Indian Journal of Science and Technology, Vol 8(36), DOI: 10.17485/ijst/2015/v8i36/78980, December 2015
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Classifications of Dynamic Clustering based on Soft Computing

Chatti Subbalakshmi[1*], G. Rama Krishna[2] and S. Krishna Mohan Rao[3]

[1]Department of CSE, GNITC, Hyderabad - 500008, Telangana, India
[2]Department of CSE, KL University, Guntur - 522502, Andhra Pradesh, India
[3]Department of CSE, Sidhartha Engineering College, Hyderabad - 501506, Telangana, India

Abstract

Data mining is major area in innovation of new trends in many application areas. The current applications generating data is are not static, it always changing day to day and it leads to changes in the technologies and data mining algorithms. Here, new study of data mining algorithms with dynamic characteristic. In various applications, data clustering is required for grouping of data based on similarity measures. For clustering on uncertain, incomplete and vague data soft computing approaches are efficient. In this paper, we considered dynamic characteristics of data and class structures into four cases and framed basic soft clustering methods for these cases. These methods are executed using fuzzy c means clustering in R programming language on suitable real time data set. We also focus on problems identified on each category.

Keywords: Dynamic Data, Dynmaic Clustering, Softcomputing

1. Introduction

In recent years all applications are online and generating data are changing over time. Finding new trends on these data bases traditional data mining algorithms[1] are not suitable. Due to classical algorithms takes static inputs and values can't be changes as the data changes. Hence dynamic data mining comes into picture, a new research area which combines traditional data mining with dynamic characteristics. Mainly three approaches are present in literature defined in[2]. We need to perform complete data mining task from the scratch due to the change in data. It takes computational cost of mining each time. Hence, most of dynamic data mining approaches fallow the updating of initial system according arrival new data.

In data mining, the data segmentation is one of the popular task to support the different application requirements. The data segmentation is process of grouping given data set into segments or also called clusters. Therefore the other name of data segmentation is cluster analysis[3]. Many clustering strategies are present in the literature. One of method is partition clustering, where given data is divided in specified number of clusters. The k-means[4] and k-mediod[5] algorithms are popular methods in partition clustering techniques. These methods need to mention number of clusters i.e., K value in prior to execution of algorithm. User need knowledge on application data set to decide k value based on type of data sets, which is not possible on unknown data set. In literature many modified k-means algorithms are available on dynamic data sets[6].

Intelligent applications, expert systems and biologiacal data applications are generate uncertain and vague data. For doing analysis on these types of data hard clustering is inefficient. Therefore soft computing is introduced to do analysis on uncertain and vague data[7]. Fuzzy set theory[8], Rough set theory[9], Neural Networks[10], Evaluation computing[11] and Genetic algorithms[12] are components of Soft computing.

Cluster analysis can be done either hard or soft computing. In case of hard clustering, no overlapping between the clusters. But in soft clustering, data point can be belongs to more than one clusters. i.e., overlapping between the clusters are allowed. Fuzzy clustering is one of the soft clustering approaches defined based on fuzzy

*Author for correspondence

set theory[13]. Rough clustering is belongs to group of soft clustering algorithms and it is defined based on rough set theory[14].

In this paper, we defined four data clustering methods based on characteristics of data and class structure. All these methods are executed using fuzzy clustering on different application data bases. The paper is organized as, classifications of dynamic clustering are given in section two, evaluations of these methods are given in section three and conclusions are mentioned in Section 4.

