Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 346 – 353

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

A Method to Find Optimum Number of Clusters Based on Fuzzy Silhouette on Dynamic Data Set

Chatti Subbalakshmi[a,*], G Rama Krishna[b], S Krishna Mohan Rao[c], P Venketeswa Rao[d]

[a]*Guru Nanak InstitutionsTechnical Campuc,Dept. of CSE, Hyderabad, Telangana, India*
[b]*K L University, Dept. of CSE, Vijayawada, AP, India*
[c]*SiddharthaEngineering College, Dept. of CSE, Hyderabad, India*
VNR VignanaJyothi Institute of Engineering and Technology, Dept. of CSE, Hyderabad, India

Abstract

Data analysis plays amajor role in innovation of new trends in many applications.In most of the current applicationdatabases is being updated day to day. In order to adopt these changes, there is a need to update the present technologies and data mining algorithms in support of changing data. In many of the clustering algorithms the user has to specify the optimum number of clusters prior to execution, for static databases this value remains constant whereas, in the case of dynamic databases the value should be changed. In this paper,we implemented a method to find optimal number of clusters based onfuzzy silhouette on dynamic data by comparing traditional clustering on synthetic data and dynamic customer segmentation.

Keywords:cluster analysis;dynamic data; optimum number of clusters;fuzzy silhouette index;

1. Introduction

In recent years all applications are online and the data is changing over the time. For finding new trends on these data bases most of the traditional data mining algorithms[1] are not suitable. As they take static inputs which is not changed over data changes. Here it takes new research area that is dynamic data mining[2, 3]which is a combination of traditional data mining techniques with dynamic characteristics. In classical data mining, the data segmentation is

[*] Chatti Subbalakshmi. Tel. : 9032312260;
E.-mail address: subbalakshmichatti@gmail.com

doi:10.1016/j.procs.2015.02.030

one of the popular tasks to support the different application requirements. The data segmentation is process of grouping given data set into segments also called as clusters and assigns class label. The other name of data segmentation is cluster analysis[4].

Many clustering strategies are availablein the literature. One of the categories of clustering is partition method, where the given data base is divided into specific number of partitions. The k-means and k-mediod algorithms are popular methods in partition clustering techniques[5]. These methods perform the clusters based on center of clusters. These methods need to mention number of clusters (*k*) prior to execution of algorithm. The user needs to have knowledge on application data set to decide optimal number of cluster (k) value based on the type of data sets. Many methods are present to decide the right number of clusters, some of them are Thumb rule[6], Cross-Validation[7], the Elbow method[8], information criterion approach[9], kernel matrix[10].Most of the methods need to perform clustering process to decide the right number of clusters.

In this paper, we have selected Silhouette cluster validity criterion index to find the right number of clusters.For static data bases, the optimal number of clusters is always a static value. So one can use hard Silhouette value. Whereas dynamic data is uncertain,henceone uses fuzzy to represent the uncertainty on dynamic data. We find fuzzy silhouette index criterion to decide the optimum number of clusters on dynamic data and perform the clustering.Section 2.1 contains related work on classic k-means clustering and section 2.2 about Silhouette index. In Section 3,we present a general view of our method in 3.1, detail steps in 3.2. The application of the proposed method on simulated data and dynamic customer segmentation results is present in section 4.1 and 4.2.

2. Related Work

Many data mining algorithms have been developed over the past few decades of which one of the popular classic clustering algorithms is k-means partition clustering algorithm. Due its simplicity it is convenient for on-line applications. Many new approaches are defined in literature to make the k-means into dynamic aspect[11, 12]. Hence we consider classic or static k-means algorithm, its merits and demerits in section 2.1. In the next section, detail description about the Silhouette cluster validity index is provided.

