295	0.443	0.452	0.423
	III-Random *k*	0.7621	0.6936	0.484	0.294	0.413	0.435	0.408
Breast cancer	I	0.9621	0.956	0.94	0.652	0.653	0.849	0.673
	II-Fixed *k*	0.9892	0.9856	0.972	0.653	0.954	0.836	0.873
	II-Random *k*	0.9892	0.9863	0.973	0.653	0.943	0.812	0.870
	III-Fixed *k*	0.9867	0.9835	0.969	0.649	0.952	0.831	0.849
	III-Random *k*	0.9813	0.9756	0.964	0.653	0.932	0.812	0.844
20 Newsgroup	I	0.8214	0.789	0.61	0.6	0.6	0.6	0.565
	II-Fixed *k*	0.8418	0.831	0.782	0.6	0.6	0.6	0.6
	II-Random *k*	0.8514	0.783	0.72	0.6	0.6	0.6	0.6
	III-Fixed *k*	0.8596	0.823	0.726	0.6	0.6	0.6	0.6
	III-Random *k*	0.8623	0.818	0.6	0.6	0.6	0.6	0.6

Table 3 Performance comparison among examined clustering methods

Ensemble factor	Methods	CA		NMI		AR	
		B	*W*	*B*	*W*	*B*	*W*
I	HFALCEF	193	22	153	58	169	58
	FALCEF	187	28	146	53	164	53
	LCE	170	35	137	65	149	61
	CO + SL	34	180	81	136	47	163
	CO + AL	72	143	114	93	84	120
	CSPA	105	92	132	82	109	95
	HGPA	21	193	19	208	24	201
II-Fixed *k*	HFALCEF	238	2	232	4	226	08
	FALCEF	225	3	228	5	212	09
	LCE	208	4	204	9	201	13
	CO + SL	26	171	37	149	28	163
	CO + AL	131	46	134	42	134	39
	CSPA	86	91	68	101	82	103
	HGPA	64	80	93	92	93	91
II-Random *k*	HFALCEF	243	2	221	6	228	06
	FALCEF	235	2	214	8	219	08
	LCE	209	4	203	12	201	11
	CO + SL	17	173	38	159	28	170
	CO + AL	97	80	94	63	117	53
	CSPA	76	113	47	131	51	125
	HGPA	67	121	41	139	49	132
III-Fixed *k*	HFALCEF	225	5	216	06	208	08
	FALCEF	219	5	203	08	197	09
	LCE	197	7	191	13	182	16
	CO + SL	17	167	37	142	32	162
	CO + AL	115	43	119	43	139	29
	CSPA	73	87	53	94	61	106
	HGPA	71	82	56	101	65	101
III-Random *k*	HFALCEF	232	2	218	08	211	08
	FALCEF	227	3	206	09	196	09
	LCE	203	6	191	13	181	11
	CO + SL	15	178	34	146	32	163
	CO + AL	81	67	81	59	103	43
	CSPA	52	112	36	121	45	123
	HGPA	51	117	38	134	63	131