2. Categorization of Dynamic Clustering

The static data does not change over time, whereas dynamic data changes. For example, in customer segmentation the behavior of customer changes always. Hence we need to adopt the changes to find new patterns. In static environment, once we perform clustering on initial data set results are same for all the times. But in dynamic environment, always patterns change as data changes. In literature many strategies are presented to perform clustering on dynamic data[15,16]. In this paper, we selected group partition clustering using soft computing. The many partition algorithm results are dependent on inputs like data set (D) and number of clusters (k). Here we consider four cases with combination of data (D) and class (k) with static and dynamic aspects. The classification of dynamic clustering is given in Table 1. These methods are executed based on fuzzy c means clustering algorithm[17] and observations are given by analyzing the results. To find right number of clusters, we used silhouette cluster validity index is used[18,19]. Basically fuzzy c means algorithm requires the number of clusters (c value) and data set D. We framed dynamic clustering methods by changing these values to show the three dynamic classifications and presented in following sections as mentioned in Table 2.

2.1 Traditional Clustering

Clustering on stationary data using fixed classes is called traditional clustering. Both the groups of classical partition clustering algorithms; hard clustering (k-means and k-medoids) and soft clustering (fuzzy c-means and rough k-means clustering), the data set (D) and number of clusters (k) known in advance to execution of algorithm. It groups the data into k number of clusters depends on similarity between data points. The distance measure is to calculate the similarity between the data objects and mean of the cluster represents the cluster centroid. Based on number of clusters it performs clustering and results are depends on k value and initial cluster centroids. The k value is decided before execution of algorithm and always it is depends on data base characteristics. Until there is no change in data set, we can use same k value for all the time and continue with same results. Here, static parameters are number of clusters (k) and data set (D) assume to be does not change in this case. In almost classical clustering algorithms takes static input parameter and performs clustering depends on these values. Therefore classical clustering algorithms are not efficient on changing data set. Hence, we need go for dynamic clustering strategies. In the next section we present three dynamic clustering approaches.

Table 1. Classifications of dynamic clustering

	Static class	Dynamic class
Static data	Traditional clustering	Dynamic clustering
Dynamic data	Dynamic clustering	Dynamic clustering

Table 2. Cases of dynamic clustering

Type of clustering	Range of values	Data size	No of clusters
Static	Constant	Constant	Constant
dc-1	Constant	Changes	Changes
dc-2	Changes	Changes	Constant
dc-3	Changes	Changes	Changes

2.2 Dynamic Clustering – Static Data using Dynamic Class

In this category, data is considered as static means fixed in data values and range of values. As performing the clustering process the class structure (number of classes) are changes. The basic frame work is given below, which performs clustering that assigns static objects to dynamic classes. Here cluster structure changes over time due to new incoming data. After adding new data, the data size increases which may effect on cluster number and they may not fit in existing clusters. For that new cluster is created. Hence, identify the constraint of changes in cluster structure before adding the new data.

The basic frame work is given as,

Step 1: Find the optimal number of clusters (k) value for initial data set (D).

Step 2: Perform initial clustering with (k) number of clusters.

For each new incoming data repeat the following steps,

Step 3: Find the changes in cluster structure with respect to new data D_{new}.

The new data object (x_i) need changes in cluster if below two conditions are satisfied,

Condition 1: New data x_i is far away from all current cluster centers.

Condition 2: All membership value of new data to cluster centers is close to 1/number of clusters (1/c).

Step 4: If above two conditions are satisfied, create new cluster; otherwise move cluster centers.

Step 5: Re-cluster the data as per changes.

2.3 Dynamic Clustering – Dynamic Data with Static Class

In this method, the behavior of data (data size, data value and range of values) is considered as dynamic i.e., changes over time and the numbers of clusters are defined.

The basic frame work as,

Step1: Fix the required number of clusters (k).

Step 2: Perform clustering with (k) no of clusters on initial data (D).

For each new incoming data repeat,

Step 3: Add new data with exiting data.

Step 4: Re-cluster complete data with same number of clusters (k).

