International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 / https://doi.org/10.14445/22315381/IJETT-V70I9P204
Volume 70 Issue 9, 34-46, September 2022

Original Article

Hybridization of Fuzzy Label Propagation and Local Resultant Evidential Clustering Method for Cancer Detection

M. Aruna[1], S. Sukumaran[2], V. Srinivasan[3]

[1]*Department of CT & IT, Vellalar College for Women, Erode, India*
[2]*Department of Computer Science, Erode Arts and Science College, Erode, India.*
[3]*Department of School of Computer Science, VET Institute of Arts and Science (Co-Education) College, Erode, India*

[1]*Corresponding Author : arunasrini2005phd@gmail.com*

Received: 19 June 2022 Revised: 30 July 2022 Accepted: 11 August 2022 Published: 06 September 2022

Abstract - *Clustering is a data analysis technique that divides information into numerous homogeneous groups. They are clustering algorithms like Centroid, Density-based Distribution and Hierarchical based Clustering. These algorithms provide better performance for only spherical clusters and acquire high-time complexity issues. So, a Belief-Peaks Evidential Clustering (BPEC) method is efficiently used to deal with non-spherical clusters and improve the clustering performance. However, if the number of clusters is too great, the complexity of BPEC becomes exorbitant. Motivated by these challenges, a hybrid of Fuzzy Label Propagation and Local Resultant Evidential Clustering Method (FLPLRECM) is proposed to handle the sparse and helps to form more precise clusters. Initially, to select Cluster Centres (CCs) based on data dispersion and local density, an adaptive CCs selection approach is proposed. This model provides a Symmetric Neighborhood Graph (SNG) to all Data Points (DPs) with other points along with the standard deviation (SD)/kurtosis, local densities of each point are computed by utilizing the reverse k-Nearest Neighbors (k-NN). To deal with the high dimensional dataset, the Centrality (CE) and Coordination (CO) metrics are introduced to classify the DPs as interior points (ips), inner boundary points (ibps), boundary points (bps), or noise DPs for improving the cluster formation. The intensities and orientations of DPs near the fuzzy CCs and at the fuzzy cluster boundaries are assessed. First, a helpful initial fuzzy cluster assignment is made for each remaining point based on the distances between each CC and its neighbours. After then, neighbour's labels are used to refine each point's own till the fuzzy partition remains the same. The developed method will provide more precise clusters with less computational time, efficiently used for the analysis of the cancer detection system.*

Keywords - *Belief-Peak based clustering, Centrality, Coordination, Symmetric Neighborhood Graph, K nearest neighbour.*

1. Introduction

The introduction Clustering is the grouping of specific objects based on their characteristics and similarities. Existing clustering algorithms can be split into two types based on their ideologies: Hierarchical, Partition and Density clustering. A dataset is divided or merged into a series of nested partitions using hierarchical clustering [1]. The nested partitions' hierarchy might be agglomerative (bottom-up) or divisive (top-down). Clustering begins with each individual object in a single cluster and progresses through the closest pairs of clusters until all entities are grouped together in a single cluster.

Partition clustering [2] has been elongated to investigate more difficult data configurations throughout time, emerging in the ideas of hard [3], soft or fuzzy [4], approximate or rough [5], possibilistic [6] and credal [7]. The type of data supplied is another key variation between these clustering algorithms. Data items, in which a list of attributes accurately characterizes each object, and closeness (or relative) data, with only bilateral matches or divergences, are provided, are both partially considered to be two typical categories of data. All you need are the right measurements to transform raw object data into domain data. As a result, clustering algorithms that deal with proximity data are more general than clustering techniques that only engage with the data layer.

Density peaks clustering (DPC), a hard partition clustering approach based on quick density peak search and find, was recently introduced in [8]. Many alternative DPC approaches have been developed to improve clustering performance. A flexible core fusion-density peak clustering method was designed to discover the clusters in any shape

or intensity data. According to [10], the subordinate can be used to interpret co-relative density adjustments. The authors apply this interpretation to the problem of reducing DPC's acuity to density kernels by detecting cluster centres. In [11], a fuzzy kernel and a density-based k-NN statistic are presented for DPC to improve cluster splitting and reduce the occurrence of outliers.

To make the most of the uncertainty and ambiguity inherent in data structures, the BPEC [12] is presented to construct a credal partition for data that permits both solo and composite clusters. The evidence theory [13] extends the idea of density [8] to belief, which indicates the possibility of each object becoming a cluster centre. The entity with the largest delta nearest to the entities with the highest belief counts as the cluster centre. The user generates a belief-delta decision graph and then sets two minimum thresholds for belief and delta. The targets with higher belief and delta are used to determine the locations of the cluster centres.

The persistent entities are consigned using a modified Evidential C-Means (ECM) [14] technique to construct a credal division. However, BPEC and distance-based computational approaches are straightforward and can readily disregard sample correlation and similarity. The manual configuration has a significant impact on the clustering outcomes. As a result, BPEC's clustering effectiveness on high-dimensional datasets is poor. To address the BPEC difficulties, the suggested FLPLRECM algorithm was created to handle sparse datasets and aid in forming more exact clusters. Furthermore, this strategy effectively delivers the optimal collection of characteristics to increase clustering performance, which has the best significance and most effective cancer detection research in the Breast cancer Wisconsin dataset.

Initially, an adaptive CCs selection approach is created that effectively selects CC based on data dispersion and local density. The fragments technique is combined based on the structural similarity concept to strengthen its competence in cluster centroid recognition and better identify potential abnormalities. In addition, the suggested approach refines the clusters at each iteration by combining the SD/kurtosis with Symmetric Neighbourhood (SN) relationships of DPs with other DPs. The cut-off distance determines the boundary between two clusters with varying densities, even though there will be one. An SN of DPs replaces the cut-off distance. The network that connects each point's SN is known as an SN graph (SNG). Outliers are locations in the SN that have fewer than two neighbours. It uses a depth-first search on an SNG to distribute each vertex to a correct cluster without cutting off distance.

Then, to better cluster the high-dimensional dataset, the CE and CO metrics are used to categorize DPs as ips, ibps, bps or noisy DPs. The integration of magnitudes and orientations of DPs adjacent to fuzzy CC and at the fuzzy cluster boundaries will be assessed next. The distances between each CCs and its neighbours produce an instructive primary fuzzy cluster consignment for each enduring point. The fuzzy label of every enduring point will then be modified repeatedly by integrating the labelled data of its neighbours until the fuzzy segregation becomes constant. The suggested technology will provide more exact clusters in less time, which can be utilized to analyze cancer detection systems more effectively.

2. Literature Survey

Aboubi et al. [15] developed a new BAT-CLARA technique for clustering large data sets. This method was based on the bat behaviour and partitioning of k-medoids. This novel method was compared to well-known partitioning methods like PAM, CLARA, CLARANS, and CLAM, as well as a recently discovered algorithm. However, it has the disadvantages of consuming a lot of time and using up a lot of memory.

Kaur and Ojha [16] proposed defining a normal subscriber of a movable operator and explained a framework for checking the design and symmetries in pseudo anonymized Call Data Records (CDR). It describes the difficult task of automatically generating expressive information from accessible data using a machine learning approach for clustering without including prior experience of the interface context in the network. The outcomes of clustering mining are used to gain better insights into the client's behaviours and to attract their illustrative profiles.

Chang et al. [17] created the Deep Adaptive Clustering (DAC) approach to cluster images using the single-stage convolutional network. The method was inspired by the basic concept that the association between pairs of images was binary and that the binary pairwise-classification issues were the method's optimal target. The cosine proximity between label attributes assesses the pairwise correlations, and the images are depicted by label features retrieved by a Convolutional Neural Network (CNN). In addition, DAC requires that the learnt label features be one-hot vectors.

