Combining K-Harmonic mean and Hierarchical algorithms for robust and efficient data clustering with cohesion self-merging

Jyothi Bankapalli,
Asst.Professor
Department of computer science and engineering,
Gayatri Vidya Parishad College of Engineering for Women,
Andhra Pradesh, India.
jyothi.maya@gmail.com
R.Venu Babu,
Asst.Professor,
Department of Information Technology,
GITAM University,
Andhra Pradesh, India.
Venubabu.r@gmail.com
S. Anjali Devi,
Asst.Professor,
Department of Computer Science & Engineering,
Gayathri Vidya Parishad College of Engineering for Women,
Andhra Pradesh, India.
swarnaanjalidevi@gmail.com

Abstract-- **A cluster is a collection of data objects that are similar to one another within the cluster and are dissimilar to the objects in the other cluster. Data clustering is studied in statistical, machine learning and data mining. The dissimilarity between two clusters is defined as the distance between their centroids or the distance between two closest (or farthest) data points. The measures are vulnerable to outliers and removing the outliers is yet another difficult task. A new similarity measure, cohesion is used to measure the inter cluster distance. By using cohesion measure, a new two phase clustering algorithm is designed called as "Cohesion Based Self Merging Algorithm (CSM)". This is a combination of partitional (K-Harmonic mean) clustering and hierarchical clustering (CURE) algorithms. CSM partitions the input dataset into several small clusters in the first phase and then continuously merges the sub clusters based on similarity measures cohesion in a hierarchical manner in the second phase. Run time behaviors of these algorithms are analyzed and compared using the existing method combining k-mean and hierarchical algorithms for robust and efficient data clustering with cohesion-self merging algorithm.**

Keywords: Data mining, K-harmonic means clustering algorithm, Hierarchical clustering algorithm, Cohesion Similarity measure.

I. INTRODUCTION

Data mining, is the process of automatically searching large volumes of data for patterns using tools such as classification, association rule mining, clustering. The term 'Data mining [1] refers to the finding of relevant and useful information from database. Data mining involves many different algorithms to accomplish different tasks. All these algorithms attempt to fit a model to the given data. The algorithms examine the data and determine a model that is closest to the characteristics of the data being examined. Clustering is similar to classification except that the groups are not predefined, but rather defined by data alone. Clustering is also known as Unsupervised Learning. Clustering means partitioning or segmenting the data into groups that might not be disjointed. The most similar data is grouped into clusters. Clustering is a process of partitioning a set of data (or objects) in a set of meaningful sub-classes called clusters. The goal of clustering is to discover both the dense and sparse regions in a dataset. Clustering is the classification of objects into different groups, or more

precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset share some common trait-often proximity according to some defined distance measure.
A good clustering method will produce high quality clusters in which

-The intra-class similarity is high.
-The inter-class similarity is low.

1.1 Categories of major clustering algorithms:
The major clustering methods can be classified into the following categories.

1.1.1 Partitioning method [2]: A partitioning method first creates an initial set of k partitions, where parameter K is the number of partitions to construct. Then it is uses an iterative relocation technique, which attempts to improve the partitioning by moving objects from one group to another group. Partitioning methods encompasses K-mean, K-medoids, K-Harmonic mean, CLARANS and their improvements.

1.1.2 Hierarchical method [3]: This method creates a hierarchical decomposition of the given set of data objects. The method can be classified as either bottom-up (agglomerative) or top-down (divisive), it is based on how the decomposition is formed. To compensate for the rigidity of merge or split, the quality of hierarchical agglomeration can improved by analyzing object linkages at each hierarchical partitioning or by first performing micro clustering and then operating on the micro clusters with other clustering techniques like iterative relocation that is BRICH.

Hierarchical clustering algorithms:

There are two basic approaches for generating a hierarchical clustering.

- Agglomerative
- Divisive

1.1.2.1Agglomerative: starts with the points as individual clusters and, at each step, merges the closest pair of clusters. This requires defining a notion of cluster proximity.

1.1.2.2 Divisive: start with one, all inclusive cluster and, at each steps, split a cluster until only singleton clusters of individual points remain. In this case, we need to decide which cluster to split at each step and how to do the splitting.

1.1.2.3 Single link clustering algorithm:
The single link [5] or MIN version of hierarchical clustering, the proximity of two clusters is defined as the minimum of the distance (maximum of the similarity) between any two points in the two different clusters. Using graph terminology, if we start with all points as singleton clusters and add links between points one at a time, shortest link first, then these single links combine the points into clusters. The single link technique is good at handling non-elliptical shapes, but is sensitive to noise and outliers.

1.2 Introduction to clustering attributes:

In general, there are two types of attributes associated with input data in clustering algorithms, that is Numerical attributes and Categorical attributes.

