Improved Unsupervised Multi-level Clustering for Segmentation of Remote Sensing Imageries Containing Embedded Objects

Ved Prakash Singh[1,2][0000-0002-2281-5687], Jimson Mathew[2], Kevin Patel[3], and Sunil Kumar Patel[4][0000-0002-9821-0852]

[1] India Meteorological Department, Ministry of Earth Sciences, Bhopal, Madhya Pradesh - 462011, India
kvpssc@gmail.com

[2] Department of Computer Science and Engineering, Indian Institute of Technology, Patna, Bihar - 801106, India
mathew.jimson@gmail.com

[3] Department of Computer Science and Engineering, Pandit Deendayal Energy University, Gujarat - 382007, India
kevin18patel@gmail.com

[4] Department of Computer Science and Engineering, National Institute of Technology Durgapur, West Bengal - 713209, India
sunilpatel.bsb@gmail.com

Abstract. In recent times, clustering in geospatial imgeries has become an indispensable part of many research projects and plays an important role in studying the object evolutions and movements in the geospatial (GS) domain. Usually, geospatial datasets are captured by satellite and Radar remote sensing, which are visualized on a geospatial platform (GIS) as an imagery i.e. colour values of pixels of an image contain the domain-parametric information. Some of the popular problems in the GS domain are identification of weather bearing clouds in Radar images, separation of natural elements in earth surface images, *etc.* Segmentation of such imageries is still a complicated task as the presence of high-resolution images with the erratic variation from one image to another, exhibit the unlabeled objects of irregular shape. Existing arts identify the suitable algorithms for clustering of overlapped objects based on a single feature-set or attribute-set, where DB-scan and K-means are found as best unsupervised algorithms for clustering of datasets containing overlapped objects. The objective of current research is to devise an unsupervised algorithm which can perform the multi-level clustering of images on the basis of different disjoint feature-sets such as RGB values and spatial values (geo-coordinates) individually. The results of clusters generated at each level were evaluated through DB index, CH index and Silhouette score. Final clustering evaluation was done through an experiment, for which satellite, Radar, and crop images were considered.

Keywords: DB-scan · K-means · Multi-level clustering · Geospatial imageries · Radar · Satellite · Unsupervised machine learning.

1 Introduction

Segmentation of remote sensing images means distinguishing different objects on the basis of its features, which is a challenging task especially when the data is captured in real-time as heterogeneous and multi-dimensional in nature. In general, clustering is a process of categorizing the data on the basis of similarity. This can be extended in the geospatial domain as grouping the data points by extracting meaningful information from an image. Image classification can be used to interpret the contents of images like agriculture survey maps, Radar imageries, pathogenic images, satellite imageries, *etc.* The image processing techniques present at time are not enhanced enough to deal with high resolution images like satellite images because it requires fast processors and powerful GPUs for computations. In the last few years, several researches performed on clustering of remote sensed data due to its wide range of applications in natural and social sciences. Existing literature reviews of recent times suggest that there are some approaches used for satellite image processing like K-means, DB-scan, K-medoid, Hierarchical clustering, *etc.*

This work talks about multi-level clustering of Radar and satellite images with the application of the most suitable clustering techniques *viz.* DB-scan and K-means. In the first level, DBSCAN was applied for the clustering of images because DBSCAN is multivariate mode clustering. Basically, it looks for data points that are characteristically close together and partitions data accordingly. It's based on nearest-neighbors methods, which uses distance function (could be Euclidean or something else) which doesn't require the predefined number of K i.e. the number of clusters. It explores and creates neighborhoods around those points to cluster data that falls within that neighborhood into a single cluster. Multivariate outliers may not be clustered, as they are not close enough to other points to be included in any of the clusters. In the second level, DBSCAN or a different algorithm is applied again on next feature-set to get the probable number of optimized clusters (K) are found and then applied K-means on each cluster generated in the first level to get sub-clusters. The proposed method overcomes the issues of the existing clustering algorithm. It can find the number of segments by using level-wise clustering with the extraction of the different features at each level. After the re-clustering, the multi-level merging method has been applied to join each cluster with its similar cluster to avoid outliers and get an optimal segmented image from the original image. The proposed algorithm is tested using the validity indices DB index and CH index which are used specially, for which datasets contain overlapped objects in nature like geo-satellite images. .