2.1. Hard K-means algorithm

In hard clustering, data set is divvied into distinct clusters, where each data object belongs to exactly one cluster. One of hard clustering algorithm is k-means. The k-means algorithm was first proposed by Stuart Lloyd in 1957 as technique for pulse-code modulation. First k-means was used by James Macqueen in 1967[13].K-means partition the data into k number of clusters based on similarity between the data objects. It uses the distance measure to find the similarity and it needs the number of cluster in prior to execution of algorithm. This considered as major problem of k-means algorithm, due to user has to work out on k value and fix it. By using any cluster validity index user can decide right number of clusters. K value is dependent on characteristics of data set and it may change over data sets.The k-means algorithm is simple and time efficient. The complexity of algorithm is $O(ntk)$, where n is size of data set, $t$is the number of iterations and k is the number of clusters. As the k value increases, the number of iterations also increases. But it is suitable to online or dynamic data due to its simplicity and efficiency. Here we call it static, as the major input of algorithm is number of clusters (*k*) which is constant even when there is change in data. Hence, the user needs to identify the changes in the data and then he has to find the right number of clusters on current data, finally he has to execute the algorithm. This process repeats when there is change in data.

2.2. Fuzzy clustering

In fuzzy clustering algorithm, data can belongs to more than one cluster and the membership value represents the association with clusters and data point. In this clustering, user has to find the membership values for data point and using them to assign data point to one or more clusters. One of the popular fuzzy clustering is the Fuzzy C-Means (FCM) algorithm (Bezdek in 1981)[15] and it is mostly used on uncertain data sets.

2.3. Silhouette cluster validity index

Silhouette was first described by Peter J. Rousseeuw in 1986[14]. It is method of interpretation and validation of crisp cluster data. This technique provides the graphical representation of how well each object lies within its cluster. The silhouette value for an attribute *i*given in equation.1

$$s(i) = \frac{a(i)b(i)}{\max\{a(i),b(i)} \quad (1)$$

Where *a(i)* is the average dissimilarity of i^{th} data point with all other data within the same cluster. The *b(i)* be the minimum average dissimilarity of i^{th} data point to any other cluster which *i*is not a member. The *s(i)* is between -1 and +1.If s(i) value is +1, then it shows the data point is correctly clustered and its value is near to -1 then *i* would be more appropriate if it was clustered in its neighboring cluster. The average *s(i)* over all data of a cluster is a measure of how tightly grouped all the data in the cluster are. The average *s(i)* over all data of the entire data set is measure of how correctly the data has been clustered. If there are too many or too few clusters, as may occur when a poor choice of k is used in the k-means algorithm, some of clusters will typically display much narrow silhouettes than the rest. Thus silhouette plots and averages may be used to determine the natural number of clusters within a data set. The average silhouette values are works on crisp cluster boundaries like k-means clustering algorithm which is for hard clustering. For fuzzy clustering, the extended average silhouette index proposed in (Campello&Hruschka 2006) which integrates fuzzy values with silhouettes. Afterwards the generalized intra-inter silhouette index to fuzzy clustering proposed in (Rawashdeh&Ralescu).

3. Proposed method

In this section, we present a method to find the optimal number of clusters based on fuzzy silhouette cluster validity index on dynamic data. By study of literature, we selected the fuzzy clusteringdue to the degree of membership of an object to the cluster found provides strong tool for changing data sets. To find optimal number of cluster on dynamic data we use generalized intra-inter silhouette index.

3.1Global view of the proposed method

Here we present the overall steps to perform the fuzzy clustering at optimal number of cluster on dynamic data set.Execute the following steps for number of clusters (k) = 2 to (n-1)/2, where n is the size of data set.

step1: Perform initial fuzzy-c means clustering with k number of clusters.
Step 1: Calculate the intra cluster similarity intra$_i$using equ.2
Step 2: Calculate inter cluster similarity to other clusters inter$_i$equ.3
Step 3: Find the compactness distance a$_j$using equ .4
Step 4: Find the separation distance b$_j$ using equ.5
Step 5: Calculate intra-inter Silhouettes value using equ.6
Step 6: Calculate cluster average silhouette using equ.7
Step 7: Find optimal cluster number at maximum cluster average silhouette
Step 8: Display cluster results.