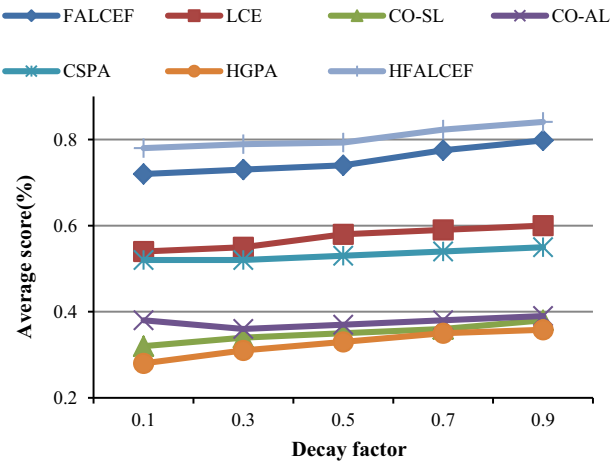


Fig. 2 Ensemble Factor I versus decay factor

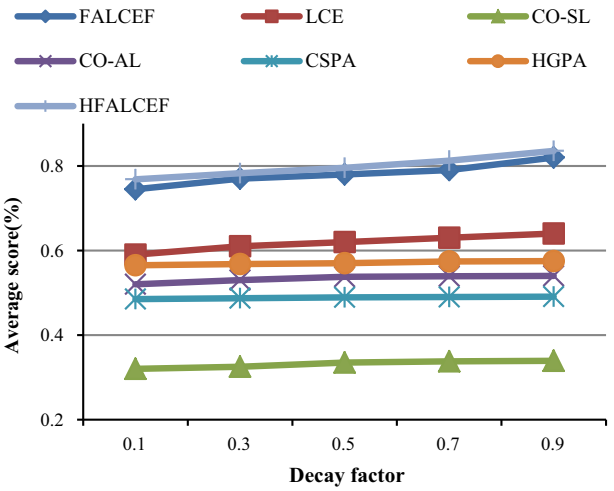


Fig. 3 Ensemble Factor II-Fixed k versus decay factor

Variance shows the separation of clusters, which is given by:

$$V_i = \sum_{m \neq n} \frac{(p_i - C_{mn})^2}{\left(\frac{K(K-1)}{2}\right) - 1} \tag{7}$$

where p_i—cluster cohesion value, C_{mn}—data points distance, K—number of data points in the clusters.

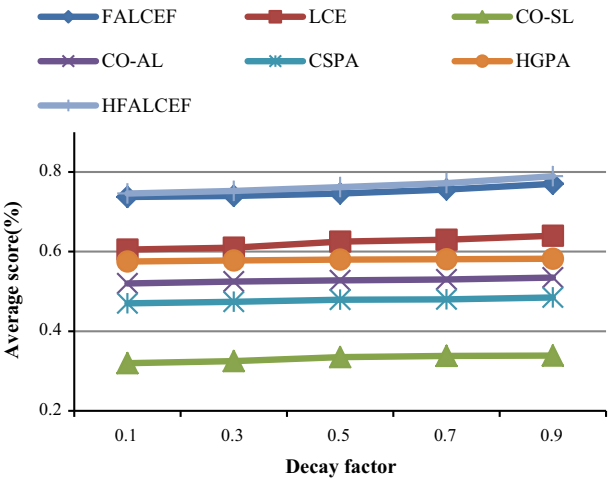


Fig. 4 Ensemble Factor II-Random k versus decay factor

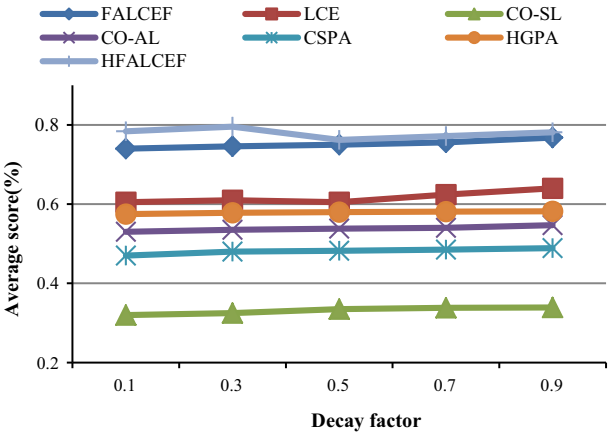


Fig. 5 Ensemble Factor III-Fixed k versus decay factor

From Table 5, variance measurement for the datasets and comparison between the proposed and other system is shown in Fig. 12.

Precision is defined as fraction of relevant retrieved datasets or positive predictive value. This is determined in terms of retrieval of true positive (TP) and false positive (FP) values, which is given by:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{8}$$