2.4 Dynamic Clustering – Dynamic Data with Dynamic Clusters

In this case, the characteristic of data (size, value and range) is changes over time. i.e., data is dependent on time and hence, class structure also changes over time. This classification is called temporal data clustering on temporal data bases. Some of the methods of temporal data clustering are Hidden Markov Model (HMM), Motion model and time series. The method is framed to cluster on these data set using soft clustering algorithms (fuzzy c-means and subtractive clustering) and fix the time interval (t) for each cycle. For fuzzy c-means algorithm, cluster number is defined for each interval (t), but in subtractive clustering algorithm is not needed.

The defined frame work of the method using fuzzy c-means as,

Initial clustering:

Step 1: find initial number of clusters (k).

Step 2: execute clustering with (k) value.

Repeat for each time interval (t),

Step 1: find optimal number of clusters.

Step 2: perform the clustering on entire data.

3. Evaluation of Dynamic Clustering Methods

In this section, dynamic methods are executed based on fuzzy c means[16] and to find optimal number of clusters, silhouette coefficient[17] is used. All methods are implemented in R programming software. To evaluate these methods, we generated synthetic random samples by changing data size, data values and range of values to show the dynamic characteristics and results are given. The novelty of methods is showed by applying these methods on suitable real time application data sets which are collected from the UCI public repository.

3.1 Dynamic Method-1 Results

The method is executed on synthetic data which generated random sampling method to meet our assumptions in data characteristics. To show the novelty of method, customer segmentation application data is selected, where frequently the customer purchase behavior changes season wise and which affects the customer group formation.

3.2 Synthetic Data Results

The synthetic data with two variable (x, y) is generated using permutation sampling and bootstrap sampling methods in R software. Initially, 50 data objects (x, y) are generated within constant range 0 to 50 for both variables. As per method, we applied silhouette index on these data to find the right numbers of clusters, which results are three clusters. The cluster centers and their silhouette width are given Table 3. In next cycle, 50 more new data objects are created randomly and identified changes in cluster with respect to new data and results are given in Table 4.

After applying condition-1 on new data points, 19 points are having distance between the cluster center to its are greater than ½ min{distance of cluster centers} which

showing far away from the three cluster centers and not fitting in these clusters and all membership values are near to 1/number of clusters (1/3). It shows that, new clusters are needed to create. For that, execute silhouette index on complete data set to find optimal number of clusters and perform fuzzy c-means algorithm.

3.3 Results on Customer Segmentation Data

To show novelty of this method, we choose the customer segmentation problem. Here, the buying behavior of customers always changes over time. But, the customer details will not be changes mostly. Therefore, data characteristics remains unchanged, but patterns are going to change. We executed this method on wholesale customer segmentation data set consist of nine attributes in three cycles.

Table 3. Cycle -1 center of fuzzy clusters

Clusters	X	Y	Cluster Silhouette width value
c1	25.66626	10.12006	0.6850912
c2	39.23399	38.02186	0.5796464
c3	9.225216	27.04826	0.4192992
		Avg. sil.	0.5644897

Table 4. Cycle -2 results

old clusters = 3	Distance Between clusters	New clusters = 2	Silhouette width
(C1,C2)	31.0256889	C1	0.5372314
(C1,C3)	23.5981284	C2	0.4334830
(C2,C3)	31.9522513	Avg. sil.	0.4823058
½{min distance}	15.51284		

Initially 20 random samples are generated from actual data set and applied silhouette index to find optimal number clusters. The results are given in Table 5 and two is right number of clusters with high silhouette value.

In next cycle 20 more new objects are added and find the changes in clusters with respect to new data. We identified some of data points require moving of existing cluster centers without changing cluster number. The results are given in Table 6.

3.4 Dynamic Method-2 Results

3.4.1 Synthetic Data Results

We created two dimensional data by change in range of values and data size for each cycle, and executed fuzzy c-means algorithm with four number of cluster for all cycles. The cluster results are given in Table 7. From the results, identified that as change in data, optimal number of clusters also changing. As we fixed cluster number (c = 4), Silhouette cluster validity index is decreasing and shows that data points not properly grouped. But as requirements of applications, cluster number can be constant. The cluster results are given Figure 1.