2 Prior Art

2.1 Symmetry Based Cluster Validity Index: Application to Satellite Image Segmentation:

In 2006, Saha *et al.* [10] claims an appropriate partitioning technique to identify the correct number of clusters and the for the image segmentation like satellite land-cover. This is a very difficult problem in remote sensing images. Earlier research was done on point symmetry-based similarity measures. Most of the validity measures use a certain geometrical structure for clustering but in case if there are different structures exist, it causes variance and leads to failure. Second research has been done by Chou et al. on ps distance and it takes variability of cluster shapes into account on the basis of second research they have proposed the S-INDEX. According to that if the number of clusters (k) varied within some range then an underlying clustering technique like K-means or DBSCAN is used to partition the data. The value of k corresponding to the minimum value of S index will indicate the correct number of clusters in the data set.

2.2 Some connectivity based cluster validity indices:

Research work in 2012 & 2018 by Saha *et al.* [9] [11] is on the measurement of connectivity using the concept of a relative neighborhood graph. The property of connectivity significantly improved the capabilities of indices in identifying the appropriate number of clusters. Single linkage and K-means partition algorithms have been used on 11 datasets, in which 8 datasets are artificially generated and three are of real life data. Result shows that the DUNN index performs best as compared to all other existing indices. For distinct non-overlapping clusters, hierarchical clustering algorithm is used whether for overlapped cluster K-means and DBSCAN performs best. Similarly, in 2014 Mitra *et al.* [6] proposed the use of deep visual features into multi-objective based multi-view search results clustering.

2.3 GAPS: A clustering method using a new point symmetry-based distance measure:

GAPS uses a new point symmetry-based distance measure, proposed by Bandyopadhyay et al. (2007 and 2008) [2] [3]. The algorithm is able to detect both convex as well as non-convex clusters.. Kd-tree nearest neighbor is used to reduce the complexity of finding the closest symmetric point. SMKC-means is proposed for the circular invariant clustering of vectors. Fuzzy c-shells clustering methods have been proposed which are well established for detecting and representing hyperspherical (specially ellipsoid clusters). Fuzzy clustering is defined as the extraction of a smooth curve from the unordered noisy data. Fuzzy c-means is used to calculate the distance between cluster centers and data are determined by the density of the data itself. Non-convex clusters can be easily identified by these algorithms but highly overlapped clusters can not be identified by this.

GA performs search in complex, large and multimodal landscapes. It uses newly proposed point symmetry distance rather than Euclidean distance. This enables the proposed algorithm to detect both convex and non-convex as long as the cluster does have some symmetry property.

2.4 Gene-Based Clustering Algorithms:

Research work in 2020 by Martin [7] compared 3 clustering algorithms used in gene-based bioinformatics research to understand disease networks, protein-protein interaction networks, and gene expression data. Denclue, Fuzzy-C, and Balanced Iterative and Clustering using Hierarchies (BIRCH) were the 3 gene-based clustering algorithms tested in relation to analyze omics data, which include but are not limited to genomics, proteomics, metagenomics, transcriptomics, and metabolomics data. Fuzzy clustering is a hard clustering type while Partitioning clustering is called soft. The reason for that is while in Partitioning clustering, one data point may have only in one cluster, in Fuzzy clustering we have the probabilities of a data point for each cluster and they may belong to any cluster at this probability level. Choosing a fuzzy parameter too big may cause the result to be distorted. Unlike Denclue and Fuzzy-C which are more efficient in handling noisy data, BIRCH can handle datasets with outliers and have a better time complexity.

2.5 Evolution-Based Clustering Algorithms:

The evolution based algorithm proposed by R. Aslanzadeh *et al.* in 2017 [1] was tested on a standard dataset (BSDS 300) of images, and the region boundaries were compared with different people segmentation contours. Results show the efficiency of the algorithm and its better performance to similar methods. As claimed by the authors, in 70% of tested images, results are better than ACT algorithm, and in 100% of tested images, they had better results in comparison with VSP algorithm that was proposed by W. Vanzella and V. Torre [12] in 2006. Since, its results were mainly discussed with the datasets having truth labels of various segments in the images, therefore, this algorithm has been tested again with computing various similarity and individuality indices for final clusters to compare it with the proposed multi-level clustering which is focused for the datasets without any truth label.