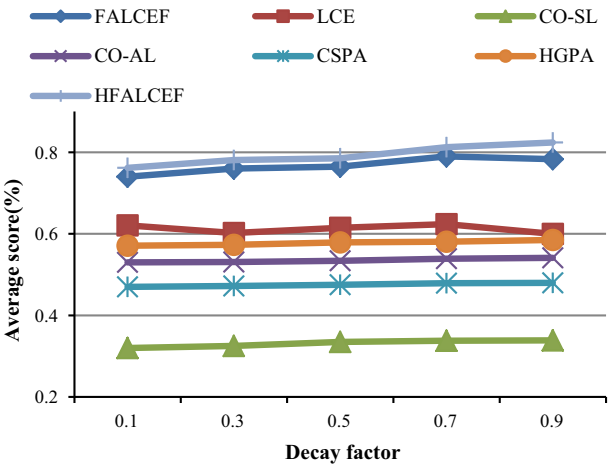


Fig. 6 Ensemble Factor III-Random k versus decay factor

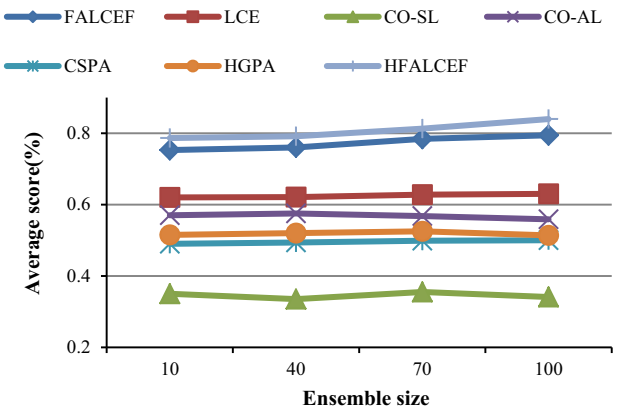


Fig. 7 Ensemble Factor II-Fixed k versus ensemble size

From Table 6, precision measurement for the datasets and comparison between the proposed and other system is shown in Fig. 13.

Recall is defined as the fraction of TP and the sum of retrieval TP and false positive (FP), which is given as:

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FP}) \tag{9}$$

From Table 7, recall measurement for the datasets and comparison between the proposed and other system is shown in Fig. 14.

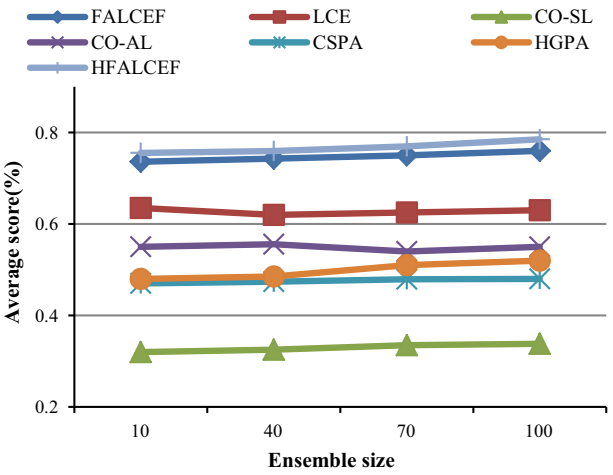


Fig. 8 Ensemble Factor II-Random k versus ensemble size

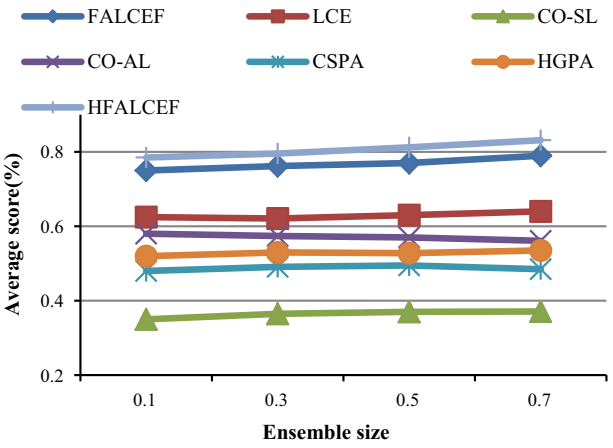


Fig. 9 Ensemble Factor III-Fixed k versus ensemble size

5 Conclusion and future works

This paper presents a high effective novel clustering ensemble method. The clustering ensemble method transforms the clustered categorical data matrix into refined matrix by direct application of graph partitioning technique. The optimal clustered result from the above ensemble methods is finally classified using ANN algorithm. This method resolves the degradation issues during cluster formation. The degree of relationship between the solutions of generated base clusters is explored. The result shows that HFFA obtains better clustered results over UCI datasets. The problem in constructing the refined matrix is resolved effectively using