Wang et al. [18] proposed a new chaotic starling Particle Swarm Optimization (PSO) technique to resolve the clustering issues. In this process, KMDD (clustering by K-means using both density and distance-oriented measures) was a two-phase clustering algorithm created to quickly locate clusters with various forms and intensities in temporal datasets. On the other hand, this approach has a slow convergence rate.

Zhou et al. [19] developed a clustering model based on a new Semi-supervised Evidential Label assignment approach which incorporates domain knowledge and performs clustering. After the propagation process has stabilized, the communities of each node can be recognized in this graph. The graph's prior knowledge of the node's relation with labels was expressed using mass functions. Using evidentiary label propagation principles, the labels were then propagated from labelled to unlabelled nodes. However, this strategy produces different results if the labelled data contains a different parameter.

Narayana and Vasumathi [20] created a similarity-based clustering approach that involves calculating and merging similarities between and within characteristics. Clustering the attributes by their similarities using the K-medoids clustering algorithm helped keep the computations to a minimum. In addition, the best characteristics were chosen using the Bee Colony (BC) optimization method. Unfortunately, there is a great deal of computational inaccuracy in this approach.

Pang et al. [21] introduced a MapReduce-based multilevel subspace clustering. This clustering worked well while dividing a large-scale dataset into smaller ones with the same data elements. During the localized clustering step, PAPU produces sub-clusters based on specific attribute subspaces from distinct sections, facilitating parallel computing. To increase the reliability of estimated clustering results, PAPU employs the hierarchical clustering method to combine sub-clusters iteratively across the whole global clustering stage, regardless of cluster size.

Budiaji and Leisch [22] devised a simple and quick k-medoids algorithm for updating medoids by minimizing the overall distance between clusters. A generalized distance function is constructed with the distance as an input to the technique to maximize the variance of the distances, especially for a combined variable dataset. Because different distances produce different outputs, the variation of the distances is an important aspect of a partitioning method. Because of their time complexity, K-Mediods are more expensive than K-Means Procedure.

Ping et al. [23] proposed a new efficient technique for handling the support vector clustering cluster labelling problem (SVC). The proposed approach examines the topology of the functions that describe SVC cluster outlines and looks for interconnection paths between crucial points that separate various cluster contours. The suggested algorithm incorporates a new quick way of discovering and classifying crucial points and evaluating their interaction structures.

Meng et al. [24] use belief peaks within a linear label assignment scheme. This method revealed the underlying data structure by counting clusters precisely and producing a fuzzy breakdown. The label assignment scheme is a useful alternative method in belief-peaks clustering due to its explicit convergence and linear overhead. An excessive amount of time was needed for this technique.

3. Proposed Methodology

This section quickly explains the main contribution of this work, which is the incorporation of magnitudes and directions of DPs close to CC and at cluster boundaries. Two new local metrics, CE and CO, are employed to represent these disparities more effectively. These new metrics assist in the classification of DPs as ips, ibps, bps, or noise DPs at a coarse level. It's worth noting that these two measurements are used to create SNG. The CE is a variable that can be anywhere between 1 and 1. When CE > 0, a data point is more likely to be ips; when CE < zero, it is more likely to be a border point. It is much simpler to separate DPs from inner cluster areas and border cluster areas with a more apparent meaning. The CO indicates how well a data point fits in with its surroundings. A CO value of CO > 0 shows that DPs are oriented in the same position as their neighbours and are most likely near the border. Along with these new criteria, the value of the neighbourhood association is computed.

3.1. BPEC method

In BPEC [12], a new element of perception $C = \{C, \neg C\}$ is interpreted to determine if an object is a CCs (C) or not $(-C)$ for a given set D of n data classes. The following is a summary of the core concept for detecting CCs. The collection of K nearest neighbours (KNN) of the object o_i is denoted by$NN_K(D_i)$. Every $NN_K(D_i)$ neighbour $o_j$contributes to the evidence that entity $o_i$is a CCs. The set of c clusters is specified as$\Omega = \{\omega_1, \omega_2 \cdots, \omega_c\}$. A hard partition of Dataset $D$is applied to find which cluster each object belongs to. A mass function m_{ij}^C can be used to represent the actual information. A normalized mass function m_i^C and its corresponding belief function Bel_i^C can be generated by merging these mass functions using Dempster's rule.

In considering the concept of belief functions, the combination of mass functions plays a crucial role. Allow two mass functions, m_1 and m_2. The un-normalized mass function is the conjunctive conjunction of m_1 and m_2 in Equations 1 and 2.

$$m_{1\cap2}^{\Omega}(A) = \sum_{B\cap C=A} m_1^{\Omega}(B) m_2^{\Omega}(C), \forall A \subseteq \Omega \quad (1)$$

The normality requirement $m^{\Omega}(\emptyset) = 0$can be regained if necessary by dividing each mass $m_{1\cap2}^{\Omega}(A)$by $1 - m_{1\cap2}^{\Omega}(\emptyset)$. The procedure that results is called Dempster's rule of combination:

$$m_1^{\Omega} \oplus_2 (A) = \frac{m_{1\cap 2}^{\Omega}(A)}{1-m_{1\cap 2}^{\Omega}(\emptyset)} \ , \emptyset \neq A \subseteq \Omega \quad (2)$$

Both rules, which are associative and progressive, allow the vacuous mass to function as a single neutral element. If an item has a higher degree of belief $Bel_i^C(\{C\})$ Than its neighbours and is also at a reasonably big distance from other objects with higher degrees of belief, it is referred to as a CCs.

3.2. Cluster Center Selection with Adaptive Strategy

Centres of clusters are chosen by manually employing two-dimensional decision graphs for accurate clustering. Ideally, the CC would be chosen from a pool of locations with a high local density, where the next location with a higher density is only a short drive away. However, this strategy has several limitations and variability in practice. It's tough to choose the right number of CC when the distribution of these sites is comparable. An adaptive CC selection SD/kurtosis of the distance function using the data distribution and the local density is created to eliminate the uncertainty and complexity of directly identifying CC.

Kurtosis is defined as the systematized fourth population period about the mean, $Kurt[X] = \beta_2 = \frac{E(X-\mu)^4}{(E(X-\mu)^2)^2} = \frac{\mu_4}{\sigma^4}$ Where E is the expectation function, μ is the mean, μ_4 is the fourth moment about the mean, and σ is the SD. The normal distribution has a kurtosis of 3, and $\beta_2 - 3$ is often used so that the preferred normal distribution has a kurtosis of zero $\beta_2 - 3$ is denoted as γ_2. A sample which is equivalent to β_2 can be obtained by replacing the population moments with the sample moments, which results in Equation 3

$$b_2 = \frac{\sum(X_i-\bar{X})^4/n}{(\sum(X_i-\bar{X})^4/n)^2} \quad (3)$$

Where b_2 represents the sample kurtosis, $\bar{X}$ The bar represents the sample mean, and n represents the number of data. On the one hand, the normal variation of all variables is determined and employed as a measure of the arithmetical measures for data's dispersal level to some extent. Then, if the distance between the sample and the adjacent larger density point (LDP) is higher than or equivalent to the weighted average deviation, it is designated a CC.