3.2 Detailed description of proposed method

Here we present mathematical equations for each step to find calculations as mentioned in previous section.
The following notations are used in equations:
c:number of classes
n:number of objects
x_j: feature vector of object *j*, *j= 1,…,n*
u_{ij} degree of membership of object j to class*i*, *i=1…c; j=1…n*

Let $d_{jk} = d(x_j, x_k)$denote the distance between the data points x_j and x_k; $1<=j \neq k <= n$. Let $u_i$denote a cluster, and u_{ij} be the membership of x_j to cluster u_i; $1<=i<=c$. The intra-score for $d_{jk}$with respect to cluster u_i is

$$intra_i(d_{jk}) = (u_{ij} \wedge u_{ik}) \qquad (2)$$

The inter-score for d_{jk} with respect to cluster u_r and u_s ; $1 \leq r < s \leq c$, is

$$inter_{rs}(d_{jk}) = (u_{rj} \wedge u_{sk}) \vee (u_{sj} \wedge u_{rk}) \qquad (3)$$

The compactness distance a_j *is,*

$$a_j = min\left\{\frac{\sum_{k=1}^{n} intra_{\,i}(j,k).d_{jk}}{\sum_{k=1}^{n} inter_{\,rs}(j,k)};\ \sum_{k=1}^{n} intra_i(j,k) > 0, 1 \leq i \leq c\right\} \qquad (4)$$

The separation distance b_j *is,*

$$b_j = min\left\{\frac{\sum_{k=1}^{n} inter_{\,rs}(j,k).d_{jk}}{\sum_{k=1}^{n} inter_{\,rs}(j,k)>0,};\ \sum_{k=1}^{n} inter_{rs}(j,k) > 0, 1 \leq r < s \leq c\right\}(5)$$

The silhouette value is,

$$S_j = \frac{b_j - a_j}{max\,\{a_j, b_j\}}, -1 < s_j < +1 \qquad (6)$$

The average silhouette value is,

$$Sil(x) = \sum_{j=1}^{n} \frac{s_j}{n} \qquad (7)$$

The average silhouette value gives the overall clustering quality of the entire data set. It returns the vector of silhouette values, one for each point in data set. If one point has a silhouette value near 1, then it is good clustering. If the silhouette value is near to -1 indicates the poor clustering and finally if its value is 0 shows the intermediate case.

4. Application of the proposed dynamic clustering to random sample data

We have implemented the proposed dynamic clustering in R, in order to demonstrate its effectiveness in different applications. In this section, we generated random sample data with size of 100 objects and each object is described by two features called (x , y). We executed sample data using k-means clustering, fuzzy clustering and compared with our dynamic clustering algorithm.

4.1 Proposed algorithm to simulated data

We executed proposed algorithm in R to show the effectiveness of dynamic clustering by comparing hard clustering algorithm. To simulate dynamic behavior of data set, initially we generated 50 random samples objects with two features in first cycle. In each cycle we added 50 more new objects generated from the subset of initial data set.

Cluster results in Cycle 1
From the results table.1, we identified that optimal number of clusters are 3 with maximum silhouette cluster validity index value is 0.9610195 and next best fit clusters are 18 with index value is 0.9131944.

Table.1 Fuzzy cluster average silhouette values for initial data set with 50 objects

Number of clusters	Cluster Avg silhouette	Number of clusters	Cluster Avg silhouette	Number of clusters	Cluster Avg silhouette
2	0.1318083	10	0.1543917	18	**0.9131944**
3	**0.9610195**	11	0.8023529	19	0.8912467
4	0.5619048	12	0.0000000	20	0.6548875
5	0.4375000	13	0.0000000	21	-0.2499395
6	-0.5135135	14	0.0000000	22	0.2729979
7	0.6392525	15	0.6751959	23	-0.5588235
8	0.6788165	16	0.1025223	24	0.0000000
9	0.8324176	17	-0.3000000		

Cluster results in Cycle 2

In cycle two 100 new objects are generated and our proposed method is applied. From results table -2, it shows that the maximum silhouette value 1 at three optimal numbers of clusters is 5, 36 and 46.

