ISVS3CE: Incremental Support Vector Semi-Supervised Subspace Clustering Ensemble and ENhanced Bat Algorithm (ENBA) for High Dimensional Data Clustering

D. Karthika, K. Kalaiselvi

Abstract: ***In the recent work, Incremental Soft Subspace Based Semi-Supervised Ensemble Clustering (IS4EC) framework was proposed which helps in detecting clusters in the dataset. IS4EC framework also increases the results of clustering by reducing the intra-cluster distance and increasing the inter-cluster distance with increased cluster quality. It cannot attain acceptable results while handling high dimensional data. However, decreasing the dimensional subspace becomes extremely difficult issue. In IS4EC framework, to choose the optimal ensemble members also extremely becomes challenging issue. In order to solve these issues of traditional cluster ensemble methods, first propose an Incremental Support vector Semi-Supervised Subspace Clustering Ensemble (ISVS3CE) framework which makes utilized of benefits of the random subspace algorithm and the Constraint Propagation (CP) algorithm. Here the centroid values were selected by using the Support Vector Machine (SVM) classifier. In the ISVS3CE framework, Incremental Ensemble Member Chosen (IEMC) process is performed by using the ENhanced Bat Algorithm (ENBA), and the normalized cut algorithm is introduced to perform high dimensional data clustering. The ISVS3CE framework is successful for solving high dimensional data issue, at the same time as the CP algorithm is valuable for incorporating the prior information. Results demonstrate that the proposed ISVS3CE framework performs well on datasets by means of very high dimensionality, and better than the traditional clustering ensemble methods.***

Index Terms: ***Cluster ensemble, semi-supervised clustering, random subspace, cancer gene expression profile, clustering analysis, Support Vector Machine (SVM), and ENhanced Bat Algorithm (ENBA).***

Revised Manuscript Received on 30 July 2019.
***** Correspondence Author

D.Karthika*, Research Scholar , Department of Computer Science, VELS Institute of Science, Technology & Advanced Studies (Formerly VELS University), Chennai, India

Dr.K.Kalaiselvi, Professor& Head, Department of Computer Science, VELS Institute of Science, Technology & Advanced Studies (Formerly VELS University), Chennai, India

I. INTRODUCTION

Clustering is the method of segmenting the information depending on their similarity [1]. Clustering is utilized on various sectors like similarity search and web mining, segmentation, compression, pattern recognition, bioinformatics and classification. The clustering algorithms which were used in olden days, the full feature space of the information is considered. As the sets of data become bigger and tedious, hence the extractions of the current algorithm to enforce to maintain its quality and speed are necessary [2]. Real-time systems are defected from low efficiency because of the high-dimensional texture representation and the unconstrained optimization. Hence high dimensional clustering can enhance the execution time in those systems [3]. The dimensionality problem is one of the significant setbacks of clustering the data in high dimensional sets of data [4]. The dimensionality issue maximizes the cardinality of dimensions. Alternatively the dimensionality issue restricts all distances to make similar to the rendering of nearest neighbour information than high-dimensional data [5]. The challenges in dimension in clustering procedure can be taken from two views. Initially, certain attributes are utilized in order to identify the relevant information for a cluster is considered to be irrelevant. Hence considering the full feature space, the occurrence of those attributes will generate more complicated distance calculation for the process of clustering. For instance, certain clusters are placed in the subsets of certain attributes and impossible to correctly determine the cluster in the full feature space. Next, there may be certain variations among the subsets of attributes that defines a cluster and the ones define other clusters too. Hence a reduction property named global feature reduction method looks unfit in the identification of a subspace which has all the clusters. Also, it may be important to describe the clusters in overlapping method. A specific model in the clustering model may yield an oblige result for a specific output but becomes inefficient for others. Basically there exist two significant challenges adhere to clustering algorithms. Initially, various methods identify variety of structures for instance cluster size and their shape from the similar data set objects [6].

For instance, k-means that is a well-known technique which is best suited for spherical-shape clusters, while single-linkage hierarchical clustering seems to be efficient to identify the connected patterns. This is because of the fact that every single algorithm is proposed to optimize a particular criterion. Next, an individual clustering algorithm with variety of parameter settings can also show different structures on the similar set of data. A particular setting may be better for a few, but not in every sets of data. Users encounter these setbacks that continuously make the choice of a good clustering technique which is complex. Cluster ensembles have identified to be more accurate when compared to individual than single clustering algorithms [7]. More significantly, they restrict the user from taking decision on a specific clustering algorithm, thereby executing risk of a poor choice. During the classification, the sufficiency of the selected algorithm is clear from the predicted level of classifiers' accuracy in clustering process, a poor selection of algorithm may satisfy the entire work. Thus the general consensus looks to be a random selection of an ensemble which is not a dangerous option of an individual clustering process. Nowadays, cluster ensembles have developed as an optimal solution that is possible to overcome these problems and enhance the robustness of the quality of the results of clustering. The significant target of cluster ensembles is to merge variety of decisions of clustering in a way as to attain the superior level of accuracy than that of individual clustering. In recent days, cluster ensemble methods are attaining greater concentration [8], because of its applied areas of pattern recognition [9], data mining [10], bioinformatics [11], and so on. When correlated with olden days individual clustering algorithms, cluster ensemble methods are capable of integration in multiple clustering solutions gathered from various sources of information into a unified result, and render a more robust, constant and accurate final output. Traditional cluster ensemble methods have many disadvantages: (1) they don't take into consideration of using the basic knowledge provided by experts which is shown by pair related constraints. (2) Almost all the cluster ensemble techniques do not attain compromising outputs in large dimensional sets of data. To address the initial and secondary problems; Incremental Support vector Semi-Supervised Subspace Clustering Ensemble (ISVS3CE) framework is proposed which makes utilize of the benefit of the random subspace algorithm and the Constraint Propagation (CP) algorithm. Here the centroid values are selected by using the Support Vector Machine (SVM) classifier. The proposed work incremental ensemble member selection process was performed by using the ENhanced Bat Algorithm (ENBA), and the normalized cut algorithm is introduced for data clustering. Results demonstrate, that the proposed ISVS3CE framework performs well on datasets by means of very high dimensionality and better than the traditional clustering ensemble methods.

II. LITERATURE REVIEW

Amina [12] introduced a new clustering algorithm based on kernel mapping and hubness algorithm. This method identifies random shaped groups in the samples and also increases the result of clustering by reducing the inner-cluster distance and increasing the outer-cluster distance which enhances the group results. Ultsch and Loetsch[13] evaluated whether clustering can be able to be carryout by introducing Emergent Self-Organizing feature Maps (ESOM). Results of clustering were compared with traditional algorithms such as single linkage, Ward and k-means. The unsupervised based ESOM/U-matrix algorithm is a feasible, unbiased algorithm towards recognize accurate groups in the higher dimensionality of complex data. Alijamaat et al [14] introduced a new clustering for solving high dimensional space by including the object size which helps us to increase accuracy and effectiveness of traditional K-Means clustering. This algorithm is performed with two major steps. The data objects are grouped depending on their size, in fact by means of using subspaces for clustering. It causes obtaining more correct and proficient results. Assume with the purpose of the space is orthogonal and dimensions designed for each and every one objects are the identical and finally make use of time series data type since of the application. Sun et al [15] proposed a regularized k-means clustering algorithm for handling high-dimensional data issue, which is able to all together cluster related explanation and remove repeated variables. The aim is to design this clustering algorithm in a type of regularization by means of considering a lasso penalty term on cluster centers. Selection criterion is developed, to make a proper balancing among the clustering algorithm fitting and sparsity. The accuracy of this clustering is also measured via a diversity of statistical experiments with applications towards two gene microarray datasets. It is also extended to common model-based clustering framework.Dash et al [16] introduced to make use of Principal Component Analysis (PCA) algorithm for K-means clustering. They also PCA-K clustering to select the first centroids correctly to give further successful and higher clustering results. Results are compared to existing methods, it was demonstrated that the PCA-K means clustering are more efficient, with reduced processing time to perform clustering process.Ntoutsi et al [17] introduced a new density based projected clustering method for clustering of High Dimensional Data STREAMs (HDDSTREAM). This reviews both the samples and the dimensions where these samples are clustered simultaneously and continues these summaries online, as new samples turn up over time and old samples terminate suitable towards ageing. Results demonstrate the efficiency and the effectiveness of HDDSTREAM and also found so as to it might serve as a trigger for identifying extreme changes in the fundamental stream population, similar to bursts of network attacks. Fatehi et al [18] proposed a new subspace clustering algorithm in order to discover each and every cluster in all subspaces. Results on different synthetic and real datasets demonstrate with the purpose of the results of this method are considerably improved results in cluster quality and runtime when compared to other methods on high-dimensional data. Yu et al [19] proposed an incremental semi-supervised clustering ensemble (ISSCE) framework which combines the procedure of random subspace technique and the CP algorithm.