As a result, CCs are defined as DPs with a local density larger than the average of all data's local density. The constraints for choosing CC are defined by merging the two procedures above in Equations 4 and 5.

$$ExpC = \delta_i \geq \lambda\sigma(\delta_i) \quad (4)$$

$$CentreC = ExpC(\rho_i) \geq \mu(\rho_i) \quad (5)$$

where EC stands for the expected CCsestimated with Kurtis, δ_i stands for the range from the nearest LDP, $\sigma(\delta_i)$ stands for the SD of the distance from the adjacent LDP of all input data, and λ is the weight. After reducing noises, $Centre\ C$ signifies the CCs, ρ_i is the local density of each sample point, $ExpC(\rho_i)$ defines the expected CCs, local density, and $\mu(\rho_i)$ is the median of all the local populations. The CC chosen by the two phases above provides a significant distance from the nearest LDP and a large local density, avoiding errors caused by choosing noises as CCs and ensuring the impartiality of clustering outcomes.

3.3. Developed Evidential Clustering with fragment merging strategy

The proposed Evidential Clustering is combined with the fragment fusion approach to detect cluster centroids better. The latent loss calculation is used to determine local densities associated with a neighbourhood region's datasets in this example. According to the d_{ij} d ij and the $belief_i^C(\{C\})$ each pattern's local density in Equation 6

$$\rho_i = \sum_{j\in NN_k(D_i)} \varphi\left(d_{ij}^2\right) - C_d \quad (6)$$

As a result, instead of applying the heuristic technique of decision graphs, the created sets of residual mistakes are employed to construct residual fragments, which are then processed to find clusters and cluster centroids. For more appropriate local density estimation, utilize the residual error calculation to assess the density of each data point within its neighbourhood section, which may lead to a greater clustering process and centroid recognition. The residual error re_{ij} between DPs x_i and x_j is calculated in Equation 7.

$$re_{ij} = \frac{\|x_i - x_j\|}{N} \quad (7)$$

Where N is the size of the neighbourhood. It is a user-defined consistent variable used to discover N total of nearest neighbours of x_i, where an adaptive CCs selection approach based on the dispersion of the data and the local density is used to minimize the inconsistency and complexity of directly classifying CCs. The residual error of x_i can also be calculated in Equation 8 as follows

$$re_{ij} = \sum_j \frac{d(x_i,x_j)}{N}, \forall\ d\left(x_i, x_j\right) = \frac{\mu_4}{\sigma^4} \quad (8)$$

Where μ is the average, μ_4 is the fourth period about the mean, σ is the SD, and the kurtosis proximity operation is used. To construct residual fragments for cluster creation, find each data point's adjoin point and neighbourhood points after preprocessing.

3.4. Belief Peaks Clustering In Symmetric Neighborhood

Let D represent a database, i and j represent some D objects, and k represent a positive number. The SD/kurtosis distance between entity i and j is denoted by $dist(i,j)$. The k-distance of i abbreviated as $kdist(i,o)$, is the distance $dist(i,o)$ between points i and o in D, which is represented as:

- It holds that $dist(i,o') \leq dist(i,o)$, for at least k objects $o' \in D$,
- and $dist(i,o') < dist(i,o)$ for at most $(k-1)$ objects $o' \in D$

If a point j meets $dist(i,j) \leq kdist(i)$, then call j as one of the kNN of i. The kNN of i is formed by a set of points J that comprises limited points j, written as $kNN\ (i)$. The following Equation 9 is the meaning of $kNN(i)$:

$$kNN(i) = \{J \in D | dist(i,J) \leq kdist(i)\} \quad (9)$$

The point i is considered the reverse kNN of j, and a collection of points I containing finite points i is referred to as the reverse k-NN, abbreviated as $RkNN\ (j)$. $RkNN(i)$ is a function that can be determined in Equation 10 as follows

$$RkNN(i) = \{j | j \in D, i \in kNN(j)\} \quad (10)$$

The probability density around i is estimated using the junction of the Knn neighbourhood and the inverse k-NN. The neighbourhood space is referred to as the symmetric neighbourhood of i abbreviated as $SNk\ (i)$. $SNk\ (i)$ means that two people are true friends only if they agree with one other, as demonstrated by Equation 11

$$SNk(i) = \{o | o \in D, o \in \big(kNN(i) \cap RkNN(i)\big) \quad (11)$$

Generally, searching kNN of point i yields at least k outcomes, whereas RkNN yields zero, one, or many results. Consider the effect of a point from other points, so estimate the local density using the inverse kNN of a point instead of other nearest neighbours, making it easier to find CCss. The current local density is determined in Equation 12 as follows:

$$\rho_i = \sum_{j \in RkNN(i)} exp\big(-dist^2(i,j)\big) \quad (12)$$

Where $RkNN(i)$ is the point$i$'s reverse kNN. While the previous definition [25] is determined using the cut-off distanced_{co}This approach can ensure that the local density ρ_i of point, i is impacted by the allocation data of its inverse kNN. It's difficult to predict the number of cut-off distance d_{co}, which impacts the local density of nodes and CCs selection. Moreover, determining parameter k is simpler than determining the cut-off distance d_{co}.

The network created by connecting each point's symmetric neighbourhood is known as a symmetric neighbourhood graph (SNG). Deviations are locations in the symmetric neighbourhood that have lower than two neighbours. Even though there would be a boundary among two clusters with varying densities, some criteria must be included to enforce the boundary.

3.5. FLPLRECM

3.5.1. Local Gravitation

The local gravitation in the data clustering approach exemplifies the relationship between a data point and its immediate neighbours. According to the theory of gravity, the attractive force across 2-point masses (m_1, m_2) can be calculated using the given formula:

$$\vec{F}_{m_1 m_2} = G \frac{m_1 m_2}{D^2_{m_1 m_2}} \widehat{D}_{m_1 m_2} \quad (13)$$

The force between point masses 1 m_1 and 2 m_2 is denoted by $\vec{F}_{m_1 m_2}$. $D_{m_1 m_2}$ represents the distance between m_1 and m_2, and the vector $\widehat{D}_{m_1 m_2}$ determines the way of connecting the two data instance masses based on force acts, and finally, G is the gravitational constant.

3.5.2. Local Resultant Forces

The essential premise of clustering tactics is that the LRFs of DPs adjacent to CCs and those at the cluster's edge differ significantly. The LRF depicts the relationship between each data point and its immediate surroundings. Suggest two local clustering techniques based on local gravity strength to advantage the information enclosed in the LRF: 1) the CE and 2) the CO. The Data point x_i CE is calculated in Equation 14 ,15,16 and 17 as

$$CE_i = \sum_{j=1}^{k} \frac{\cos(\vec{F}_j, \vec{D}_{ij})}{k} \quad (14)$$

Where $\vec{D}_{ij}$the displacement vector from the data is point x's the $j-th$ neighbour to it, and k is the neighbour association. A data point with a CE value of CE *i* > 0 shows that most of its neighbours' LRFs are pointing in its direction. Since

$$1 \leq \cos\big(\vec{F}_j, \vec{D}_{ij}\big) \leq 1 \text{ and } -k \leq \sum_{j=1}^{k} \cos\big(\vec{F}_j, \vec{D}_{ij}\big) \leq k \quad (15)$$

CE has the properties listed below:

$$-1 \leq CE_i \leq 1 \quad (16)$$

The CO of data point x_i is calculated as follows:

$$CO_i = \sum_{j=1}^{k}(\vec{F}_i, \vec{F}_j) \quad (17)$$

The LRF of the data instance x_i is $\vec{F}_i$And the force accompanying its neighbours is $\vec{F}_j$. The CO represents a data point's familiarity with its neighbours in general. A data instance with a CO value of $CO > 0$ has an LRF that is generally in the same position as its neighbours, and it is most likely near the border.

3.5.3. Fuzzy label propagation process

Let$X_{center} = \{x_1, x_2, \ldots, x_C\}$ and $X_{Noise} = \{x_{n-1+1}, x_{n-1+2}, \ldots, x_n\}$ represent the collection of the selected CC and their outliers, respectively. It proves that the datasets contain c clusters $O_q(q = 1,2,\ldots,c)$. Moreover, l finds the outliers being ignored from the X and only ruminate the cluster forming off the remaining $n - l$ point to eliminate the impact of outliers.