Table.2Fuzzy cluster average silhouette values for 100 objects

Number of clusters	Cluster Avg silhouette	Number of clusters	Cluster Avg silhouette	Number of clusters	Cluster Avg silhouette
2	0.850000	18	0.312959559	34	0.850000000
3	0.000000	19	0.461700405	35	-0.062500000
4	0.000000	20	0.936538462	36	1.000000000
5	**1.000000**	21	0.311538462	37	-0.201010101
6	0.000000	22	0.723076923	38	0.858333333
7	0.000000	23	0.314285714	39	0.000000000
8	-0.227778	24	0.833333333	40	0.101037851
9	0.6466081	25	0.858307534	41	-0.513227513
10	0.5263533	26	0.000000000	42	0.000000000
11	0.2827313	27	0.187500000	43	0.000000000
12	0.1776765	28	-0.054511369	44	0.000000000
13	0.7979798	29	0.000000000	45	0.553846154
14	0.0023331	30	0.849267873	46	1.000000000
15	0.0000000	31	-0.409343434	47	0.000000000
16	-0.290741	32	0.411538462	48	0.891812865
17	0.0000000	33	0.000000000	49	0.000000000

Cluster results in Cycle 3

In cycle three with 150 data objects giving 21 optimal number of clusters with maximum silhouette index.

Table.3 Fuzzy average silhouette values for initial data set with 150objects

Cluster No.	**Fuzzy Avg. Silouthe Value**					
Clus 2-7	-0.27777778	0.00000000	0.62750000	0.89181880	0.27125506	0.00000000
clus 8-13	0.74195906	0.05314103	0.51583333	0.85000000	0.01481481	0.00000000
clus 14-19	0.32962963	0.00000000	0.00000000	0.00000000	-0.23461538	0.10833333
clus 20-25	0.00000000	**1.00000000**	0.00000000	-0.05632716	0.53247549	0.07273148
clus 26-31	0.35714286	0.00000000	0.00000000	0.55555556	0.00000000	0.18571429
clus 32-37	0.74000523	0.15769231	0.43137255	0.32708357	0.00000000	0.00000000
clus 38-43	0.38811189	0.93667109	0.56807041	0.22452287	0.00000000	0.70909091
clus 44-49	0.00000000	0.79047619	0.00000000	0.65000000	1.00000000	0.00000000
clus 50-55	0.73140306	0.00000000	0.25000000	0.19318182	0.00000000	0.78478261
clus 56-61	0.00000000	0.86666667	0.00000000	0.18750000	1.00000000	0.00000000
clus 62-67	-0.55555556	0.00000000	1.00000000	0.00000000	-0.30000000	0.53598485
clus 68-73	0.00000000	-0.13194444	-0.41942971	0.25000000	0.60000000	0.00000000
clus 74-75	0.55769231	0.00000000				

From Table.4 it observed that soft silhouette value giving maximum for different sizes of dynamic data sets. As data size increases it is converged at maximum value 1 with optimal number of clusters.

Table.4Optimum number of clusters

Data size	Fuzzy Avg silhouette	Number of clusters	Hard avg silhouette	Number of clusters
50	0.9610195	03	1.0000000	06
100	1.000000000	05	0.75230765	30
150	1.000000000	22	1.0000000	22
200	1.000000000	08	1.000000000	08

4.2 Proposed algorithm to Dynamic customer segmentation

In this section, we haveshown the effectiveness of the proposed algorithm presenting on dynamic customer segmentation. Customer segmentation is major requirement for customer relationship management. Several methods are present in data mining to study customer behavior. The changes in the customer number and changes in the cluster structure show changes in Customer behavior. We simulated the wholesale customer data as dynamic by taking initial sample data from complete data and then for each iteration we added subset of data. Here we show the execution of algorithm for each cycle in tables.