ISSCE framework, Incremental Ensemble Member Chosen (IEMC) procedure, and the normalized cut algorithm were introduced for grouping the high dimensional data. The IEMC procedure, normalized cut algorithm is used to provide the consensus function designed for giving more constant, and exact clustering results. Yu et al [20] proposed a new random transformation operator, which consists of the random mixture of transformation functions in the statistics dimension, the attribute dimension, and in concurrently size respectively. Here confidence assesses and the normalized cut algorithm is also introduced into the ensemble framework. Results show that this outperforms when compared to traditional transformation operators designed for various type of data samples such as gene expression and datasets from benchmark UCI repository.

III. PROPOSED METHODOLOGY

Incremental Support vector Semi-Supervised Subspace Clustering Ensemble framework (ISVS3CE) is proposed which make use of the incremental ensemble member selection step based on a local objective function and global objective function in the direction of create the ensemble set $E_s = \{f(X_1, A_1), (X_2, A_2), \dots. (X_{Y'}, A_{Y'})\}$ (where $Y' <$ Y) from the existing ensemble $\hat{E}_s$. There are two major reasons for performing ISVS3CE incremental procedure: (1) The overall objective function is helpful designed for managing global search, at the same time as the current objective function is suitable for local explore. The ISVS3CE incremental algorithm will improve the clustering results with achieve these search methods. (2) The local objective function takes into explanation the association of both subspaces and the objective function of the clustering results, which will improves the adaptively of the process. Particularly, known high dimensional dataset $FV = \{fv_1, \dots, fv_n\}$, with every feature vector fv_i (i ∈ fv_1,.......fv_n) containing m features, RSSCE make use of the random subspace algorithm to create a set of random subspaces X = (X_1,....X_Y) in the initial step. Particularly, a sampling rate $\tau \in [\tau_{min}, \tau_{max}]$ of the no. of features in the subspace over the total no. of features in the original space is first created. In the second step, the constraint propagation approach (CP) [21] is used to serve as the SSC model towards create a set of clustering results (fig1).

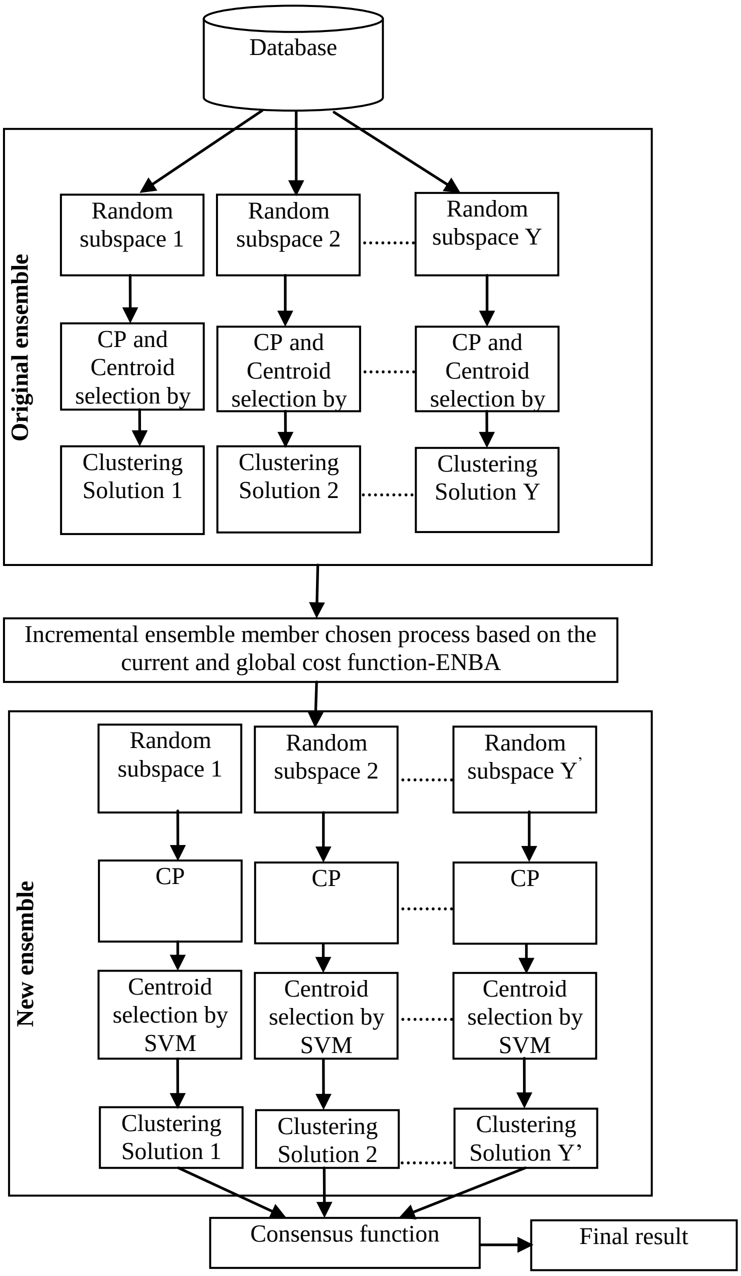


Fig. 1. Proposed Incremental Support vector Semi-Supervised Subspace Clustering Ensemble framework (ISVS3CE)

A. Semi-Supervised Subspace Clustering Ensemble framework (S3CE)

CP regards as a restricted number of must-link and cannot link constraints among pairs of feature vectors known by experts, and decomposes a constraint propagation issue into a set of SSC issues. The dataset FV = $\{fv_1, \dots, fv_n\}$be able to be formulated by an undirected weighted graph G(FV,A), where FV is denoted as the set of feature vectors related to the vertices, and A is described as the similarity matrix through a weight value a_{ij} related by means of the edge among fv_i and fv_j :

$$a_{ij} = \begin{cases} \rho_{ij} & if\ fv_i \in \mathrm{N}_c(fv_j) \\ 0\ else \end{cases} \quad (1)$$

$$\rho_{ij} = \exp\left(\frac{-\left\|fv_i - fv_j\right\|^2}{dis^2}\right) \quad (2)$$

where ‘dis’ is denoted as the set to the common Euclidean distance among each and every one of data points of c-nearest neighbors, and $N_c(fv_i)$ is described as the c-nearest neighbor set used for fv_i. Known a set of initial must-link constraints M = (fv_i, fv_j) : $b_i = b_j, 1 \leq i; j \leq n)$ and a set of original cannot-link constraints N = (fv_i, fv_j) : $b_i \neq b_j, 1 \leq i; j \leq n$ (where b_i is denoted as the label of the feature vector fv_i), the constraint matrix R = $(r_{ij})_{nxn}$ be able to be formulated as follows:

$$R = \begin{cases} +1 & (fv_i, fv_j) \in M \\ -1 & (fv_i, fv_j) \in N \\ 0\ else \end{cases} \quad (3)$$