The combinatorial hypothesis [26, 27] states that comparable points are more likely to have a similar label. Based on this premise, the class of each point x_i's neighbours x_j ($x_j \in \boldsymbol{\mathcal{N}}_K(x_j)$) are merged to estimate its label. Each neighbour's weight w_{ij} is directly equivalent to the distanced_{ij}, as is well known. Furthermore, if the neighbour x_j is a CCs, it is more probable that x_i has the same label asx_j . As a result, the amount of $Belief_j^{\mathcal{C}}$ ($\{C\}$) is equivalent to the weight w_{ij} of neighbour x_j .

A novel metric for calculating neighbour weights. For $i, j = 1,2,\ldots,n-l$,

$$w_{ij}^{*} = \begin{cases} Belief_j^{\mathcal{C}}(\{C\}) \exp\left(-\frac{d_{ij}^2}{\beta_i^2}\right) & if\ x_j \in \boldsymbol{\mathcal{N}}_K(x_i) \\ 0 \quad otherwise, \end{cases} \quad (18)$$

The cluster assessment $y_i = [y_i(O_1), y_i(O_2), \cdots, y_i(O_c)]$ of data point x_i can be called a fuzzy label, according to the concept of "fuzzy partition" in references [12] and [28]. In this case, $(O_q) \in [0,1] (0 \leq q \leq c)$ and $\sum_{q=1}^{q=c} y_i(O_q) = 1$. Fuzzy labels are used to assess the degree of ambiguity in data point cluster classifications.

- *Initial partition*

Indicate the point's original label x_i as $y_i^{(0)} = [y_i^{(0)}(O_1), y_i^{(0)}(O_2), \cdots, y_i^{(0)}(O_c)$. For any $q \in \{1,2,\cdots,c\}$, each of the CCs points $x_k \in X_{center}$ obviously belongs to the cluster O_k which is denoted in Equation 19 as

$$y_k^{(0)}(O_q) = \begin{cases} 1 & if \quad q = k \\ 0 & otherwise \end{cases} \quad (19)$$

The generally used label propagation models [29, 30] view each of the remaining points $x_u (u = c+1, c+2, \ldots, n-l)$, as having a zero vector as its label, which is unspecific. In addition, it suggests a normalized and relevant initial label for x_u Based on the allocation data of its neighbours. The actual proximity among its K nearest neighbours and the CCs point x_q is used to calculate the possibility of point x_u belonging to cluster O_q using an exponential expression to every $q \in \{1,2,\cdots,c\}$, where the variable $\eta_q^2 = \max d_{jq}^2 \ |x_j \in \boldsymbol{\mathcal{N}}_K(x_u)$. Equations 20 to 24 represent the steps of the proposed clustering method

$$y_u^{(o)^*}(O_q) = \exp(-1 \backslash K \sum_{x_j \in \boldsymbol{\mathcal{N}}_K(x_u)} w_{uj} \frac{d_{jq}^2}{\eta_q^2}) \quad (20)$$

Normalization can then be used to initiate fuzzy label $y_u^{(O)}$ of pointx_u. For each of the $q \in \{1,2,\cdots,c\}$,

$$y_u^{(O)^*}(O_q) = y_u^{(O)^*}(O_q) \backslash \sum_{q=1}^{c} y_u^{(O)^*}(O_q) \quad (21)$$

The instructive initial fuzzy partition $Y^{(0)} = \{y^{(0)}\}_{i=1}^{n=1}$ is obtained as a result.

- *Label propagation process*

Based on the stated initial partition, each point continuously upgrades its label by partially fascinating the label evidence circulated by its K nearest neighbours. $Y^{(0)}$ and weight matrix W. After the t th label propagation, let $Y^{(t)} = \{y_i^{(t)}\}_{i=1}^{n=1}$ signify the fuzzy partition. Under the synchronous updating mechanism, the label of each point x_i will be progressively changed as follows.

$$y_i^{(t)} = \alpha \sum_{x_j \in \boldsymbol{\mathcal{N}}_K(x_i)} w_{ij}\, y_j^{(t-1)} + (1-\alpha)\, y_i^{(0)} \quad (22)$$

A proportion variable $\alpha (0 < \alpha < 1)$ is used. The label propagation mechanism regulates the amount of label information absorbed from its neighbours. When $x_j \notin \boldsymbol{\mathcal{N}}_K(x_i)$*A*ccording to the specification of the weight matrix W, $w_{ij} = 0$ is defined. As a result, the label propagation mechanism can be summed up as follows:

$$Y^{(t)} = \alpha W Y^{(t-1)} + (1-\alpha) Y^{(O)} \quad (23)$$

Following the t th circulation, the fuzzy partition $Y^{(t)}$ can be obtained. The following statement will then prove that the label propagation section will be convergent and construct a permanent fuzzy partition.

Proposition 1 The linear label propagation model [31] yields a convergent fuzzy partition that is

$$Y = (1-\alpha)(I - \alpha W)^{-1} Y^{(O)} \quad (24)$$

Most importantly, the label propagation mechanism converges to a stable fuzzy division Y is represented as