Customer segmentation in cycle 1results

We generated random sample size of 32 objects from total customer data set and executed algorithm for number of clusters k=2 to (n-1)/2-1 =2 to (32-1)/2=2 to 15 using both algorithms PAM and Fuzzy clustering using Silhouette validity index. From the results of fuzzy silhouette clustering, we identified the optimal number of clusters as 10 with cluster average silhouette value 0.75935926 which is the maximum than remaining clusters from Table 5.

Table.5Fuzzy average silhouette values for initial data set with 32 objects

Clusters No	Cluster Avg silhouette	Cluster No	Cluster Avg silhouette	Cluster no	Cluster Avg silhouette	Cluster no	Cluster Avg silhouette
2	0.52008321	6	-0.15736384	10	**0.75935926**	14	0.52893654
3	-0.24720373	7	-0.16791357	11	0.253688212	15	0.31783665
4	0.00000000	8	0.53237734	12	0.19782793	16	0.00000000
5	-0.06188254	9	0.20767991	13	0.00000000		

Cycle -2 results

In second cycle 58 objects randomly generated and apply the method we get results in table.

Table.6 fuzzy average silhouette values for initial data set with 38 objects

0.45300315	-0.09941426	0.00000000	0.44285917	-0.24212839	0.21139694
0.68337689	0.32628115	0.04530589	-0.08657911	0.62535395	0.45366632
0.79874486	0.00000000	0.17425593	0.00000000	0.86741678	0.00000000
0.00000000	0.77239167	0.89083778	0.55603774	0.26151961	0.00000000
0.00000000	0.00000000	0.18731023	0.00000000		

Cycle 3 results

In third cycle 104 objects are generated and apply the method results given in table. The optimal number of clusters is 2 and silhouette value is 1 which shows the good clustering. For this data set more than one cluster is giving silhouette value 1.

Table.7 fuzzy average silhouette values for initial data set with 104 objects

1.0000000	1.0000000	1.0000000	1.0000000	0.5860130	0.7223891	1.0000000	1.00000
1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.00000
1.0000000	0.5777057	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.00000
1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.00000
0.3026127	1.0000000	0.8456069	1.0000000	0.0000000	0.3856016	0.0000000	0.00000
0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.00000
0.0000000	0.0000000	0.0000000					

Cycle 4 results

In third cycle 200 objects generated and apply the method we get results in table. The optimal number of cluster is 56 with 0.77593793 silhouette value.

Table.8 Fuzzy average silhouette values for initial data set with 200 objects

0.55976279	0.00000000	0.37925663	-0.05623117	0.03260471	-0.01628974
-0.16729967	-0.16229263	-0.28324568	0.00000000	-0.31690776	0.00000000
-0.12482701	0.56795609	-0.03906255	0.00000000	0.19008256	-0.09519126
0.25900301	-0.09910093	0.00000000	-0.29432349	0.00000000	0.65023187
-0.27729190	0.00000000	-0.07184749	-0.10998568	-0.23884071	0.34929264
0.11552685	0.04842694	0.25841285	0.63720793	0.00000000	0.06959880
0.00000000	0.00000000	0.00000000	-0.12218672	0.00000000	0.00000000
-0.24194879	0.00000000	0.00000000	0.29320750	0.00000000	-0.29316399
0.44807941	0.20297669	-0.30465480	0.00000000	0.00000000	0.00000000
0.77593793	-0.16698539	0.18234737	0.00000000	0.29380706	0.49526349
0.00000000	0.00000000	0.14479465	0.00000000	0.53172394	0.00000000
0.00000000	0.00000000	0.52536294	0.05599229	0.00000000	0.26516607
0.00000000	0.00000000	0.23757874	0.00000000	0.00000000	0.00000000
0.00000000	0.00000000	0.00000000	-0.26678068	0.00000000	0.00000000
0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000

From table. 9, results show the comparison between the crispand fuzzy silhouette values. We identified that the generalized fuzzy average silhouette values are maximum ascompare to hard silhouette in all cases of dynamic data set.