The aim of CP is towards propagate supervised data from labeled to unlabeled, which be able to be handled via label propagation depending on c-nearest neighbour graphs. The CP algorithm is able to consider as a semi-supervised binary classification issue according to the fv_j. supervised binary classification issue according to the fv_j. This SSC issue with related to fv_j in the vertical direction be able to be created as reducing a Laplacian regularized objective function

$$= \min_F \frac{1}{2}\left\|F - R\right\|_2^2 + \frac{\mu}{2} F^T LF \quad (4)$$

where L is denoted as the Laplacian matrix. In the equation (4), where c is defined as the centroid points of the cluster, which are chosen depending on the Support Vector Machine classifier. which are chosen depending on the Support Vector Machine classifier.

B. Support Vector Machine(SVM) classifier

A classification task generally involves dividing samples into training and testing sets. In this work every nearest points in the cluster is chosen as training set includes one “target value” (i.e. the distance value) and many “feature vectors” (i.e. the features). The aim of SVM [22] is to design a new classifier which determines the distance value of the test sample by considering only the feature vectors.

Known a training set of feature vectors FV = $\{fv_1, \dots, fv_n\}$ (fv_i,c_i),i= 1,...,n where $fv_i \in R^n$ and $c \in \{1, -1\}^l$, the SVMs need the solution of the following subsequent problem

$$= \min_{we,bi,\xi} \frac{1}{2}\left\|we^T we\right\|_2^2 + C\sum_{i=1}^{n} \xi_i \ subjet\ to\ c_i(we^T \phi(fv_i) + b_i) \geq 1 - \xi_i, \xi_i \geq 0 \quad (5)$$

where we, b_i is denoted as the weight and biases of the SVM model. The training vectors fv_i are mapped into a feature space via the variable φ. C >0 is denoted as the penalty parameter for computing the error value. Moreover, $K(fv_i,fv_j) \equiv \varphi(fv_i)^T \varphi(fv_j)$ is named as the kernel function. Selected centroid values then Incremental Ensemble Member Chosen (IEMC) process is carryout by using the Enhanced Bat Optimization (EBO) algorithm.

C. Incremental Ensemble Member Chosen (IEMC)

The contribution is the new ensemble, at the same time as the output is a newly created ensemble with smaller size. Particularly, IEMC includes the ensemble members one by one, and determines the objective function Lo_y designed for each clustering results I_y created with CP via respect towards the subspace X_y in the initial step. In the second step, it sorts each and every one the ensemble members in $\hat{E}_s$ in ascending order related to the resultant Δ values. The initial ensemble member (X_t, A_t) (where t = 1) is collected and inserted into the new ensemble E = (X_t, A_t). At the similar point, (X_t, A_t) is removed from the present ensemble $\hat{E}_s$ as follows

$$\hat{E}_s = \hat{E}_s \backslash \{(X_t, A_t)\} \quad (6)$$

Each ensemble member(X_y, A_y)in $\hat{E}$ is measured in turn, and the local objective function Lo_y by means of respect towards the ensemble member (X_t, A_t) in E_s is determined. Each and every one the ensemble members in $\hat{E}_s$ are sorted in ascending order related to the correct local objective function Lo_y in the fourth step. Determines the global objective functions $\Delta(I')$ and $\Delta(I)$designed for the clustering results I' and I created by E_s' and E_s, correspondingly. **IEMC** procedure make use of the global objective function Δ and the local objective function Lo_y.let us assume with the purpose of (1) M and N are the sets of must-link and cannot-link pairs, correspondingly. (2) $We^M = (we_{ij}^M)$ and $We^N = (we_{ij}^N)$ represents the weight sets which include weights related towards the must-link and cannot-link constraints, correspondingly. (3) $B = (b_1, \dots b_n)$ Is denoted as the set of labels designed for the feature vectors in FV. (4) The cost of violating constraints is particular with the generalized Potts metric. Known a clustering solution I, the global objective function Δ(I) of IEMS, which is stimulated from the cost function of Pairwise Constrained- K means(PC-K means) clustering [23], is described as follows, (7)

$$\Delta(I) = \frac{1}{2}\sum_{fv_i \in FV}\sum_{h=1}^{k} \theta(b_i = h)\, dis(fv_i, c_h) + \sum_{fv_i, fv_j \in M} we_{ij}^M \theta(b_i \neq b_j) + \sum_{fv_i, fv_j \in N} we_{ij}^N \theta(b_i = b_j)$$

where $dis(fv_i, c_h)$ is defined as the Euclidean distance among the attribute vectors fv_i and c_h, θ is defined an indicator function, θ (positive) = 1 and θ (negative) = 0. The local objective function of Lo_y is described as follows,

$$Lo_y = \sum_{\forall X_t \in E_s} \frac{S(X_y, X_t)}{\Delta(I^y)} \quad (8)$$

where $\Delta(I^y)$ is represented as the global objective function designed for the clustering results I^y, and S(X_y, X_t) is represented the similarity function among two subspaces X_y (5)and X_t calculated by distance function.

D. Enhanced Bat Algorithm (ENBA)

The function of (7) and (8) is performed via the use of Enhanced Bat Algorithm (ENBA). BA is performed with group of bats with echolocation towards intellect for difficulty and preys. In this work bat is used to collect the information of the ensemble members of clusters for ISVS3CE clustering. To make simpler the ensemble member chosen step of the ISVS3CE algorithm, the characteristics of bat sound have been described as given below [24]:

1. Each and every one of the bats make use of sound to sense distance, to determine the distance between subspaces and they also ‘distinguish’ the difference between similarity and dissimilarity objects in some manner;
2. Bats fly indiscriminately with velocity vel_i at ensemble member location mp_i with a frequency fre_{min}, modifying the wavelength λ and loudness L_0 towards investigates of optimal member collection for clustering.They are capable to routinely regulate the wavelength of their emitted pulses and adjust the pulse emission rate.
$er \in [0,1]$, depending on the proximity of their local and global cluster function;
3. It is presumed with the purpose of the loudness changes from a large L_0 to a lesser constant value L_{min}.