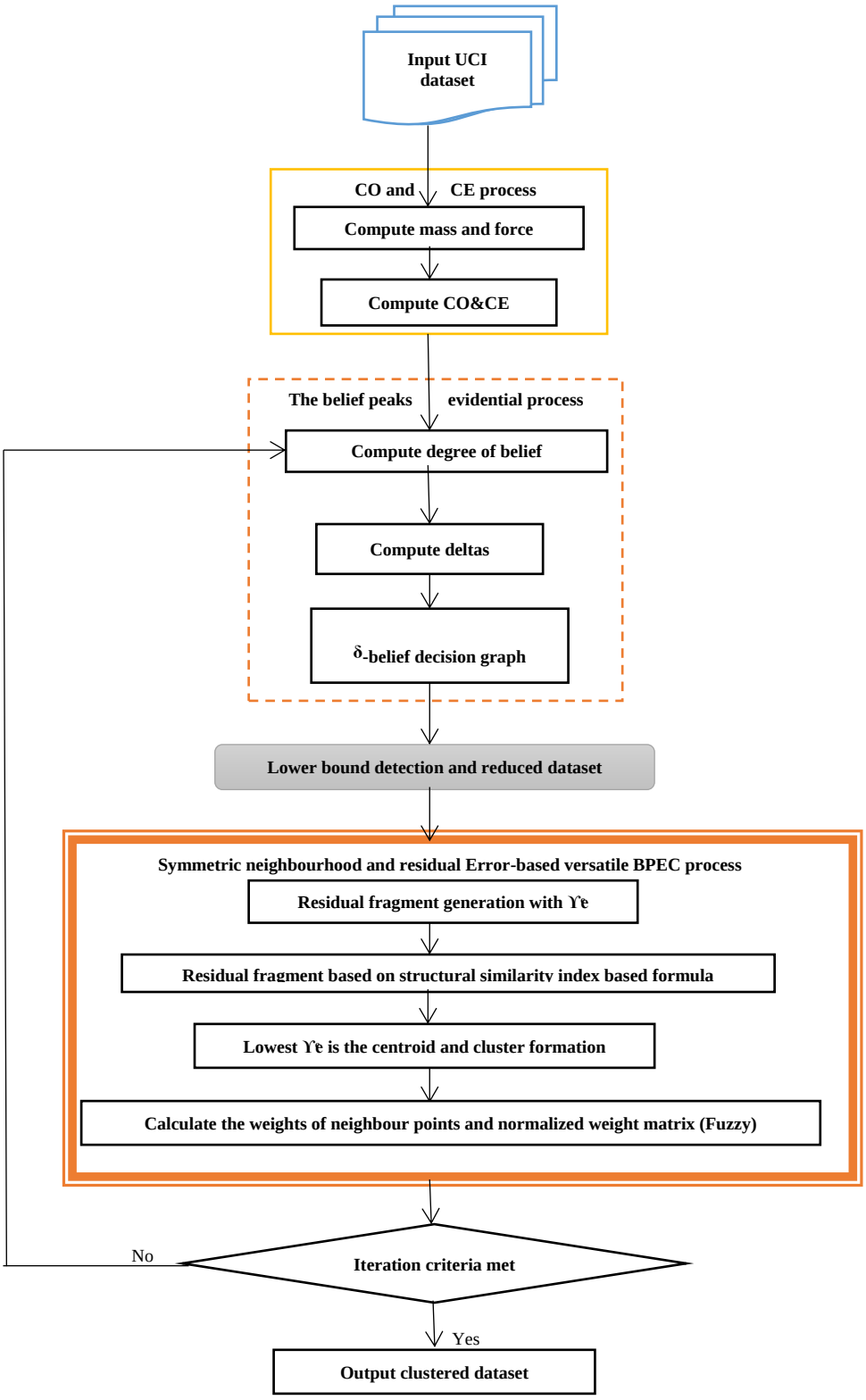

Fig. 1 Flow diagram of FBPLRE

$Y = (1-\alpha)(1-\alpha W)^{-1}\ Y^{(O)}$ completing the suggested technique. Figure 1 and algorithm 1 depict this research work's overall flow.

Algorithm 1: FLPLRECM

Input: Dataset D with n DPs and a user-defined neighbourhood size N, error vector re, sortd_re, index vector of each data point's nearest neighbour set of NN

Output: Clustering results

1. Compute the degrees of belief for all entities $(Bel_i^C(\{C\}))$

$$belief_i^C(\{C\}) = 1 - \prod_{x_j \in NN_k(x_i)} \left[1 - \varphi\left(d_{ij}^2\right)\right]$$

2. Determine each sample's local density ρ_i according to the d_{ij} and the $belief_i^C(\{C\})$ as $\rho_i = \sum_{j \in NN_k(o_i)} \varphi\left(d_{ij}^2\right)$
3. Determine the distance from each sample's nearest LDP δ_i corresponding to ρ_i.
4. Using the ρ_i and the distance from the $\delta_i = \min_{j:\rho_j > \rho_i} \left(\frac{\mu_4}{\sigma^4}\right)$ construct a two-dimensional decision graph.
5. Compute the deltas (δ_i) for all objects using the following formula:

$$\delta_i = \begin{cases} \min\limits_{j:belief_j^C(\{C\}) > belief_i^C(\{C\})} \{d_{ij}\} & \text{x}_\text{i} \text{ does not have the highest belief} \\ \max\limits_{1 \le j \le n} \{d_{ij}\} & otherwise \end{cases}$$

6. Create δ –belief decision graph and find the d_{min} and belief$_{min}$ lower limits.
7. Choose the CC
8. Determine the shortest distance between the remaining sample points and the CCss, then place the other data samples in the CCs closest to them.
9. Compute Residual Fragment Generationre →For each data pointx_i and sortd_re and obtain $NNset$do

$$re_{ij} = \frac{\|x_i - x_j\|}{N} \text{and} re_{ij} = \sum_j \frac{d(x_i, x_j)}{N}, \forall\ d(x_i, x_j) = \frac{\mu_4}{\sigma^4}$$

10. Check the adjoin point identification criterion and satisfies update APs accordingly;

$$APs = APs \cup x_i, iff \left\|x_i - x_j\right\| < C_d \ \forall\ x_j$$

11. For every locality point of data point x_i do
12. Calculate the weights of neighbour points and normalized weight matrix
13. Check the $NNset$ recognition criterion is met and then upgrade n neighbour set accordingly;
14. Determine that the NN set identification criterion is satisfied before updating $NN\ set$.

$$NN_x = \{x_j | \|x_i - x_j\| < C_d\}$$

15. Connect each data point to its neighbouring points, i.e., $x_i + 1$, and generate Apt link to build a connection.
16. Construct residual fragments and update the fragment set as needed.

$$reFx = \text{APtlink} \cup \left(NN_x \cup NN_y\right)$$

17. Calculate each fragment's structural similarity index.

$$reF_{sim}(x, y) = \frac{\left|reF_x \cap reF_y\right|}{\sqrt{|reF_x||reF_y|}}$$

18. Check $reF_{sim}(x, y) > TH$ is met, then combine fragments x and y as a single cluster.
19. Otherwise, generate a new cluster;
20. Identify the data point with the least error (cluster centroid) for each created cluster
21. Apply x_i to the cluster label of the cluster's selected centroid; also, check the un-clustered boundary point and $CO_i \ge 0$, then add x_i to its nearest cluster;
22. Calculate the Euclidean distance d_{ij} between x_i and x_j
23. Find out the set $\boldsymbol{\mathcal{N}}_K(x_i)$ of k-NN of x_i
24. $\delta - Bel$ find the CCsX_{center} and outliers X_{noise}
25. Calculate the weight matrix W by (17)-(18)
26. Calculate the initial partition $Y^{(o)}$ by (19)-(21)
27. Calculate the convergent fuzzy partition Y by (24)
28. Final $\mathcal{CL}$ is the result of the FBPLRECM algorithm

4. Results and Discussion

This research compares the results of the proposed FBPLRECM with existing BPEC [12], SELP-GDC [19], and BPEC-FPL [24] on benchmark datasets such as Diabetes 130, Drug review, Codon Usage, and Breast cancer Wisconsin to compare the results of proposed FBPLRECM with existing BPEC [12], SELP-GDC [19], and BPEC-FPL [24] in terms of Precision, F-Measure, Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI).

4.1. Precision

Precision is defined as the percentage of pairs that are appropriately placed in the identical cluster, and it is estimated in Equation 25 as follows:

$$Precision = \frac{TP}{TP+FP} \quad (25)$$

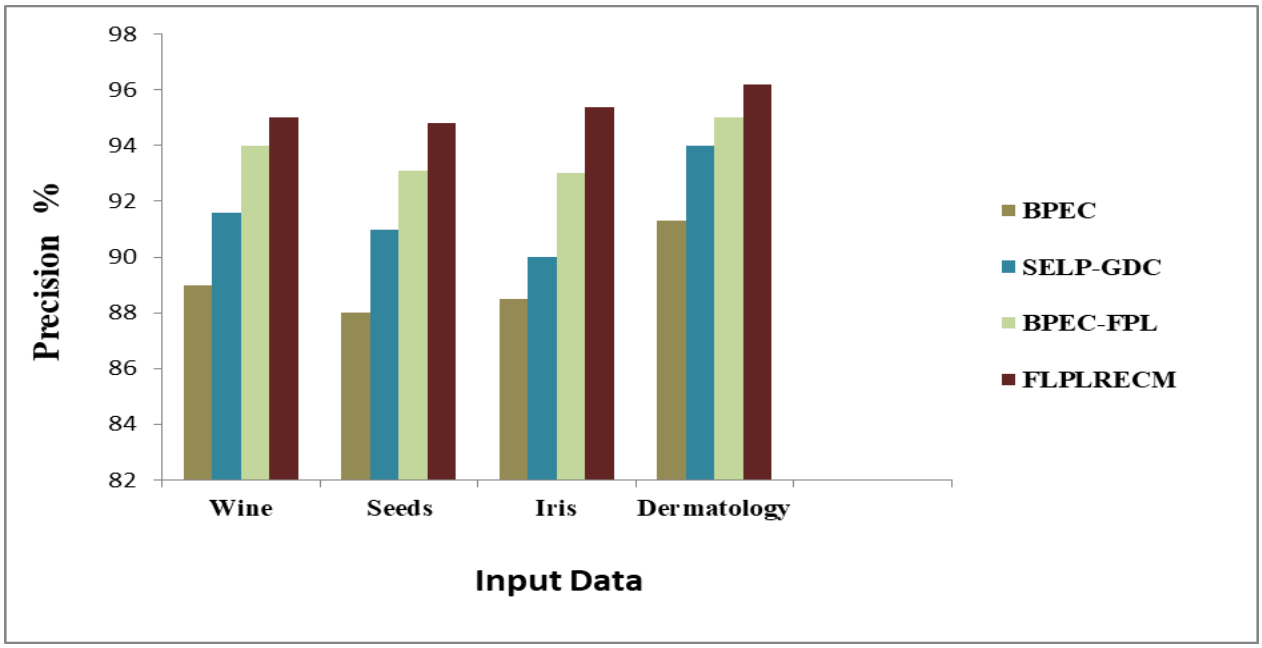

Fig. 2 Comparison of Precision Rate

Table 1. Numerical Results of Precision Rate

Input Data	BPEC	SELP-	BPEC-FPL	FLPLRECM
Diabetes 130	89	91.6	94	95
Drug review	88	91	93.1	94.8
Codon Usage	88.5	90	93	95.4
Breast cancer Wisconsin	91.3	94	95	96.2

The precision comparison results of different methods with diverse input data are provided in Table 1, which gives the numerical results of the Precision rate. In this analysis, the precision value of FLPLRECM is 6.74 %, 3.71 % and 1.06 %; 7.72%, 4.17 % and 1.82 %; 7.79 %, 6% % and 2.58 %; 5.36 %, 2.34 % and 1.26 % higher than that of the existing method like BPEC SELP-GDC and BPEC-FPL respectively for the provided Diabetes 130, Drug review, Codon Usage and Breast cancer Wisconsin. From Figure 2, the proposed FLPLRECM method can obtain a high precision rate when compared to existing methods

4.2 F-measure

It is a metric for assessing the quality or accuracy of clustering algorithms. Precision and recall are the two parameters that are used to calculate F-score. The F-Measure produces a single score that averages the effects of precision and recall problems, and it is defined in Equation 26 as

$$F - Measure = \frac{(2 * Precision * Recall)}{(Precision + Recall)} \quad (26)$$

Table 2. Numerical Results of F-measure Rate

Input Data	BPEC	SELP-GDC	BPEC-FPL	FLPLRECM
Diabetes 130	85	87	89	91
Drug review	86	87.9	89.3	90.9
Codon Usage	85.6	87	89.4	92.7
Breast cancer Wisconsin	88	89.1	92	93.5

The numerical results of the F-measure rate are shown in Table 2. Figure 3 shows the F-measure comparison results between proposed FLPLRECM with existing BPEC, SELP-GDC and BPEC-FPL on various datasets. In this analysis, the F-measure value of FLPLRECM is 7.05 %, 4.59 % and 2.24 %; 5.69 %, 3.41 % and 1.79%; 8.29%, 8.18% and 3.69%; 6.25%, 4.93% and 1.63% higher than that of the existing method like BPEC SELP-GDC and BPEC-FPL respectively for the provided Diabetes 130, Drug review, Codon Usage and Breast cancer Wisconsin. The figure shows that the proposed FLPLRECM method can obtain a high F-measure rate compared to existing methods for better clustering results.

4.3. ARI RATE

Rand index, RI, is calculated in Equation 27 by:

$$ARI = \frac{2((TP*TN)-(FP*FN))}{(TP+TN)(TN+FN)+(TP+FP)(FP+FN)} \quad (27)$$

In a true positive (TP) decision, two comparable documents are assigned to a similar cluster, whereas two dissimilar documents are assigned to separate clusters in a

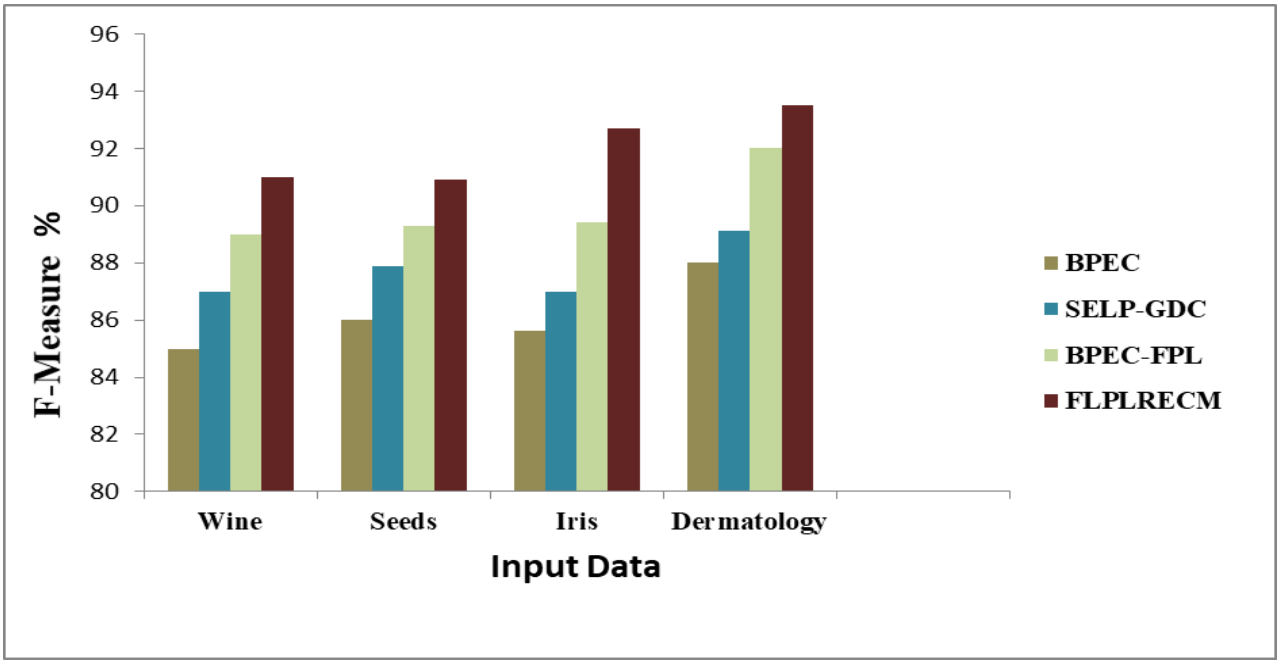

Fig. 3 Comparison of F-measure rate

true negative (TN) decision. False Positive (FP) decisions group two dissimilar documents together, whereas False Negative (FN) decisions group two similar objects together.

The refines the clusters at each iteration by using a symmetric neighbourhood relationship of DPs without cut-off distance will improve clustering results efficiently. The numerical results of the ARI rate are shown in Table 3.