3 Data and Methodology

3.1 Dataset Description:

This paper is resolving the issue of heterogeneous dataset clustering. Here, the word heterogeneous refers to the speciality of a dataset in which the parameters or attributes of the data-points are of different characteristics as well as of different scales which can't be considered together to compute distance between two points. One of such cases in the geospatial domain is the Radar datasets. Different Radar products are visualized in image form using different scaling systems and the geographical coordinates are in the space scale. One of Radar products is intensity of rainy-cloud, represented by logarithmic Radar-reflectivity (DBZ) as depicted in fig. 1. In this proposed method, along with the radar scan, satellite images, crop unhealthy organ (leaf, pathogenic leaf) images are also tested. Every dataset of different domains contains 100 images, collected from IMD-Bhopal centre of Ministry of Earth Sciences, Govt. of India.

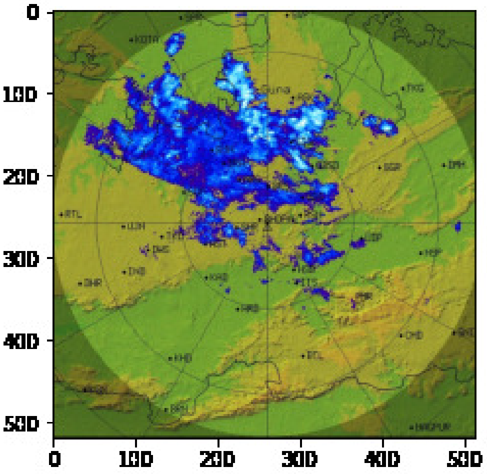


Fig. 1. Bhopal-Radar imagery depicting rainy-clouds intensity over Madhya Pradesh State

3.2 Methodology

This paper focuses on devising the clustering technique for deep clustering of heterogeneous datasets. The main issue while solving this kind of problem is that we can't analyze such heterogeneous parameters using the single technique and make decisions of the cluster based on that.

Unsupervised Multi-level Clustering Algorithm: In the proposed approach, we have followed the technique of split and merge inorder to get the most optimal solution. After splitting each object-component based on a particular feature-set, the merging process is repeated multiple times with halting conditions on the results after each iteration. Detailed algorithm 1 along with task-flow (fig. 2) is mentioned below:

Algorithm 1: Unsupervised Multi-level Clustering Algorithm

Input : Image represented with feature-set 1 (colour intensity) and feature-set 2 (pixel geo-positions).

Output: Segmented image with embedded objects separated optimally.

```
begin
    /* Level-1: Clustering on feature-set 1                                  */
    DBScan Algorithm on intensity features i.e. R,G,B channel of the image
      with eps=1 to 3 and min_points=30, hyperparameters to be chosen as
      per imagery dimensions ;
    /* Level-2: Clustering on feature-set 2                                  */
    for each cluster of above clustering result : do
        find value of K by applying DBScan using spatial features
        i.e. X,Y coordinates;
        apply K-means with the value of K to get sub-clusters;
    end for
    /* Level-3: Merging of one or more sub-clusters.                         */
    while number_of_total_cluster is reducing or DBindex is
      decreasing significantly do
        : for sub-cluster_i of image at level-2 : do
            for sub-cluster_j among all sub-clusters at level-2 : do
                if (sub-cluster_i != sub-cluster_j) and (sub-cluster_i
                  and sub-cluster_j are not merged) and (sub-cluster_i
                  and sub-cluster_j are similar in terms of intensity
                  features ) : then
                    if  is_neighbour(sub-cluster_i, sub-cluster_j) :
                      then
                        merge(sub-cluster_i, sub-cluster_j);
                    end if
                end if
            end for
        end for
    end while
end
```

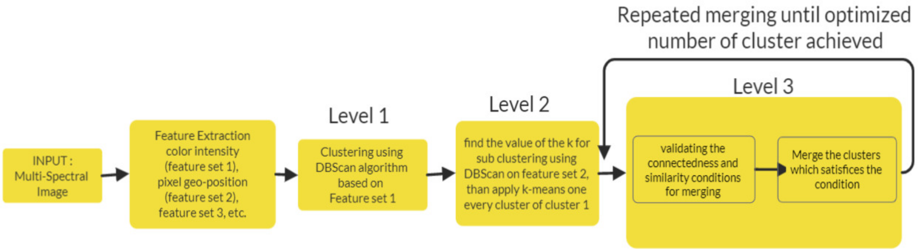

Repeated merging until optimized number of cluster achieved
INPUT : Multi-Spectral Image
Feature Extraction color intensity (feature set 1), pixel geo-position (feature set 2), feature set 3, etc.
Level 1
Clustering using DBScan algorithm based on Feature set 1
Level 2
find the value of the k for sub clustering using DBScan on feature set 2, than apply k-means one every cluster of cluster 1
Level 3
validating the connectedness and similarity conditions for merging
Merge the clusters which satisfices the condition