It is carryout depending on two major steps: the initial step starts entire the cluster members for clustering, as each bat should be described by a location p_i^t , velocity vel_i^t , emission pulse rate er_i^t , loudness L_i^t , and frequency fre_i^t , in the solution space at iteration t. The population of bats is defined randomly where each bat denotes a possible ensemble member chosen to the ISVS3CE clustering issue. The second step of the procedure consists of the creation of new element by applying the alterations defined with the subsequent equations (9)

$$fre_i = fre_{min} + (fre_{max} - fre_{min})\beta \tag{9}$$

where $\beta \in [0,1]$ is denoted as the random vector created from a uniform distribution

$$vel_i^t = vel_i^{t-1} + (mp_i^t - mp^*)fre_i \tag{10}$$

where mp^* is denoted as the current global best member position, which is located after comparing each and every one the cluster members among all the datapoints(bats).

$$mp_i^t = mp_i^{t-1} + mp_i^t \tag{11}$$

Depending on the issue, the frequency ‘fre’ is defined to $fre_{min} = 0$ and $fre_{max} = 100$. Firstly, every bat is start with a known frequency which is created uniformly from $[fre_{min}, fre_{max}]$. For the local explore of member chosen, once a clustering solution is chosen between the present best local and global member procedure, a new member designed for each bat is found locally with random walk where $\varepsilon \in [-1,1]$ is a scaling factor, at the same time as $L^t = < L_i^t >$ is the standard loudness of each and every one the bats at time step t.

$$mp_{new} = mp_{old} + \varepsilon L^t \tag{12}$$

Moreover, the loudness L_i and the rate er_i of pulse emission are changed depending on the following equations:

$$L_i^{t+1} = \alpha L_i^t \tag{13}$$

$$er_i^{t+1} = er_i^0[1 - \exp(-\gamma t)] \tag{14}$$

where α and γ are denoted as the constants. The major modification of the proposed BA algorithm is to begin a fixed loudness A, exchange of a variety of loudness L_i^t. In adding together it moreover includes a mutation operator in order towards increase the range of the population towards improves the investigate success of member selection procedure. This operator in the ENBA algorithm gives a new pair of optimizing of Lim_1 and Lim_2 parameters, depending on the recent work [25-26]. In this stage, if a random value is lower when compared to Lim_1, then a result mp_{mv}^t is randomly choose from the population as shown in Eq. (16).

$$mv_r = rand * NP \tag{15}$$

$$mp_v^t = mp_{v_r}^t \tag{16}$$

where the r ∈ (1, 2… NP); More, if a random value is lower than Lim_2, is introduced into the steps of the current member chosen results, drawing the exploration back to a increased position between to the top and worst results established as a result far. The vary of the mutation operator is defined in Eqs. (17) and (18).

$$mp_v^t = 0.5 \times (mp_{worst}^t - mp_v^t) \times rand\,[0.1) \tag{17}$$

$$mp_v^t = 0.5 \times (mp_v^t - mp_{best}^t) \times rand\,[0.1) \tag{18}$$

The equations (17-18), where mp_v^t is defined as the new cluster ensemble member results belongs to t^th iteration; mp_v^t is defined as the random solution selection from equation Eq. (16),and mp_{worst}^t , mp_{best}^t is described as the worst and optimal solutions, respectively. Otherwise, the randomization rule relates to insert population diversity, significant to increase the probability of discovering the global clustering results. Consequently, the randomization rule generates a new range for the i^th example in the unit mp_v^t as described in Eq. (19).

$$mp_v^t = Lb + rand\,(Ub - lb) \tag{19}$$

An n×n consensus matrix ‘O’ is then formulated by merging the entire adjacency matrices (O^1,O^2, ,...,O^Y) as follows:

$$O = \frac{1}{Y}\sum_{y=1}^{Y} O^Y \tag{20}$$

Normalized cut algorithm (Ncut) [27] is used as the consensus function towards the partition of feature vector set ‘FV’ depending on the matrix O.

Proposed algorithm
Incremental Support vector Semi-Supervised Subspace Clustering Ensemble framework (ISVS3CE)
Input: *a high dimensional datasetFV;*
Ensure:

1. *Original ensemble generation*
2. *Create Y random subspaces (X_1,X_2,....,X_Y)*
3. *Create the semi-supervised clustering models $A_1, \dots A_Y$ using CP*
4. *Call incremental ensemble member chosen process in ENBA*
5. *New ensemble creation*
6. *Create Y' random subspaces (X_1;X_2,...,$X_{Y'}$) (Y' < Y)*
7. *Create SSC models (X_1;X_2,...,$X_{Y'}$) using CP;*
8. *Obtain consensus matrix O by adding the clustering solutions (I_1, I_2,,..., $I_{Y'}$) created by the SSC methods;*
9. *Consensus functions for the clustering results by the normalized cut;*

Output: *the labels of the samples in FV.*

IV. EXPERIMENT AND RESULTS

The results of ISVS3CE, ISSCE and SSCE are measured using datasets as shown in Table 1 (where ‘n’ represents the amount of data samples, ‘m’ represents the amount of features, and k represents the amount of classes), which consists of 3 datasets from UCI machine learning repository .

Table 1. Dataset samples

Dataset	n	m	K
Libras Movement	360	90	15
RobotExecution1	89	90	4
Breast Cancer Wisconsin	569	32	2

Normalized Mutual Information (NMI)[28] and Adjusted Rand Index (ARI) [28] are used to measure the performances of various clustering methods. The result of the proposed IS^{4}EC clustering algorithm is assessed by the typical value and the related standard deviation of NMI and ARI, correspondingly, subsequent to 20 runs. Known the ground truth result I by means of k clusters $I = \{C_1, C_2, .. C_k\}$ and the result I' obtained by ISSCE with k$_0$ clusters $I' = \{C'_1, \dots C'_k\}$, make use of NMI[28] in the direction of calculate the value of the final clustering result, which is described as follows:

$$NMI(I, I') = \frac{2H_1(I, I')}{H_2(I) + H_2(I')} \quad (21)$$

$$H_1(I, I') = \sum_h \sum_l \frac{|c_h \cap c'_l|}{n} \log \frac{n|c_h \cap c'_l|}{|c_h||c'_l|} \quad (22)$$

$$H_2(I) = \sum_h \frac{|c_h|}{n} \log \frac{n|c_h|}{n} \quad (23)$$

$$H_2(I') = -\sum_h \frac{|c'_l|}{n} \log \frac{n|c'_l|}{n} \quad (24)$$

where $h \in \{1, \dots k\}, l \in \{1, \dots k'\}$, n is the full amount number of data samples, and $|.|$ represents the cardinality of the cluster. The adjusted Rand index ARI(I,I') [28] is described as follows:

$$ARI(I, I') = \frac{\sum_{h=1}^{k} \sum_{l=1}^{k'} \binom{|c_h \cap c'_l|}{2} - p_3}{\frac{1}{2}(p_1 + p_2) - p_3} \quad (25)$$