Table.9Optimum number of clusters

Data size	Fuzzy Avg silhouette	Number of clusters	Hard avg silhouette	Number of clusters
32	**0.75935926**	10	0.617862208	10
58	**0.89083778**	22	0.64285909	20
104	**1.00000000**	02	1.00000000	02
200	**0.77593793**	56	0.71404183	44

5. Conclusion and Future work

In this paper we consider the optimal number of clusters in dynamic environment. As data changes, mining algorithm also have to adopt those changes and it has to modify its parameters. We select partition clustering algorithms, where user has to give number of clusters in prior. If data is uncertain in some cases, number of clusters can be decided at run time. In this view we proposed a method to find the optimal number of clusters for agiven dynamic data set and display the results. We can further reduce the complexity of method without executing entire process when ever new data arrived as feature work.

References

1. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, 2001.
2. Raghavan, A. Hafez, Dynamic data mining, in: R. Loganantharaj, G. Palm, M. Ali (Eds.), Intelligent Problem Solving— Methodologies and Approaches: Proc. of Thirteenth International Conference on Industrial Engineering Applications of AI & Expert Systems, Springer, New York, 2000, pp. 220–229.
3. Fernando Crespo, Richard Weber, A methodology for dynamic data mining based on fuzzy clustering, Fuzzy Sets and Systems, volume 150, issue 2, 1 March 2005, pages 267-284.
4. (1939). *Cluster Analysis: Correlation Profile and Orthometric (factor) Analysis for the Isolation of Unities in Mind and Personality*. Edwards Brothers.
5. J. A J.A. Hartigan (1975). *Clustering algorithms*. John Wiley & Sons, Inc.
6. KantiMardia et al. (1979). *Multivariate Analysis*. Academic Press.
7. "Finding the Right Number of Clusters in k-Means and EM Clustering: v-Fold Cross-Validation". *Electronic Statistics Textbook*. StatSoft. 2010. Retrieved 2010-05-03.
8. Robert L. Thorndike (December 1953). "Who Belong in the Family?". *Psychometrika* 18 (4): 267–276.
9. Catherine A. Sugar and Gareth M. James (2003). "Finding the number of clusters in a data set: An information theoretic approach". *Journal of the American Statistical Association* 98 (January): 750–763.
10. Honarkhah, M and Caers, J (2010). "Stochastic Simulation of Patterns Using Distance-Based Pattern Modeling". *Mathematical Geosciences* 42 (5): 487–517.
11. C.-Y. Lee, E.K. Antonsson, Dynamic partitional clustering using evolution strategies, Proc. of the Third Asia Paci4c Conf. on Simulated Evolution and Learning, Nagoya, Japan, 25–27 October 2000.
12. Aaron, B. ; Dept. of Comput. Sci., Texas State Univ., San Marcos, TX, USA ; Tamir, D.E. ; Rishe, N.D. ; Kandel, A., Dynamic incremental k-means clustering, CSCI 2014 International conference.
13. Lloyd, S. (1982). "Least squares quantization in PCM". *IEEE Transactions on Information Theory*28 (2): 129–137. doi:10.1109/TIT.1982.1056489.
14. Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". *Computational and Applied Mathematics*20: 53–65. doi:10.1016/0377-0427(87)90125-7
15. Ahmed, Mohamed N.; Yamany, Sameh M.; Mohamed, Nevin; Farag, Aly A.; Moriarty, Thomas (2002). "A Modified Fuzzy C-Means Algorithm for Bias Field Estimation and Segmentation of MRI Data". *IEEE Transactions on Medical Imaging*21 (3): 193–199. doi:10.1109/42.996338