Fig. 2. Task flow of proposed multi-level unsupervised clustering method

3.3 Clustering Validity Indices

To check the performance of final clusters generated on different datasets and to compare the end results of conventional and evolution based clustering methods with our proposed level-wise clustering method, we have used following cluster validity indices applicable for the image datasets with no truth label available for their segments:

1. Davies Bouldin score (DB index)
   The score is defined as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances, described by [5]. Thus, clusters which are farther apart and less dispersed will result in a better score. The minimum score is zero, with lower values indicating better clustering. Its limitation are that this index is generally higher for convex clusters than other concepts of clusters, such as density based clusters like those obtained from DBSCAN. Also, the usage of centroid distance limits the distance metric to Euclidean space.
2. Calinski Harabasz score (CH index)
   This index is also known as the Variance Ratio Criterion. The score is defined as ratio of the sum of between-cluster dispersion and of within-cluster dispersion. Caliński [4] tells that if the ground truth labels are not known, then it can be used to evaluate the model, where a higher Calinski-Harabasz score relates to a model with better defined clusters. The score is higher when clusters are dense and well separated, which relates to a standard concept of a cluster. The score is fast to compute.
3. Silhouette score If the ground truth labels are not known, evaluation must be performed using the model itself. The Silhouette Coefficient given by [8] is an example of such an evaluation, where a higher Silhouette Coefficient score relates to a model with better defined clusters. The Silhouette Coefficient is defined for each sample and is composed of two scores:

   - The mean distance between a sample and all other points in the same class.
   - The mean distance between a sample and all other points in the next nearest cluster.

   The Silhouette Coefficient s for a single sample is then given as:

   $$s = \frac{(b-a)}{max(a,b)} \tag{1}$$

   The Silhouette Coefficient for a set of samples is given as the mean of the Silhouette Coefficient for each sample.