3.4.2 Breast Cancer Data Results

Here, we selected breast cancer data where classes (Benign/Malign) are defined but data values may be

Table 5. Result of step 1

Cluster	Sil width	Cluster no	Sil width
2	0.676344	5	0.356714
3	0.508388	6	0.261475
4	0.454025	7	0.232145
5	0.37516	8	0.203021

Table 6. Results of two cycles on whole customer segmentation data

clus-1 center	channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_ Paper	Delicassen
Old	1.655071	3	7914.64	5712.485	7986.423	1423.608	2774.497	1962.236
New	2	3	23139.76	8559.293	11445.28	1022.658	16027.01	6852.629
Moved	1.791593	3	13940.72	6839.247	9355.433	1264.913	8019.821	1654.839
clus-2 center	channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_ Paper	Delicassen
Old	1.966694	3	22048.7	7686.176	11017.52	1952.139	3979.721	2650.81
New	1	4.684762	3896.161	3896.161	14778.86	1044.702	3932.077	782.3553
Moved	2.966694	5.425168	23556.25	6220.299	12472.3	1601.167	3961.293	1928.142

changes between cases to case. The method is executed on Breast Cancer Data (WBCD) in three cycles given in Table 8. The cluster points are given Figure 2 to 4.

3.5 Dynamic Method-3 Results on Stock Exchange Data

This method is executed on stock exchange, where data changes day wise. We collected data monthly and set time interval as month. We applied the dynamic clustering method by incrementing these data. As initial clustering, the first month data is clustered with two optimal numbers of clusters given in Table 9.

In second cycle, we added second month data to first month, agin we calculted right clusters and performed

Table 7. Cluster results

	Cycle-1	Cycle-2	Cycle-3
Data range	0:15	0:30	0:50
Data size	50	100	150
Optimal cluster number & avg cluster sil width	4	7	2
	0.5013412	0.6246845	0.6169781
Fixed cluster number and avg cluster sil width	4	4	4
	0.5013412	0.5913356	0.5702474

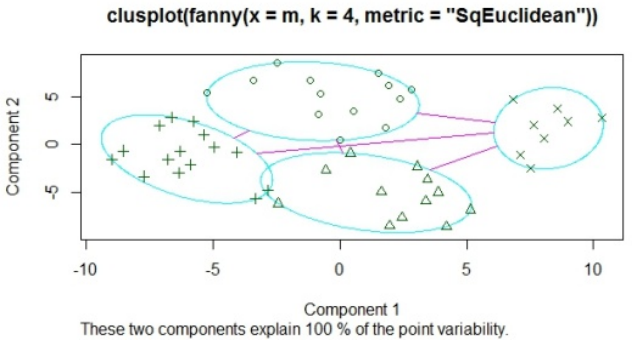

Figure 1. Cycle-2 cluster points.

Table 8. Results of dynamic fuzzy c-means clustering algorithm

	Cycle-1	Cycle-2	Cycle-3
Data size	367	437	468
Clus Avg Sil width	0.69340811	0.7204372	0.72818879
Right c value	2	2	2
Dunn_coeff	0.8041438	0.8187395	0.8222193
Belgian	207	263	286
Malignant	160	174	182

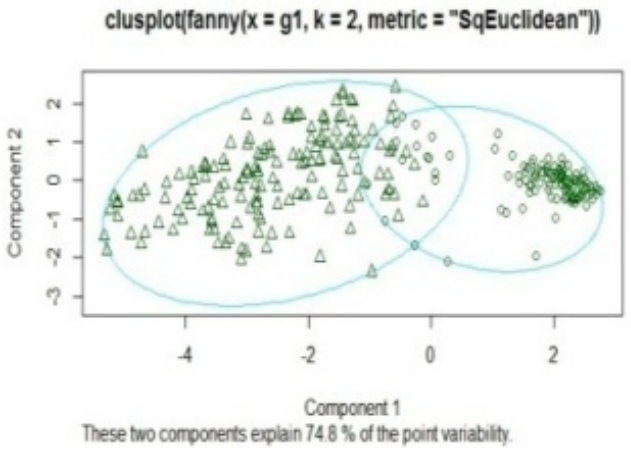

Figure 2. Cycle-1 cluster points.