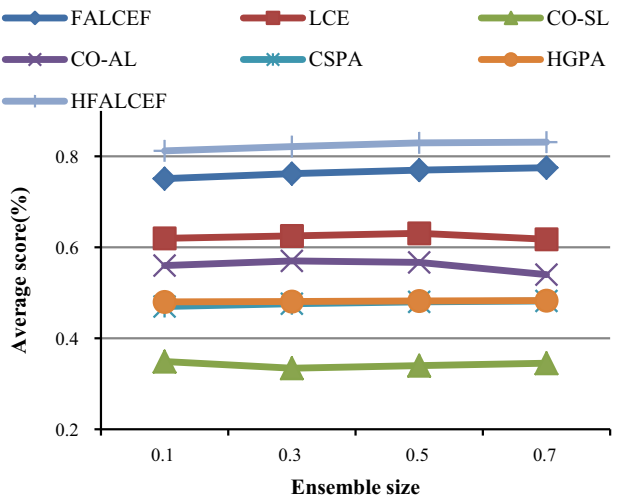


Fig. 10 Ensemble Factor III-Random *k* versus ensemble size

Table 4 Comparison of the performance in cohesion clustering method for datasets

Dataset	HFALCEF	FALCEF	LCE	CO + SL	CO + AL	CSPA	HGPA
Zoo	0.812	0.7818	0.7512	0.7315	0.7218	0.7185	0.7083
Lymphography	0.8105	0.7815	0.7618	0.7481	0.7321	0.7258	0.7156
Primary tumor	0.8361	0.8153	0.7952	0.7712	0.7589	0.7328	0.7265
Breast cancer	0.8454	0.8258	0.8021	0.7829	0.7682	0.7581	0.7337
20 Newsgroup	0.8685	0.8418	0.8206	0.8102	0.7819	0.7663	0.7418

Table 5 Variance performance comparison for datasets

Datasets	HFALCEF	FALCEF	LCE	CO + SL	CO + AL	CSPA	HGPA
Zoo	0.732	0.7128	0.7024	0.6958	0.6817	0.6638	0.6582
Lymphography	0.7481	0.70819	0.6921	0.6886	0.6785	0.6571	0.6423
Primary tumor	0.7563	0.7236	0.7181	0.7078	0.6889	0.6614	0.6582
Breast cancer	0.7281	0.7176	0.7048	0.6912	0.6837	0.6451	0.6386
20 Newsgroup	0.7894	0.7618	0.7486	0.7318	0.7263	0.7158	0.7043

WTQ similarity algorithm. The experimental results with various ensemble types, datasets and validity measures suggest that proposed link-based clustering ensemble obtains superior results than conventional cluster ensemble and categorical data algorithms.

Additionally, this work can further be extended on similarity measurement using fuzzy clustering ensemble, subspace clustering ensemble and other optimization ensemble algorithms like bee colony, harmony search, genetic and fuzzy-based methods over categorical, real and synthetic datasets with huge dimensionality. Also,