4 Results

4.1 Level-1 Clustering for grouping of data points based on most impactful feature-set

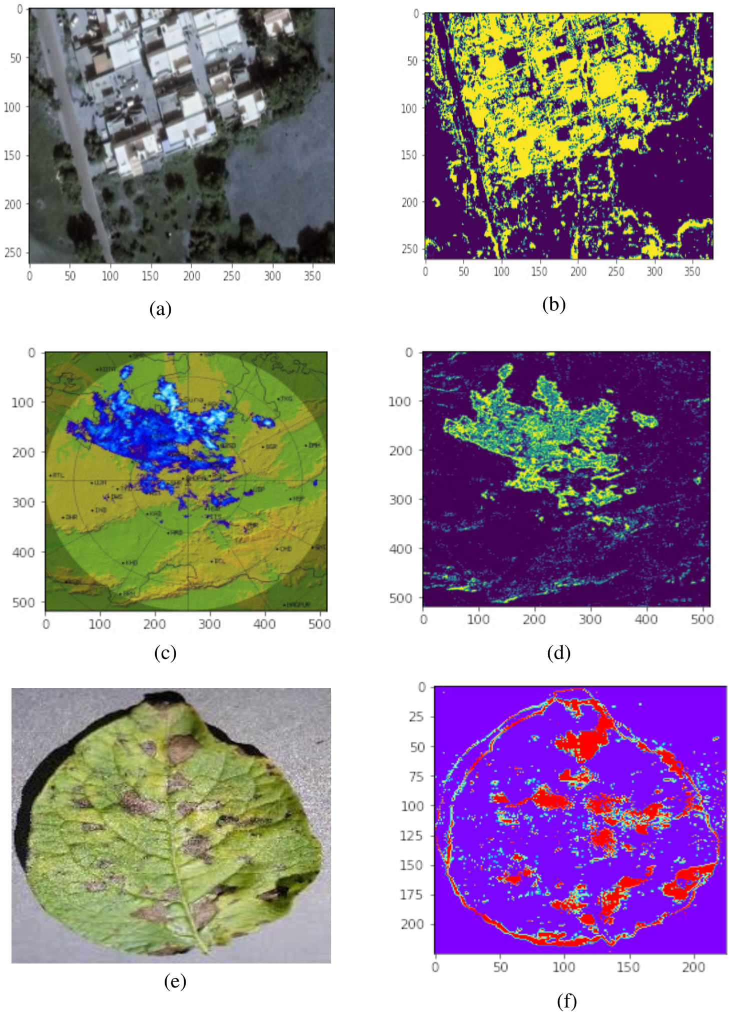


Fig. 3. (a) Original satellite imagery, (b) Segmented satellite imagery after level-1 clustering, (c) Original Radar imagery, (d) Segmented Radar imagery after level-1 clustering, (e) Original crop (potato) unhealthy leaf image, and (f) Segmented crop (potato) unhealthy leaf image after level-1 clustering

We used different type image datasets such as satellite image, Radar scans, Pathogenic images as well as crop landscape images. Thus, we have different parameters available based on the domain of the dataset for the first level of clustering but after computing the value of cluster validity indices, the observation is made that for all of the datasets the colour feature is the most impactful out of all the rest parameters for each dataset.

Hence, In the first level clustering, the images were segmented based on colour features which are the intensity of Red, Green and Blue channels. For clustering, various unsupervised algorithms are already tested with a single set of attributes used together to compute inter-point distance or check similarity of two data-points. The datasets which we are using in this work, contain overlapped or embedded objects from the color perspective. Thus, out of all we are using DBSCAN for the first level of clustering as it does not require the number of clusters (K) to be known for clustering and as it performs the best for the overlapped dataset. Fig. 3 (a) and (b) show the level-1 clustering results on satellite captured earth-surface imagery; Fig. 3 (c) and (d) show the same on Radar scan imagery, while Fig. 3 (e) and (f) show the same on unhealthy potato crop leaf image.

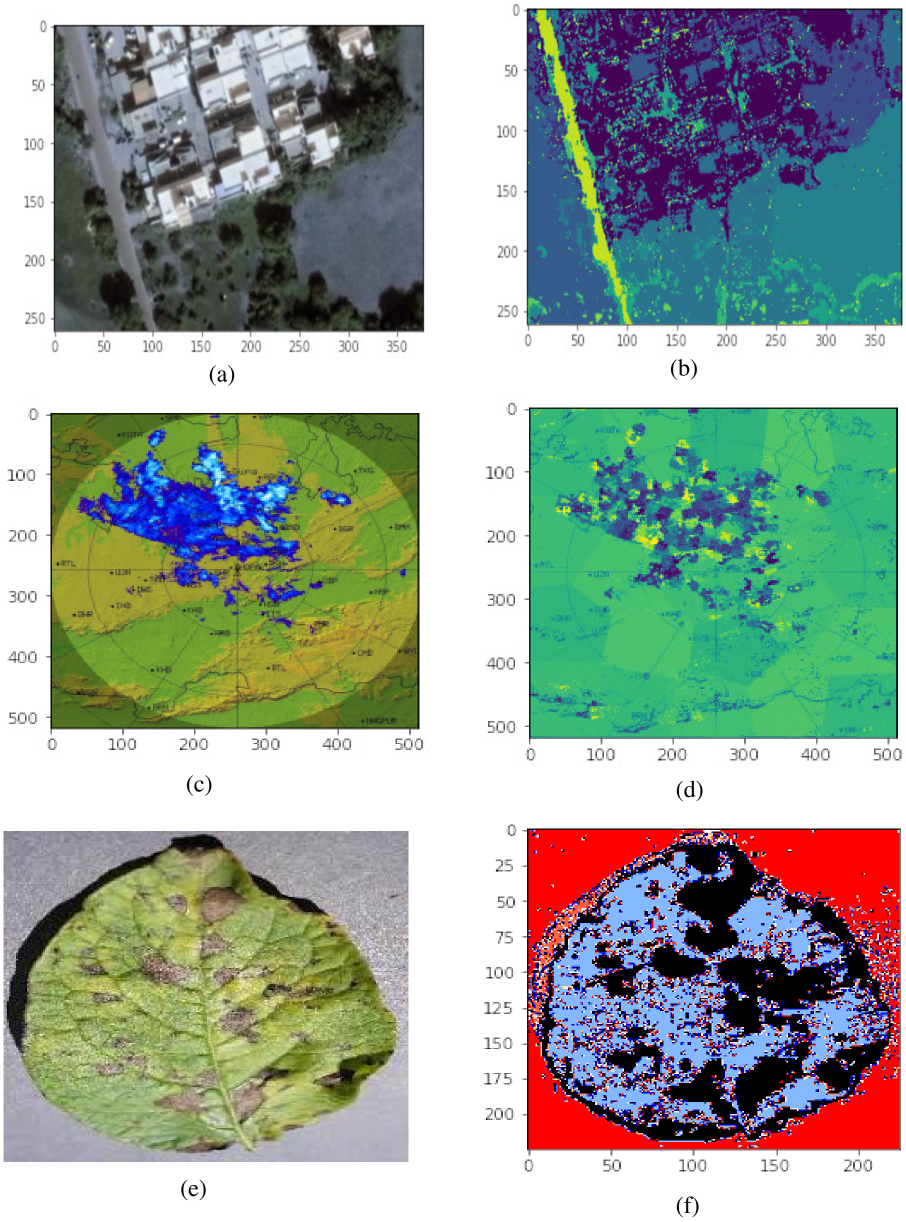
4.2 Level-2 Clustering for sub-clustering of level-1 clusters based on the another feature-set