$$p_1 = \sum_{h=1}^{k} \binom{|C_h|}{2} \quad (26)$$

$$p_2 = \sum_{h=1}^{k'} \binom{|C'_l|}{2} \quad (27)$$

$$p_3 = \frac{2p_1p_2}{n(n-1)} \quad (28)$$

$$Recall(i, j) = N_{ij}/N_i \quad (29)$$

$$Precision(i, j) = N_{ij}/N_j \quad (30)$$

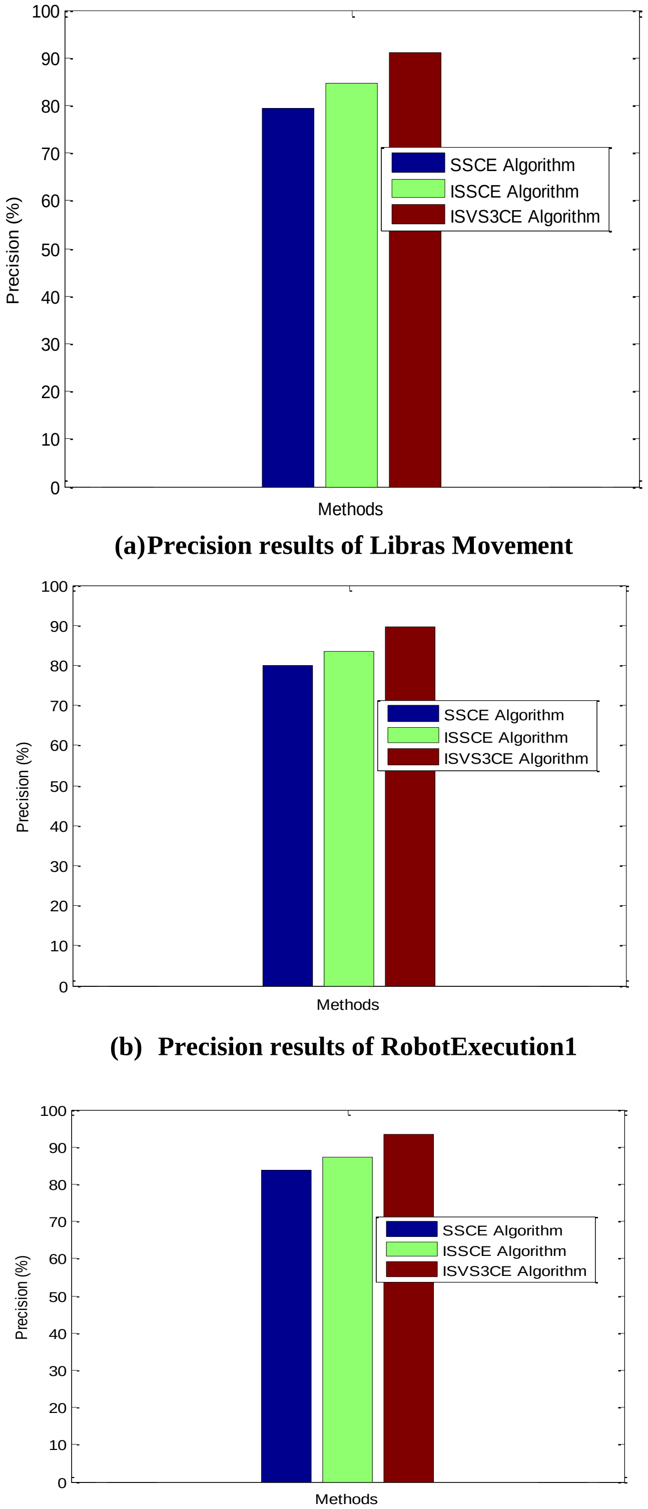
where N$_j$ is denoted as the no. of parts of cluster j and N$_i$ is denoted as the no. of parts of class i, N$_{ij}$ is denoted as the no. of parts of class i in cluster j. F-measure is defined as the harmonic mean of recall and precision. It is calculated by using the following the functions

$$F(i, j) = \frac{2.(Recall(i,j) * Precision(i,j))}{Recall(i,j) + Precision(i,j)} \quad (31)$$

A larger ARI value relates to a higher clustering solution quality. In the subsequent experiments, we primary learning the result of the parameters. After that, the effect of the incremental ensemble member chosen procedure is explored. Subsequently, the proposed ISVS3CE clustering algorithm is compared with single ISSCE, and SSCE, on the datasets. At last, a set of nonparametric tests are implemented in the direction of evaluate numerous SSCE methods over different datasets. Table 2 shows the clustering results obtained by various clustering methods related to various metrics on each and every one the datasets.

Table 2. Performance comparison metrics vs. clustering methods

Dataset	Metrics	SSCE	ISSCE	ISVS3CE
Libras Movement	**Precision (%)**	79.3224	84.7323	91.2007
	Recall(%)	79.1667	84.4444	91.1111
	F-measure(%)	79.2445	84.5881	91.1559
	Accuracy (%)	79.1667	84.4444	91.1111
	ARI	0.6027	0.6898	0.8175
	NMI	0.7110	0.7746	0.8675
RobotExecution1	**Precision (%)**	79.8818	83.4749	89.6174
	Recall(%)	80.5964	84.0926	90.2982
	F-measure(%)	80.2375	83.7827	89.9565
	Accuracy (%)	80.5982	84.2748	90.1976
	ARI	0.5474	0.6287	0.7563
	NMI	0.4915	0.5687	0.6995
Breast Cancer Wisconsin	**Precision (%)**	83.8623	87.4038	93.4950
	Recall(%)	83.8060	87.7162	93.0900
	F-measure(%)	83.8342	87.5597	93.2921
	Accuracy (%)	84.5725	88.1387	93.6823
	ARI	0.4080	0.5227	0.7084
	NMI	0.3440	0.4426	0.6291



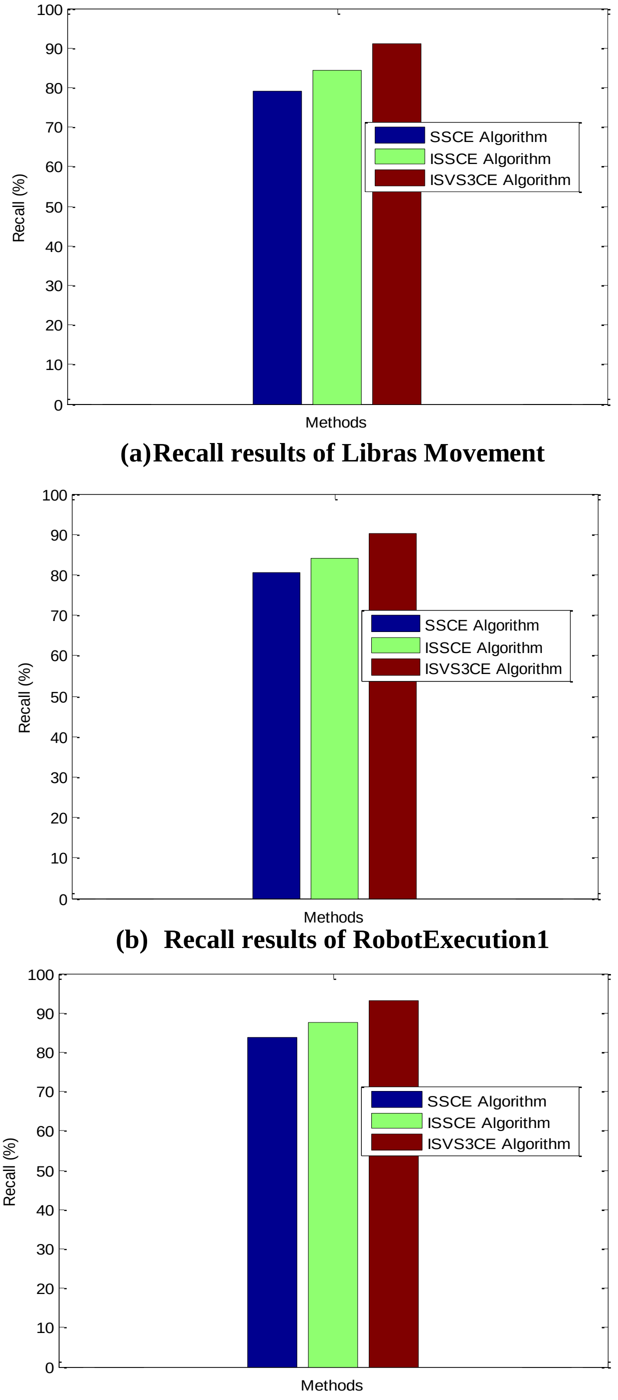
(a) Precision results of Libras Movement

(b) Precision results of RobotExecution1

(c) Precision results of Breast Cancer Wisconsin

Fig. 2. Precision comparison vs. many clustering algorithms

Fig. 2(a-c) shows the performance comparison results of precision with respect to several methods such as SSCE, ISSCE and ISVS3CE in terms of three datasets such as Libras Movement, RobotExecution1 and Breast Cancer Wisconsin. The proposed ISVS3CE algorithm provides higher precision results of 91.2007% for Libras Movement dataset whereas other clustering methods such as SSCE and ISSCE provides precision results of 79.3224%, and 84.7323% methods respectively.