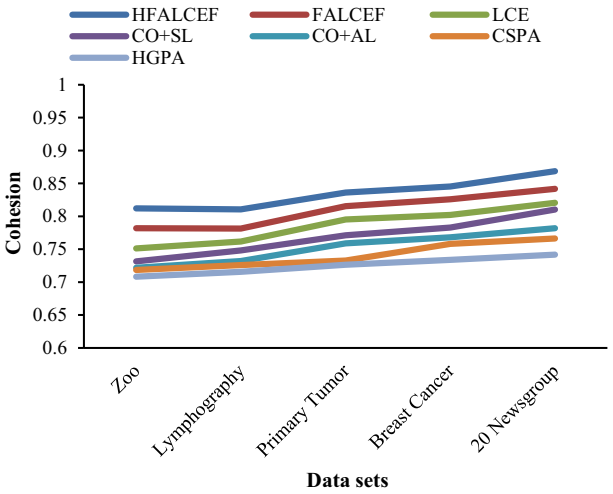


Fig. 11 Cohesion comparison versus datasets

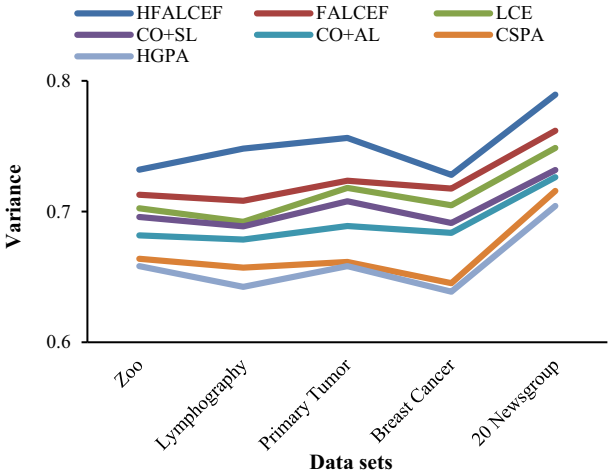


Fig. 12 Variance performance comparison versus datasets

Table 6 Precision comparison for datasets

Datasets	HFALCEF	FALCEF	LCE	CO + SL	CO + AL	CSPA	HGPA
Zoo	0.752	0.7328	0.7258	0.7136	0.7095	0.7018	0.6925
Lymphography	0.7681	0.7485	0.7335	0.7481	0.7423	0.7292	0.7184
Primary tumor	0.7854	0.7682	0.7437	0.7355	0.7256	0.7155	0.7082
Breast cancer	0.7719	0.7564	0.7352	0.7246	0.7185	0.7059	0.6949
20 Newsgroup	0.7987	0.7819	0.7628	0.7468	0.7329	0.7257	0.7085

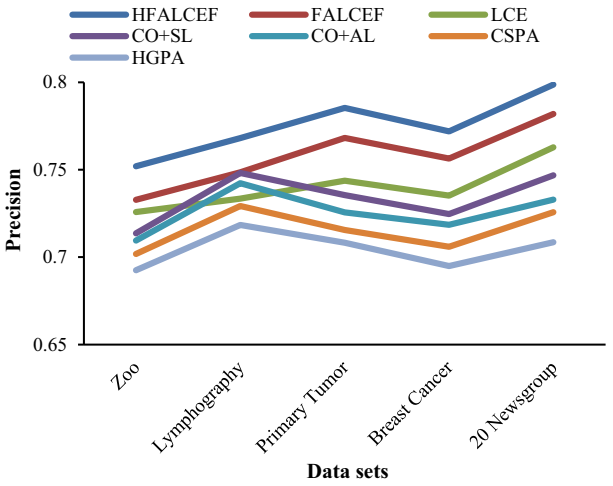


Fig. 13 Precision comparison versus datasets

Table 7 Recall comparison for datasets

Datasets	HFALCEF	FALCEF	LCE	CO + SL	CO + AL	CSPA	HGPA
Zoo	0.7953	0.7818	0.7702	0.7652	0.7581	0.73214	0.7289
Lymphography	0.8128	0.7956	0.7856	0.7669	0.7463	0.7388	0.7213
Primary tumor	0.7925	0.7782	0.7638	0.7518	0.7463	0.7413	0.7352
Breast cancer	0.8235	0.8102	0.7941	0.7852	0.7641	0.7518	0.7381
20 Newsgroup	0.8102	0.7980	0.7851	0.7743	0.7518	0.7456	0.7258

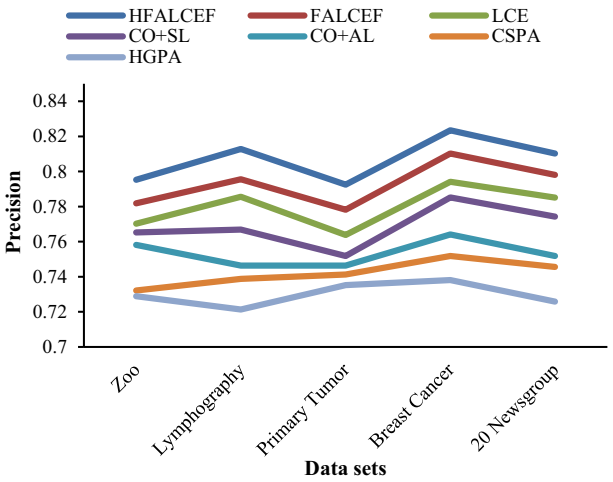


Fig. 14 Recall comparison for datasets

improvements over the execution time can be reduced with increased clustering efficiency. Further, the research can be improved on other ensemble technique with the use of linked similarity measures using multi-viewpoint similarity measurement.