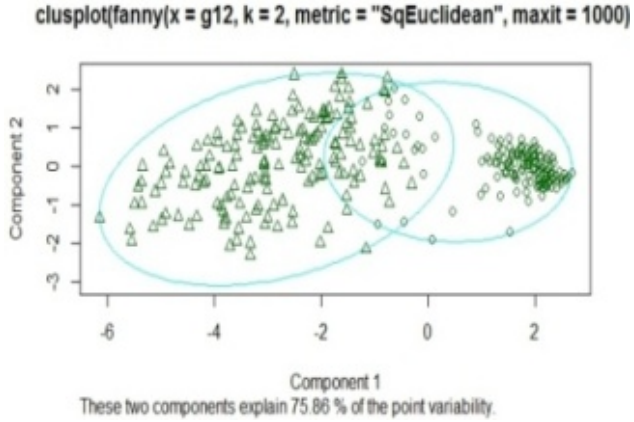

Figure 3. Cycle-2 cluster points.

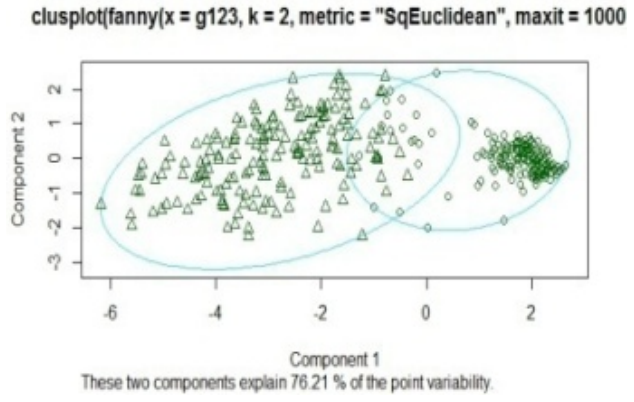

Figure 4. Cycle-3 cluster points.

the clustering. The results are given Table 10. As data size, value changes the number of clusters (3 clusters) also changed.

The output of fuzzy clustering is given Figure 5 and 6 for cycles.

Table 9. Step 1 results for finding right clusters

Cluster no	Sil width	Cluster no	No of clusters
2	0.795451	6	0.670182
3	0.741281	7	0.452615
4	0.705557	8	0.446102
5	0.6883388	9	0.4983554

Table 10. Cycle-2 results for finding right clusters

Cluster no	Sil width	Cluster no	No of clusters
2	0.724956	6	0.654952
3	0.796966	7	0.56944
4	0.706814	8	0.479083
5	0.577404	9	0.425186

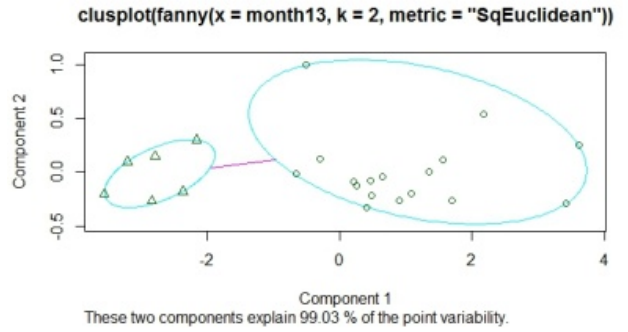

Figure 5. Cycle-1 cluster points.