In this step, the clusters are further divided based on another feature-set. Clustering of any dataset based on the color intensity can divide the data points with respect to color but for proper clustering the position of the point and the distance among the data points also plays a vital role for cluster creation as their geo-position matters in the geospatial domain. Thus, in the second level, we divide each cluster based on its coordinate position.

From the previous experiments, it is clearly examined that the K-means and K-medoid algorithms work best for clustering on non-overlapped datasets. As each cluster is separated from its neighboring cluster based on colour feature-set in the first level of clustering, the data points in each cluster is well separated from the data-points of other cluster and K-means would be probably the best for next level of clustering algorithm but the only problem in this approach, is the value of K which needs to be optimized inorder to get most correct clusters.

In the proposed method, the second level of clustering consists of 2 steps : (a) Defining the value of K : Each cluster is grouped using DBscan on X,Y coordinate data. By doing this, we are allowing the algorithm to be flexible with the individual value of K based on the number of datapoints of a particular cluster generated at level-1.

(b) On the value of K, the K-means clustering is applied on each cluster and thus every cluster is divided into sub-clusters. Fig. 4 (a) and (b) depict the sub-clusters at level-2 clustering on satellite captured earth-surface imagery; Fig. 4 (c) and (d) depict the same on Radar scan imagery, while Fig. 4 (e) and (f) depict the same on unhealthy potato crop leaf image.



(a)

(b)