Table 3. Numerical Results of ARI Rate

Input Data	BPEC	SELP-GDC	BPEC -FPL	FLPLR ECM
Diabetes 130	86	87.6	88	91.4
Drug review	85	87	89	93.4
Codon Usage	87	89	91	94
Breast cancer Wisconsin	89.1	92	93.9	95.6

The numerical results of the ARI rate are depicted in Table 3. Figure 4 shows the ARI evaluation results between proposed FLPLRECM with existing BPEC, SELP-GDC and BPEC-FPL on various datasets. In this analysis, the ARI value of FLPLRECM is 6.27%, 4.33% and 3.86%; 9.88%, 7.35% and 4.94%; 8.04%, 5.61% and 3.29%; 7.29%, 3.91% and 1.81% higher than that of the existing method like BPEC SELP-GDC and BPEC-FPL respectively for the provided Diabetes 130, Drug review, Codon Usage and Breast cancer Wisconsin. The figure shows that the proposed FLPLRECM method can obtain a high ARI rate compared to existing methods for better clustering results.

4.4. NMI Rate

The NMI formula is as described in the following Equation 28:

$$NMI(X,Y) = \frac{MI(X,Y)}{\sqrt{E(X)*E(Y)}} \quad (28)$$

The similarity measure between two arbitrary variables X and Y is $MI(X,Y)$While the entropy of arbitrary variables X and Y is E(X) and E(Y), clustering efficiency improves as the NMI value increases.

Table 4. Numerical Results of NMI Rate

Input Data	BPEC	SELP-GDC	BPEC -FPL	FLPLR ECM
Diabetes 130	88	89	90.2	0.93
Drug review	87.4	88.9	91	92.7
Codon Usage	87	89	91	94
Breast cancer Wisconsin	89	91	93	94.6

The numerical results of the NMI rate are shown in Table 4. Figure 5 explains the NMI comparison for predicting CCs of BPEC, SELP-GDC, BPEC-FPL and FLPLRECM methods. The number of data increases according to the NMI value is increased linearly. In this analysis, the NMI value of FLPLRECM is 5.68%, 4.49% and 3.10%;6.06%, 4.27% and 1.86%;8.04%, 5.61% and 3.29%;6.29%, 3.95% and 1.72% higher than that of the existing method like BPEC SELP-GDC and BPEC-FPL respectively for the provided Diabetes 130, Drug review, Codon Usage and Breast cancer Wisconsin. The figure shows that the

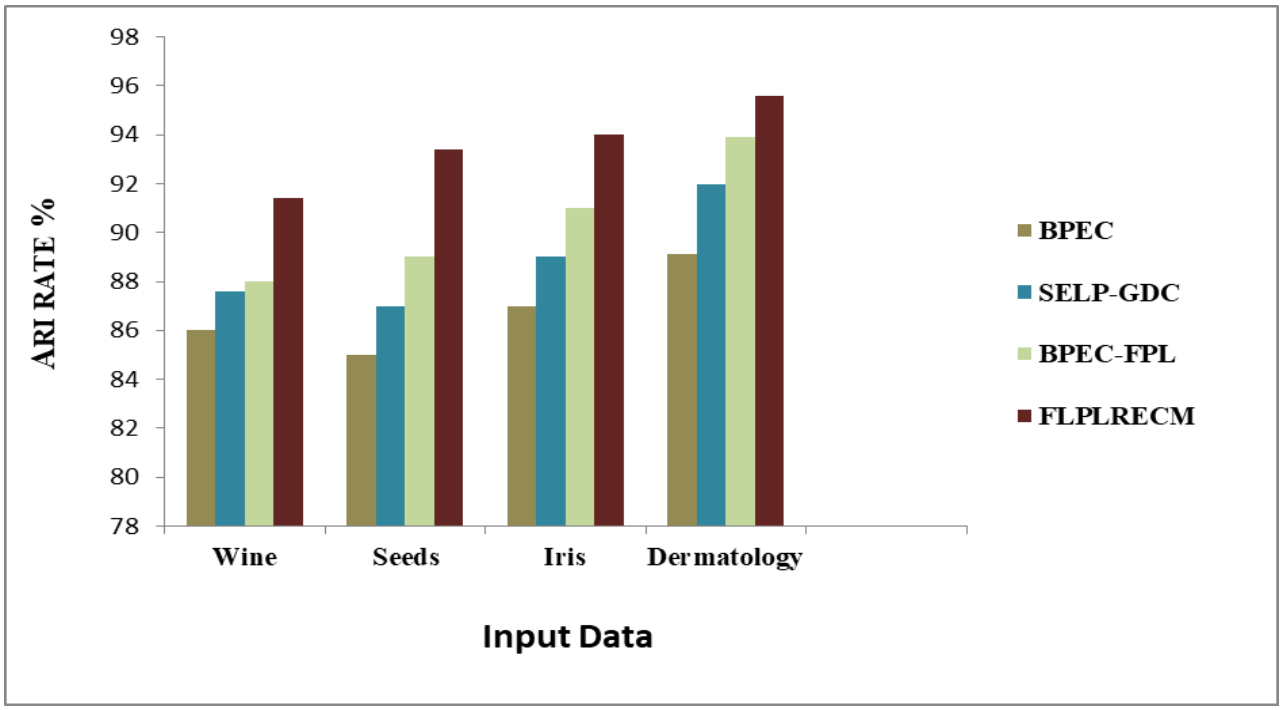

Fig. 4 Comparison of ARI Rate

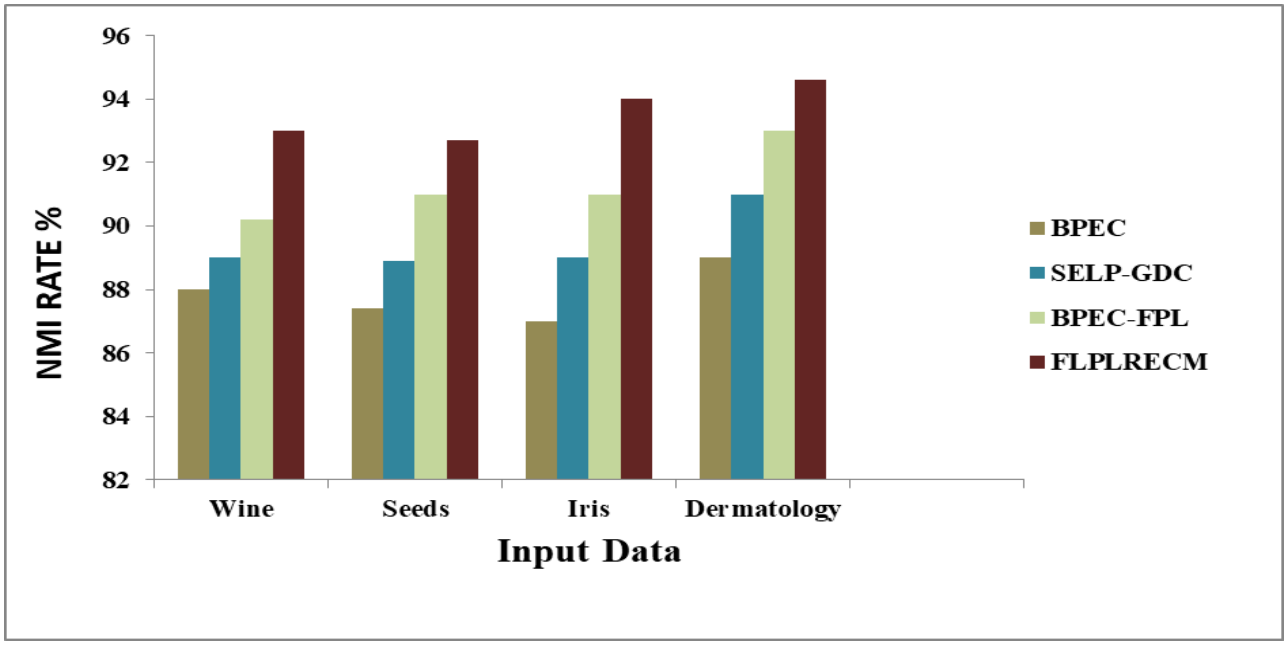

Fig. 5 Comparison of NMI Rate

proposed FLPLRECM method can obtain a high NMI rate compared to existing methods for better clustering results. Hence, the proposed method provides better clustering results with a high NMI rate.