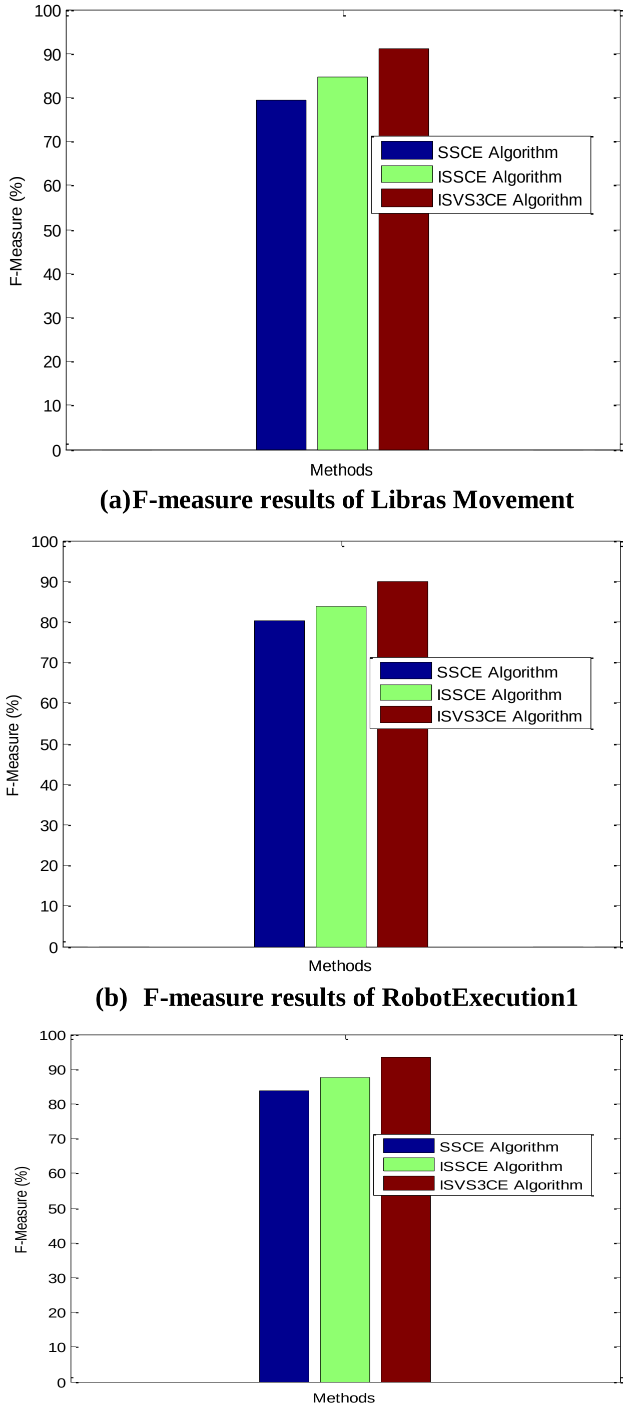
(a) Recall results of Libras Movement

(b) Recall results of RobotExecution1

(c) Recall results of Breast Cancer Wisconsin

Fig. 3. Recall comparison vs. many clustering algorithms

Fig. 3(a-c) shows the performance comparison results of recall with respect to several methods such as SSCE, ISSCE and ISVS3CE in terms of three datasets such as Libras Movement, RobotExecution1 and Breast Cancer Wisconsin. The proposed ISVS3CE algorithm provides higher recall results of 91.1111% for Libras Movement dataset whereas other clustering methods such as SSCE and ISSCE provides recall results of 79.1667%, and 84.4444% methods respectively.



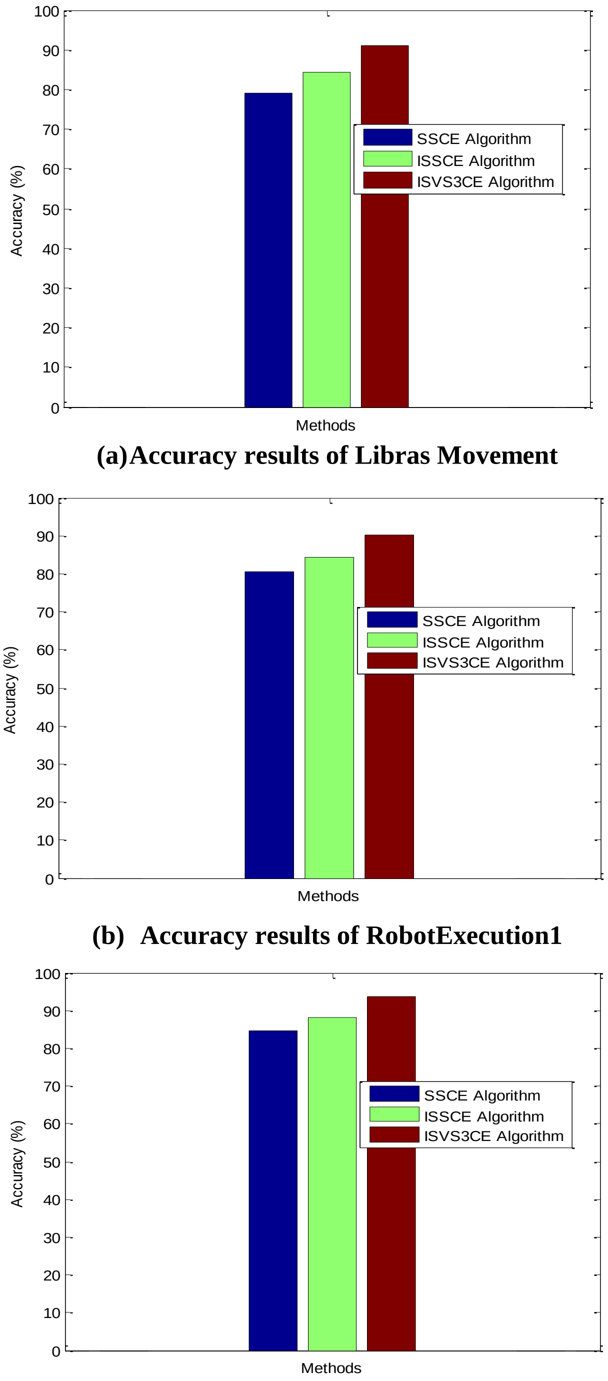
(a) F-measure results of Libras Movement

(b) F-measure results of RobotExecution1

(c) F-measure results of Breast Cancer Wisconsin

Fig. 4. F-measure comparison vs. many clustering algorithms

Fig. 4(a-c) shows the performance comparison results of f-measure with respect to several methods such as SSCE, ISSCE and ISVS3CE in terms of three datasets such as Libras Movement, RobotExecution1 and Breast Cancer Wisconsin. The proposed ISVS3CE algorithm provides higher f-measure results of 91.1559% for Libras Movement dataset whereas other clustering methods such as SSCE and ISSCE provides f-measure results of 79.2445%, and 84.5881% methods respectively.