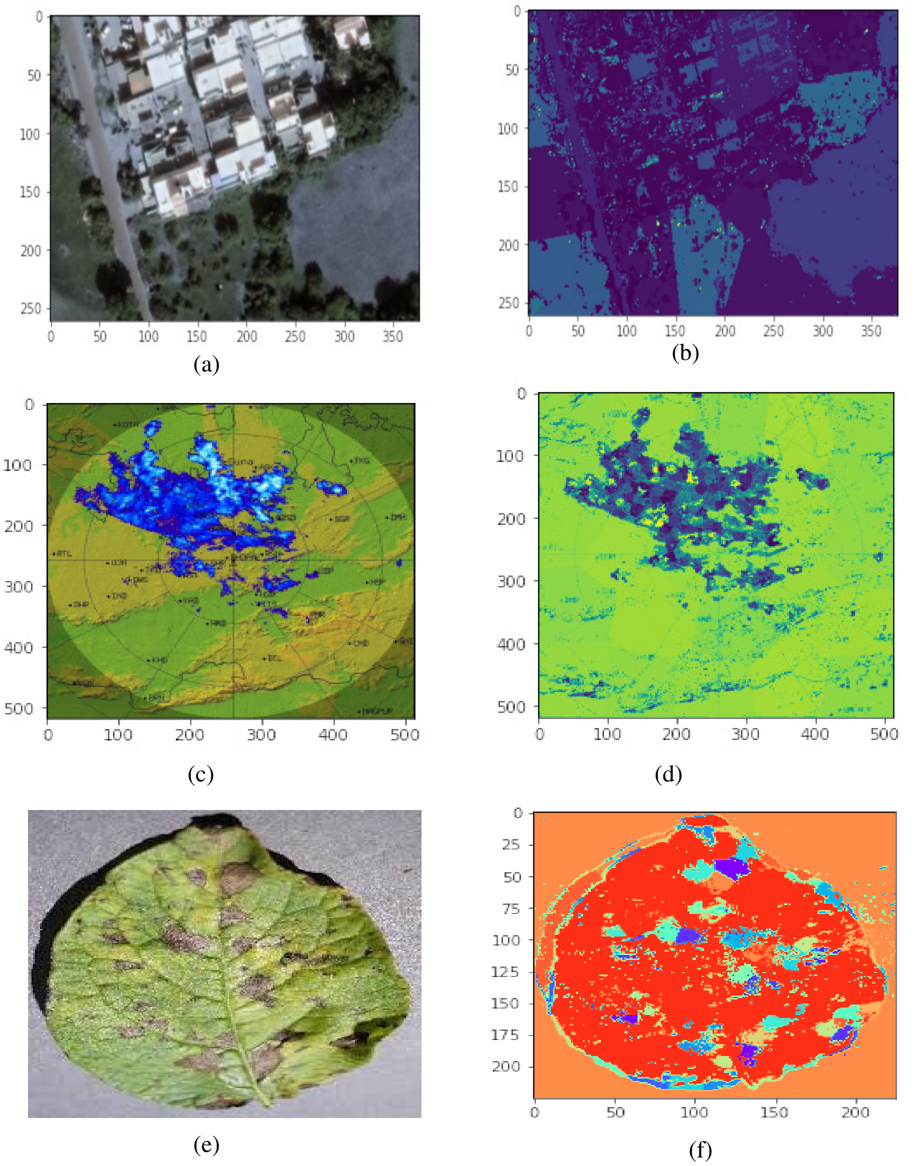
(c)

(d)

(e)

(f)

Fig. 4. (a) Original satellite imagery, (b) Segmented satellite imagery after level-2 clustering, (c) Original Radar imagery, (d) Segmented Radar imagery after level-2 clustering, (e) Original crop (potato) unhealthy leaf image, and (f) Segmented crop (potato) unhealthy leaf image after level-2 clustering



(a) (b) (c) (d) (e) (f)

Fig. 5. (a) Original satellite imagery, (b) Segmented satellite imagery after level-3 clustering, (c) Original Radar imagery, (d) Segmented Radar imagery after level-3 clustering, (e) Original crop (potato) unhealthy leaf image, and (f) Segmented crop (potato) unhealthy leaf image after level-3 clustering

The algorithm will merge different sub-clusters if those 2 sub-clusters are satisfying the condition of the merging. The primary condition for merging is that both clusters are connected. In addition, the cluster colour intensities should also match with each other. For reducing the time-complexity, we first check whether both sub-clusters' average RGB intensities are almost the same or not and then, after that the geo-connectivity is checked between both the sub-clusters. The merging process is repeated for multiple times until the change in the number of clusters is nominal or the validity indices show a reverse trend. Fig. 5 (a) and (b) depict the final level-wise merged clustering results on satellite captured earth-surface imagery. Similarly, Fig. 5 (c) and (d) depict the final clustering results on Radar scan imagery, while Fig. 5 (e) and (f) depict the same on unhealthy potato crop leaf image.

5 Discussion:

Table 1 compares the clustering validity using DB index values for first-level, second-level and final-level results generated from the proposed algorithm along with existing algorithms for each dataset. DB-index values are quite high for the conventional algorithms *viz.* DBSCAN, K-Means and recent Evolutionary clustering [1]. Though, it is higher for level-1 clustering but reduced further after applying multi-level merging for all three datasets. This implies that the proposed algorithm converges at every level towards more optimal results and overall, it performed better than the conventional methods of clustering for the datasets containing embedded objects especially in geo-spatial domain (i.e. satellite imagery).

Table 1. Comparison of cluster validity-index (DB index) for conventional clustering algorithms and Level-1, 2 & 3 of proposed algorithm for different datasets.