References

1. Andritsos P, Tsaparas P, Miller RJ, Sevcik KC (2004) LIMBO: scalable clustering of categorical data. In: EDBT, pp 123–146
2. Bai L, Liang J (2014) The k-modes factor clustering plus between-cluster information for categorical data. Neurocomputing 133:111–121
3. Billard L, Diday E (2003) From the statistics of data to the statistics of knowledge: symbolic data analysis. J Am Stat Assoc 98(462):470–487
4. Chen K, Liu L (2005) The best K for entropy-based categorical data clustering. Georgia Institute of Technology, Atlanta, pp 1–11
5. Dhillon IS (2001) Co-clustering documents and words using bipartite spectral graph partitioning. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM
6. Ganti V, Gehrke J, Ramakrishnan R (1999) CACTUS—clustering categorical data using summaries. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp 73–83
7. Gibson D, Kleinberg J, Raghavan P (2000) Clustering categorical data: an approach based on dynamical systems. VLDB J 8(3–4):222–236
8. Grabmeier J, Rudolph A (2002) Techniques of cluster algorithms in data mining. Data Min Knowl Discov 6(4):303–360
9. Guha S, Rastogi R, Shim K (1999) ROCK: a robust clustering algorithm for categorical attributes. In: 15th IEEE International Conference Proceedings on Data Engineering, pp 512–521
10. Huang Z (1998) Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min Knowl Discov 2(3):283–304
11. Iam-on N, Garrett S (2010) LinkCluE: a MATLAB package for link-based cluster ensembles. J Stat Softw 36(9):1–36
12. Iam-On N, Boongeon T, Garrett S, Price C (2012) A link-based cluster ensemble approach for categorical data clustering. IEEE Trans Knowl Data Eng 24(3):413–425
13. Jain AK (2010) Data clustering: 50 years beyond K-means. Pattern Recognit Lett 31(8):651–666
14. Kuncheva LI, Hadjitodorov ST (2004) Using diversity in cluster ensembles. In: IEEE International Conference on Systems, Man and Cybernetics, vol 2. IEEE, pp 1214–1219
15. Nanda SJ, Panda G (2014) A survey on nature inspired metaheuristic algorithms for partitional clustering. Swarm Evol Comput 16:1–18
16. Nguyen DT, Chen L, Chan CK (2012) Clustering with multiviewpoint-based similarity measure. IEEE Trans Knowl Data Eng 24(6):988–1001
17. Pavithra M, Chandrakala D (2013) Measurement of similarity using link based cluster approach for categorical data. In: International Conference on Information Communication and Embedded Systems (ICICES). IEEE, pp 507–516
18. Shabanzadeh P, Yusof R (2014) A new method for solving supervised data classification problems. Abstr Appl Anal. https://doi.org/10.1155/2014/318478
19. Strehl A, Ghosh J (2003) Cluster ensembles a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 3:583–617
20. Tan P-N, Steinbach M, Kumar V (2006) Introduction to data mining. Pearson Addison Wesley, Boston
21. Wolpert DH, Macready WG (1995) No free-lunch theorems for search. Working paper 95-02-010, Santa Fe Institute
22. Xue H, Chen S, Yang Q (2009) Discriminatively regularized least-squares classification. Pattern Recognit 42(1):93–104
23. Yang X-S (2009) Firefly algorithms for multimodal optimization. Stochastic algorithms: foundations and applications. Springer, Berlin, pp 169–178
24. Zaki JM, Peters M (2005) CLICKS: mining subspace clusters in categorical data via K-partite maximal cliques. In: 21st International Conference on Data Engineering, IEEE Proceedings (ICDE 2005), pp 355–356