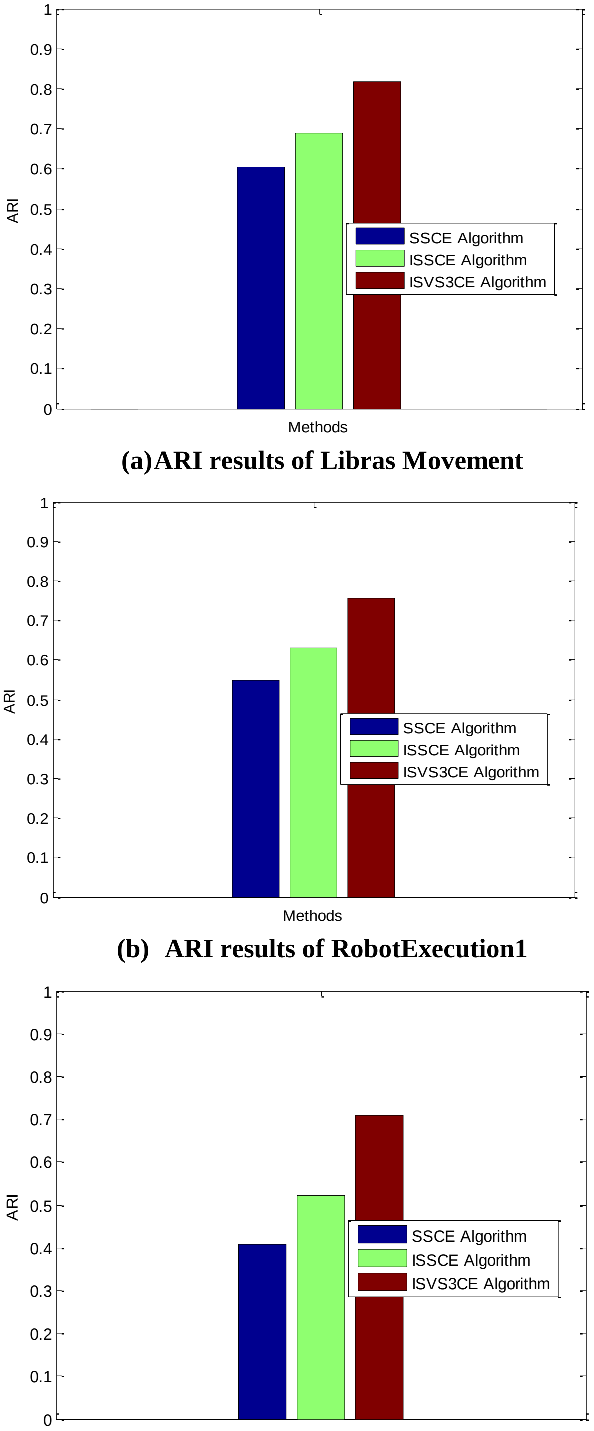
(a) Accuracy results of Libras Movement

(b) Accuracy results of RobotExecution1

(c) Accuracy results of Breast Cancer Wisconsin

Fig. 5. Accuracy comparison vs. many clustering algorithms

Accuracy comparison results with respect to several methods such as SSCE, ISSCE and ISVS3CE in terms of three datasets such as Libras Movement, RobotExecution1 and Breast Cancer Wisconsin are shown in fig. 5(a-c). From the fig. 5(a) it concludes that the proposed ISVS3CE algorithm provides higher accuracy results of 91.1111% for Libras Movement dataset whereas other clustering methods such as SSCE and ISSCE provides accuracy results of 79.1667%, and 84.4444% methods respectively.



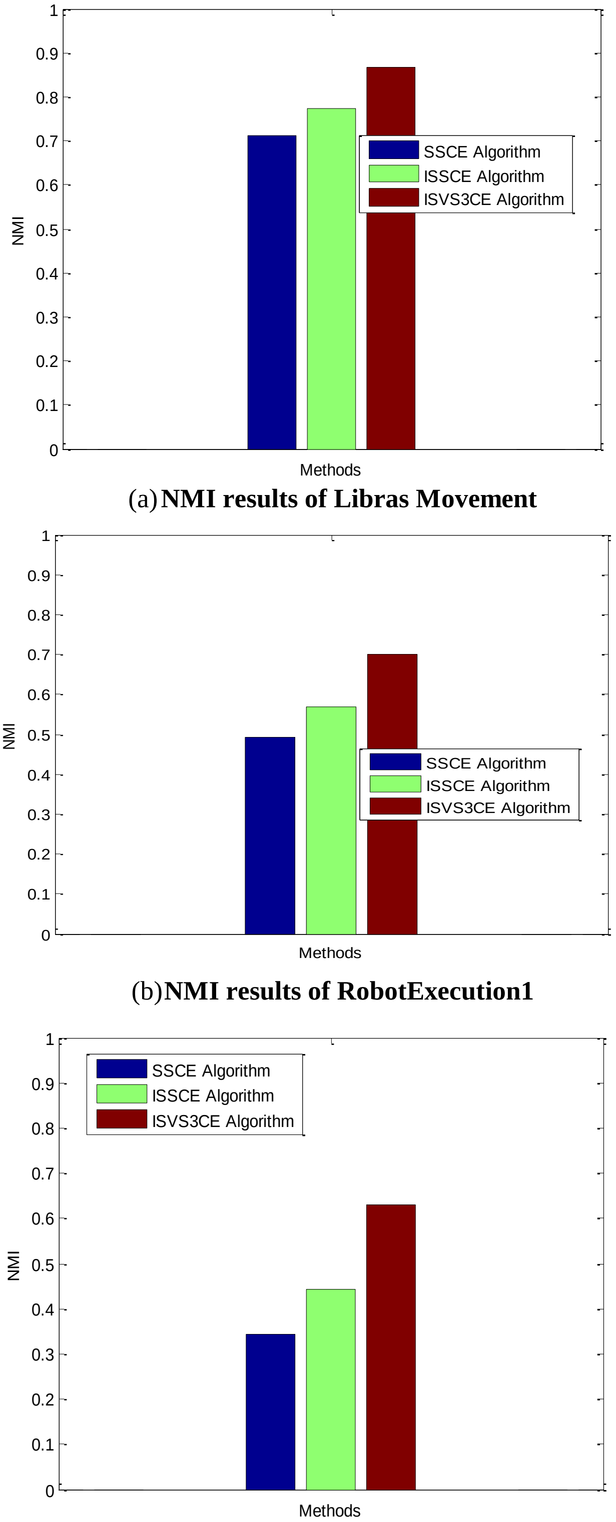
(a)ARI results of Libras Movement

(b) ARI results of RobotExecution1

(c)ARI results of Breast Cancer Wisconsin

Fig. 6. ARI comparison vs. many clustering algorithms

ARI comparison results with respect to three datasets such as Libras Movement, RobotExecution1 and Breast Cancer Wisconsin for several methods such as SSCE, ISSCE and ISVS3CE are shown in fig. 6(a-c). From the fig. 6(a) it concludes that the proposed ISVS3CE algorithm provides higher ARI results of 0.8175 ,whereas SSCE and ISSCE clustering methods provides ARI results of 0.6027, and 0.6898 respectively for Libras Movement dataset.



(a)**NMI results of Libras Movement**

(b)**NMI results of RobotExecution1**

(c)**NMI results of Breast Cancer Wisconsin**

Fig. 7. NMI comparison vs. many clustering algorithms

NMI comparison results with respect to Libras Movement, RobotExecution1 and Breast Cancer Wisconsin datasets for clustering methods such as SSCE, ISSCE and ISVS3CE are shown in fig. 7(a-c). From the fig. 7(a) it concludes that the proposed ISVS3CE algorithm gives higher NMI results of 0.8675,whereas SSCE and ISSCE clustering methods gives lesser NMI results of 0.7110, and 0.7746 for Libras Movement dataset.

V. CONCLUSION

In this paper, we semi-supervised clustering ensemble approach is proposed, which was called as the Incremental Support Vector Semi-Supervised Subspace Clustering Ensemble framework (ISVS3CE).It includes of two major steps: ensemble creation and new ensemble creation. In this initial step Ensemble creation was followed by using random subspace clustering algorithm. In the second step, some changes are carryout in this work. In second step firstly random subspace clustering is followed by Constraint Propagation (CP) approach. Then new optimal Centroid chosen was performed by support vector machine (SVM) classifier. From then Incremental Ensemble Member Chosen (IEMC) is performed by ENhanced Bat Algorithm (ENBA).Here the global objective function is useful for performing global search, at the same time as the local objective function is useful for local search. The similarity function is used to calculate the amount to which two sets of features in the subspaces are comparable to each other. Experimentation is conducted on three datasets from the benchmark UCI repository. Future work will focus on how to choose parameter changes based on the structure of the datasets.