Algorithm	DBSCAN	K-Means	Evolution based Clustering	Proposed multi-level unsupervised Clustering		
Dataset				Level-1 Clustering	Level-1 Clustering	After final merging
Radar Image	19.1232	25.1211	8.9093	13.4322	9.2551	4.3117
Satellite Image	23.9798	29.5541	9.0221	11.8773	7.2131	2.3149
Crop organ Image	17.7165	20.0621	11.0214	12.8846	8.6731	3.5478

Further, Table 2 shows the CH index for the clustering results obtained at different levels from the proposed algorithm. High values of CH indices at the final level indicate that end clusters are dense and well separated.

Similarly, Table 3 shows the Silhouette score for the clusters obtained at each level of proposed algorithm. This score should be ideally 1, if all clusters are obtained correctly. Moving from negative values (wrong clustering of given sample) to positive values (correct clustering of given sample) indicates that

Table 2. Calinski-Harabasz Index for Level-1, 2 & 3 of proposed algorithm for different datasets.

Level of proposed Clustering	Level-1 Clustering	Level-2 Clustering	After final merging
Dataset			
Radar Image	900.1999	851.1321	950.1214
Satellite Image	854.7000	803.9994	936.4514
Crop organ Image	857.1991	764.2113	871.9111

proposed level-wise method is conversing towards more correctness with each level. End results obtained from these datasets were experimentally verified with manual observations and found valid in most of the cases.

Table 3. Silhouette Coefficient for Level-1, 2 & 3 of proposed algorithm for different datasets.

Level of proposed Clustering	Level-1 Clustering	Level-2 Clustering	After final merging
Dataset			
Radar Image	-0.2014	-0.1011	0.5741
Satellite Image	-0.3211	-0.1547	0.6733
Crop organ Image	-0.2775	-0.1000	0.5112

6 Conclusion

Most of the geo-spatial datasets which are used in the domain of natural sciences, space and biological sciences carry multiple dimensions with disjoint sets of characteristics. Conventional clustering algorithms usually fail to perform better when objects are highly variable in size and shape as well as embedded with each other in such datasets. Finding the irregular shaped objects with no truth labels available is very difficult with general clustering mechanisms in the applications, such as identifying rainy clouds or trees in Radar or satellite images respectively or diagnosing unhealthy spots in crop organs, which can be extracted from the proposed algorithm with satisfactory accuracy. Further, use of other available channels or feature-sets in these multi-spectral datasets may enhance the accuracy of results and the same needs to be tested.

7 Acknowledgement

Authors would like to thank India Meteorological Department, Bhopal for providing the Radar, Agricultural and Satellite datasets for our study and experimental validation on expert observations.

References

1. Aslanzadeh, R., Qazanfari, K., Rahmati, M.: An efficient evolutionary based method for image segmentation. Computer Vision and Pattern Recognition, arXiv preprint arXiv:1709.04393 (2017)
2. Bandyopadhyay, S., Saha, S.: Gaps: A clustering method using a new point symmetry-based distance measure. Pattern recognition **40**(12), 3430–3451 (2007)
3. Bandyopadhyay, S., Saha, S., Maulik, U., Deb, K.: A simulated annealing-based multiobjective optimization algorithm: Amosa. IEEE transactions on evolutionary computation **12**(3), 269–283 (2008)
4. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics-theory and Methods **3**(1), 1–27 (1974)
5. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence (2), 224–227 (1979)
6. Mitra, S., Hasanuzzaman, M., Saha, S., Way, A.: Incorporating deep visual features into multiobjective based multi-view search results clustering. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 3793–3805 (2018)
7. Nwadiugwu, M.C.: Gene-based clustering algorithms: comparison between denclue, fuzzy-c, and birch. Bioinformatics and biology insights **14**, 1177932220909851 (2020)
8. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)
9. Saha, S., Bandyopadhyay, S.: Some connectivity based cluster validity indices. Applied Soft Computing **12**(5), 1555–1565 (2012)
10. Saha, S., Bandyopadhyay, S., Maulik, U.: A new symmetry based cluster validity index: Application to satellite image segmentation. In: 9th International Conference on Information Technology (ICIT'06). pp. 121–124. IEEE (2006)
11. Saha, S., Mitra, S., Kramer, S.: Exploring multiobjective optimization for multiview clustering. ACM Transactions on Knowledge Discovery from Data (TKDD) **12**(4), 1–30 (2018)
12. Vanzella, W., Torre, V.: A versatile segmentation procedure. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) **36**(2), 366–378 (2006)