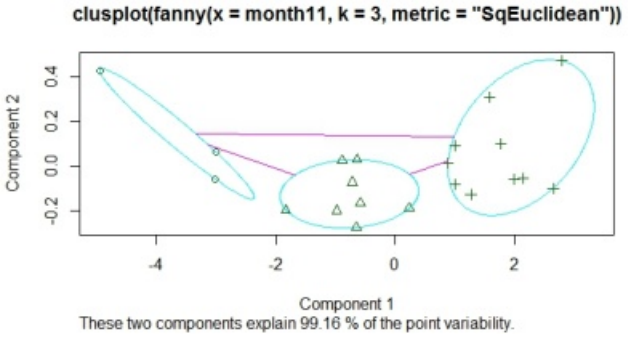

Figure 6. Cycle-2 cluster points.

4. Conclusion

In recent years many application data base are changing frequently. Hence, we need to adopt these changes in data analysis by applying dynamic aspects to classic methods. In this view, many methodologies are proposed in literature. In this paper, we considered the changing behavior of data as, change in data value, data size and range. It implies, change in patterns data analysis. We framed three cases and proposed dynamic clustering for these using one of the soft computing method fuzzy clustering. These methods can be further enhancing by using other soft computing methods.

5. References

1. Ian HW, Eibe F, Mark AH. Data Mining: Practical Machine Learning Tools and Techniques. 3rd ed. Elsevier; 2011 Jan 30. ISBN: 978-0-12-374856-0.
2. Crespo F, Weber R. A methodology for dynamic data mining based on fuzzy clustering. Fuzzy Sets and Systems. 2005 Mar; 150(2).
3. Hartigan JAJA. Clustering algorithms. John Wiley and Sons, Inc; 1975.
4. Park HS, Jun CH. A simple and fast algorithm for K-medoids clustering. Expert Systems with Applications. 2009; 36(2):3336–41.
5. Hartigan JA, Wong MA. Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society, Series C28. 1979; 1:100–8.
6. Hongyang, He J. Application of dynamic clustering algorithm in the center selection RBF nueral networks. WGEC'09 3rd International Conference Genetic and Evaluation Computing; 2009. p. 488–91.
7. Lotfi AZ. Fuzzy logic, neural networks, and soft computing. Communication of the ACM. 1994 Mar; 37(3):77–84.
8. Zadeh LA. Fuzzy sets. Information and Control. 1965; 8(3):338–53.
9. Zdzisław P. Rough sets. International Journal of Parallel Programming. 1982; 11(5):341–56.
10. Warren M, Pitts W. A logical calculus of ideas immanent in nervous activity. Bulletin of Mathematical Biophysics. 1943; 5(4):115–33.
11. Back Th, Schwefel H-P. An overview of evolutionary algorithms for parameter optimization. Evolutionary Computation. 1993; 1(1):1–23.
12. Del Moral P. Non linear filtering: Interacting particle solution. Markov Processes and Related Fields. 1996; 2(4):555–80.
13. Mohamed NA, Sameh MY, Nevin M, Aly AF, Thomas M. A Modified Fuzzy C-Means Algorithm for Bias

Field Estimation and Segmentation of MRI Data. IEEE Transactions on Medical Imaging. 2002; 21(3):193–9.
14. Lingras P, Peters G. Applying rough set concepts to clustering, rough sets: Selected methods and applications in management and engineering. Springer; 2012.
15. Peters G, Weber R, Nowatzke R. Dynamic rough clustering and its applications. Journal of Applied Soft Computing. 2012; 12(2012):3193–207.
16. Nock R, Nielsen F. On weighting clustering. IEEE Trans on Pattern Analysis and Machine Intelligence. 2006; 28(8):1–13.
17. Subbalakshmi C, Ramakrishna G, Rao SKM, Rao PV. A method to find the optimal number of clusters based on fuzzy silhouette on dynamic data set. ICICT; 2014 Dec 3-5. Procedia Computer Science. 2015; 46:346–53.