REFERENCES

1. Bouveyron, C. and Brunet-Saumard, C., 2014. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, *71*, pp.52-78.
2. McWilliams, B., Heinze, C., Meinshausen, N., Krummenacher, G. and Vanchinathan, H.P., 2014. LOCO: Distributing ridge regression with random projections. stat, 1050, p.26.
3. Chen, Y., Tang, S., Bouguila, N., Wang, C., Du, J. and Li, H., 2018. A Fast Clustering Algorithm based on pruning unnecessary distance computations in DBSCAN for High-Dimensional Data. *Pattern Recognition,pp.1-42*.
4. Zakerzadeh, H., Aggarwal, C.C. and Barker, K., 2014, Towards breaking the curse of dimensionality for high-dimensional privacy. In *Proceedings of the 2014 SIAM International Conference on Data Mining* (pp. 731-739). Society for Industrial and Applied Mathematics.
5. Esmin, A. A., Coelho, R. A., & Matwin, S. (2015).A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data. Artificial Intelligence Review, 44, 23-45.
6. D. Greene, A. Tsymbal, N. Bolshakova, and P. Cunningham. Ensemble clustering in medical diagnostics. Technical Report TCD-CS-2004-12, Department of Computer Science, Trinity College, Dublin, Ireland, 2004.
7. H. Xue, S. Chen, Q. Yang, Discriminatively regularized least-squares classification, Pattern Recognit. 42 (2009) 93–104.
8. H.G. Ayad, M.S. Kamel, “Cumulative Voting Consensus Method for Partitions with Variable Number of Clusters”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 30, Issue 1, pp.16–173, 2008.
9. N. Iam-On, T. Boongoen, S. Garrett, C. Price, “A Link-Based Approach to the Cluster Ensemble Problem”, IEEE Transactions on Pattern Analysis and Machine Intelligence , vol. 33, no. 12, pp. 2396-2409, 2011.
10. N. Iam-On, T. Boongoen, S. Garrett, C. Price, “A Link-Based cluster ensemble approach for categorical data clustering”, IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 3, pp. 413-425, 2012.
11. Zhiwen Yu, Hantao Chen, Jane You, Guoqiang Han, Le Li, ”Hybrid Fuzzy Cluster Ensemble Framework for Tumor Clustering from Biomolecular Data”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 10, no. 3, pp. 657-670, 2013.
12. Amina, M., 2015, A Novel Approach for Clustering High Dimensional Data Using Kernal Hubness. Fifth International Conference on Advances in Computing and Communications (ICACC), pp. 94-97.
13. Ultsch, A. and Loetsch, J., 2017. Machine-learned cluster identification in high-dimensional data. Journal of biomedical informatics, 66, pp.95-104.
14. Alijamaat, A., Khalilian, M. and Mustapha, N., 2010, A novel approach for high dimensional data clustering. Third International Conference on Knowledge Discovery and Data Mining, pp. 264-267.
15. Sun, W., Wang, J. and Fang, Y., 2012. Regularized k-means clustering of high-dimensional data and its asymptotic consistency. Electronic Journal of Statistics, 6, pp.148-167.
16. Dash, B., Mishra, D., Rath, A. and Acharya, M., 2010. A hybridized K-means clustering approach for high dimensional dataset. International Journal of Engineering, Science and Technology, 2(2), pp.59-66.
17. Ntoutsi, I., Zimek, A., Palpanas, T., Kröger, P. and Kriegel, H.P., 2012, April. Density-based projected clustering over high dimensional data streams. In Proceedings of the 2012 SIAM International Conference on Data Mining (pp. 987-998). Society for Industrial and Applied Mathematics.
18. Fatehi, K., Rezvani, M., Fateh, M. and Pajoohan, M.R., 2018. Subspace Clustering for High-Dimensional Data Using Cluster Structure Similarity. International Journal of Intelligent Information Technologies (IJIIT), 14(3), pp.38-55.
19. Yu, Z., Luo, P., You, J., Wong, H.S., Leung, H., Wu, S., Zhang, J. and Han, G., 2016. Incremental semi-supervised clustering ensemble for high dimensional data clustering. IEEE Transactions on Knowledge and Data Engineering, 28(3), pp.701-714.
20. Z. Yu, H.-S. Wong, J. You, G. Yu, G. Han, “Hybrid Cluster Ensemble Framework based on the Random Combination of Data Transformation Operators”, Pattern Recognition, vol. 45, no. 5, pp. 1826-1837, 2012.
21. Z. Lu, Y. Peng, “Exhaustive and Efficient Constraint Propagation: A Graph-Based Learning Approach and Its Applications”, International Journal of Computer Vision, vol. 103, no. 3, pp. 306-325, 2013.
22. Widodo, A. and Yang, B.S., 2007. Support vector machine in machine condition monitoring and fault diagnosis. Mechanical systems and signal processing, 21(6), pp.2560-2574.
23. S. Basu, A. Banjeree, E.R. Mooney, A. Banerjee, R.J. Mooney, “Active Semi-Supervision for Pairwise Constrained Clustering”, Proceedings of the Fourth SIAM International Conference on Data Mining ,pp. 333-344, 2004.
24. Yang X-S (2010) A new metaheuristic bat-inspired algorithm, in Nature inspired cooperative strategies for optimization (NICSO 2010). Springer, pp. 65–74
25. GhanemWAHM, Jantan A (2016) Hybridizing artificial bee colony with monarch butterfly optimization for numerical optimization problems. Neural Comput Applic :1–19.
26. Wang G, Guo L (2013) A novel hybrid bat algorithm with harmony search for global numerical optimization. J Appl Math 2013.
27. Huang, S.H., Chu, Y.H., Lai, S.H. and Novak, C.L., 2009. Learning-based vertebra detection and iterative normalized-cut segmentation for spinal MRI. IEEE transactions on medical imaging, 28(10), pp.1595-1605.
28. Vinh N.X., J. Epps, J. Bailey, “Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance”, The Journal of Machine Learning Research, vol. 11, pp. 2837-2854, 2010.
29. Dr.K.Kalaiselvi and D.Karthika, “Review of Traditional and Ensemble Clustering Algorithms for High Dimensional Data”, (April 28, 2018). Available at Elsevier SSRN: https://ssrn.com/abstract=3170321
30. Dr.K.Kalaiselvi and D.Karthika, “IS4EC: Incremental Soft Subspace Based Semi-Supervised Ensemble For High Dimensional Data”, International Journal of Advanced Studies of Scientific Research, Vol. 3, No. 10, 2018.

AUTHORS PROFILE

D.Karthika is a Research scholar in the Department of Computer Science, VELS Institute of Science, Technology & Advanced Studies (Formerly VELS University), Chennai, India.Interested in Machine Learning , Artificial Intelligence and Big Data Analytics.Broad area of research is Big Data Analytics focusing on High Dimensional Data Clustering.

Dr.K.Kalaiselvi is a Professor & Head, Department of Computer Science, VELS Institute of Science, Technology & Advanced Studies (Formerly VELS University), Chennai, India. She does research in Information Science, Artificial Neural Network and Artificial Intelligence.She is also excellence in Machine Learning and Big Data Analytics.She is also one of the Editorial Member in academic Journals.