5. Conclusion

In this research work, FLPLRECM is proposed to handle the sparse dataset; additional metrics are introduced to classify the DPs for improving cluster formation. The key addition of this work is the incorporation of magnitudes and orientations of DPs around CCs and at cluster boundaries. In addition, the fuzzy label propagation approach is recommended for determining the number of clusters, outliers, and partitioned groups based on the fuzzy method. The uncertainty of the data point determines outliers. Finally, experiment datasets reveal that the proposed

method outperforms other recently developed clustering algorithms. FLPLRECM is a highly accurate and efficient clustering technique that does not necessitate user input. As a result, it's appropriate for a broad range of research and operational applications, including a cancer diagnosis.

References

[1] D. Jaeger, J. Barth, A. Niehues and C. Fufezan, Pygcluster, "A Novel Hierarchical Clustering Approach, " Bioinformatics, vol. 30, no.6, pp. 896-898, 2014.

[2] A. Dharmarajan and T. Velmurugan, "Applications of Partition-Based Clustering Algorithms: A Survey," *In 2013 IEEE International Conference on Computational Intelligence and Computing Research*, IEEE, pp.1-5, 2013.

[3] F. D. A. De Carvalho Y. Lechevallier and F. M. De Melo, "Partitioning Hard Clustering Algorithms Based on Multiple Dissimilarity Matrices, Pattern Recognition," vol.45, no.1, pp. 447-464, 2012.

[4] K. V. Rajkumar, A. Yesubabu and K. Subrahmanyam, "Fuzzy Clustering and Fuzzy C-Means Partition Cluster Analysis and Validation Studies on aSubset of Citescore Dataset," *International Journal of Electrical & Computer Engineering*, vol. 9, no.4, pp. 2088-8708, 2019.

[5] P. Lingras and G. Peters, "Applying Rough Set Concepts To Clustering, In Rough Sets: Selected Methods and Applications in Management and Engineering," Springer, London, pp.23-27, 2012.

[6] M. B. Ferraro and P. Giordani, "Possibilistic and Fuzzy Clustering Methods for Robust Analysis of Non-Precise Data," *International Journal of Approximate Reasoning*, vol.88, pp.23-38, 2017.

[7] T. Denœux and M. H. Masson, "Evclus: Evidential Clustering of Proximity Data," *IEEE Transactions on Systems, Man, and Cybernetics*, Part B (Cybernetics), vol.34, no.1, pp. 95-109, 2004.

[8] A. Rodriguez and A. Laio, "Clustering By Fast Search and Find of Density Peaks," *Science*, vol.344, no.6191, pp.1492-1496, 2014.

[9] F. Fang, L. Qiu and S. Yuan, "Adaptive Core Fusion-Based Density Peak Clustering for Complex Data with Arbitrary Shapes and Densities, *Pattern Recognition*," vol.107, pp.107452, 2020.

[10] J. Hou, A. Zhang and N. Qi, "Density Peak Clustering Based on Relative Density Relationship," *Pattern Recognition*, vol.108 , pp.107554, 2020.

[11] A. Lotfi, P. Moradi and H. Beigy, "Density Peaks Clustering Based on Density Backbone and Fuzzy Neighborhood," *Pattern Recognition*, vol. 107, pp.107449, 2020.

[12] Z. G. Su and T. Denoeux, "Bpec: Belief-Peaks Evidential Clustering," *IEEE Transactions on Fuzzy Systems*, vol. 27, no.1, pp. 111-123, 2018.

[13] G. Shafer, "A Mathematical Theory of Evidence, " *Princeton University Press*, 1976.

[14] Z. G. Liu, J. Dezert, Q. Pan and Y. M. Cheng, "A New Evidential C-Means Clustering Method," *In 2012 15th International Conference on Information Fusion*, IEEE, pp. 239-246, 2012.

[15] Y. Aboubi, H. Drias, and N. Kamel, "Bat-Clara: Bat-Inspired Algorithm for Clustering Large Applications, Ifac-Papersonline," vol. 49, no.12, pp.243-248, 2016.

[16] N. Kaur, and N. Ojha, "Robust Fuzzy Based Clustering Approach in Data Mining Using on Call Data Records," *In 2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, pp. 1111-1117, 2017.

[17] J. Chang, L. Wang, G. Meng, S. Xiang and C. Pan, "Deep Adaptive Image Clustering," *In Proceedings of the IEEE International Conference on Computer Vision*, pp. 5879-5887, 2017.

[18] L. Wang, X. Liu, M. Sun, J. Qu and Y. Wei, "A New Chaotic Starling Particle Swarm Optimization Algorithm for Clustering Problems," *Mathematical Problems In Engineering*, 2018.

[19] K. Zhou, A. Martin, Q. Pan and Z. Liu, Selp: "Semi-Supervised Evidential Label Propagation Algorithm for Graph Data Clustering," *International Journal of Approximate Reasoning*, vol.92 , pp.139-154, 2018.

[20] G. S. Narayana and D. Vasumathi, "An Attributes Similarity-Based K-Medoids Clustering Technique in Data Mining," *Arabian Journal for Science and Engineering*, vol.43, no.8, vol. 3979-3992, 2018.

[21] N. Pang, J. Zhang, C. Zhang and X. Qin, "Parallel Hierarchical Subspace Clustering of Categorical Data," *IEEE Transactions on Computers*, vol. 68, no.4, pp. 542-555, 2018.

[22] W. Budiaji and F. Leisch, "Simple K-Medoids Partitioning Algorithm for Mixed Variable Data," Algorithms, vol.12, no.9, pp.177, 2019.

[23] Y. Ping, B. Hao, H. Li, Y. Lai, C. Guo, H. Ma,.. and X. Hei, "Efficient Training Support Vector Clustering With Appropriate Boundary Information," *IEEE Access*, vol.7, pp.146964-146978, 2019.

[24] J. Meng, D. Fu and Y. Tang, "Belief-Peaks Clustering Based on Fuzzy Label Propagation," *Applied Intelligence*, vol.50, no.4, pp.1259-1271, 2020.

[25] S. Sieranoja and P. Fränti, "Fast and General Density Peaks Clustering," *Pattern Recognition Letters*, vol.128, pp.551-558, 2019.

[26] O. Chapelle, M. Chi and A. Zien, A, "A Continuation Method for Semi-Supervised Svms," *In Proceedings of the 23rd International Conference on Machine Learning*, pp.185-192, 2006.

[27] T. Yang, D. Fu and X. Li, "Semi-Supervised Classification of Multiple Kernels Embedding Manifold Information," *Cluster Computing*, vol.20, no.4, pp. 3417-3426, 2017.

[28] T. Denoeux and O. Kanjanatarakul, "Evidential Clustering: A Review, in Integrated Uncertainty in Knowledge Modelling and Decision Making - 5th International Symposium," Iukm 2016, pp.24–35, 2016.

[29] D. Liu, H. Y. Bai, H. J. Li and W. J. Wang, "Semi-Supervised Community Detection Using Label Propagation," *International Journal of Modern Physics B*, vol.28, no.29 , pp.1450208, 2014.

[30] J. Yu and S. B. Kim, "Consensus Rate-Based Label Propagation for Semi-Supervised Classification," *Information Science*, vol. 465, pp. 265 – 284, 2018.

[31] G. H. Golub and C. F, "Van Loan, Matrix Computations, Baltimore.," Md: Jhu Press, vol. 3 , 2012.