- Numerical attributes are those with a finite or infinite number of ordered values, such as the height of a person or the x-coordinate of a point on a 2D domain.
- Categorical attributes are those with finite unordered values, such as the occupation or the blood type of a person. In this paper the focus is on the numerical data.

TABLE1: Example for categorical attributes:

Sno	**Age**	**Sex**	**Region**	**Income**	**Married**	**children**	**Car**	**save act**	**current act**	**Mortgage**	**pep**
12101	48	Male	A	17546	No	1	Yes	yes	yes	Yes	yes
12102	40	female	B	30085.1	Yes	3	No	yes	no	No	yes
12103	51	Male	A	16575.4	Yes	0	No	no	no	Yes	yes
12104	23	Male	B	20375.4	Yes	3	Yes	yes	no	Yes	yes
12105	57	Male	C	50576.3	Yes	0	Yes	no	yes	Yes	yes
12106	57	Male	B	37869.6	Yes	2	Yes	no	no	Yes	no
12107	22	female	C	8877.07	No	0	Yes	yes	no	Yes	no
12108	58	female	B	24946.6	Yes	0	no	no	no	Yes	yes
12109	37	Male	C	25304.3	Yes	2	no	yes	yes	Yes	yes

TABLE 2: Example for Numerical attributes

SNO	**Age**	**Sex**	**region**	**income**	**Married**	**children**	**car**	**save act**	**current act**	**Mortgage**	**pep**
12101	48	0	0	17546	0	1	0	0	0	0	1
12102	40	1	1	30085.1	1	3	1	0	1	1	0
12103	51	0	0	16575.4	1	0	1	1	1	0	0
12104	23	0	1	20375.4	1	3	0	0	1	0	0
12105	57	0	2	50576.3	1	0	0	1	0	0	0
12106	57	0	1	37869.6	1	2	0	1	1	0	1
12107	22	1	2	8877.07	0	0	0	0	1	0	1
12108	58	1	1	24946.6	1	0	1	1	1	0	0
12109	37	0	3	25304.3	1	2	1	0	0	0	0

1.3 Similarity Measure:

Distance is a numerical description of how far apart objects are. Distance measure determines the similarity of two elements. There are number of distance measures available. Euclidean measure, Manhattan distance, and hamming distance. several alternatives, such as the distance between the closest (or farthest) points of the two clusters are proposed. However, those measures are very vulnerable to random noises (outliers). Consequently, we propose a new similarity measure, namely, cohesion, based on the joinability of two clusters, referring to the existence of a data point. Conceptually, joinability is the merging inclination of two clusters according to the existence of a shared data point.

II. PROBLEM DEFINITION

Clustering is the classification of objects into different groups or subsets (clusters), so that the data in each subset share some common traits or interesting patterns. The K Means (KM) algorithm is a popular algorithm which attempts to find K-clusters. K-Mean algorithm is a center-based clustering algorithm. Mean of the data is used to find the clusters. Expectation Maximization is another algorithm to find out the clusters. In K-Mean and Expectation Maximization algorithms initialization of the centers is a major problem the above problem is minimized by a new clustering method called the K-Harmonic Means algorithm (KHM). KHM is a center-based clustering algorithm which uses the Harmonic Average of the distances from each data point to find the clusters. K-Harmonic Means essentially insensitive to the initialization of the centers. There are number of similarity measures available but these are vulnerable for detecting outliers. A new measure cohesion is used to measure the inter cluster distance. Using cohesion similarity measure a new two phase algorithm is designed called cohesion based self merging algorithm.

Data sets are taken and clusters are obtained for the data by using both k-mean and K-harmonic means algorithms in the first phase and are compared. In this stage the clustering of elements are different because of the different centers. The resulting clusters are arranged in hierarchical format using Clustering Using Representatives (CURE) algorithms and outliers that are not similar to the center are detected.

III. EVOLUTION

3.1Cohesion-Based Self Merging Algorithm:

CSM [9] is a combination of partitional that is K-Harmonic means algorithm and hierarchical that is clustering using representatives.

Input: The input dataset, the size of the dataset is ‘n’, and the desired number of clusters ‘K’.

Output**:** The hierarchical representation of the ‘K’ clusters.

Method:

1) Apply k-harmonic means [1] on the input data set to obtain K sub clusters.
2) Apply CURE hierarchical clustering algorithm on the K sub clusters produced in step1 with cohesion as the similarity measure and stop when K clusters are obtained

3.2 Cohesion:

Cohesion is a new similarity measure [5], which is based on join ability of two clusters referring to the existence of a data point.

Join ability have the following properties.

1) Data points located closer to the boundary of the two clusters are more important
2) The merging inclination should not be determined by only a few points, I.e. the value of join ability should not vary dramatically.

This similarity measure of cohesion is robust to the existence of outliers due to the following reasons.

1) Using the cohesion measure, instead of only a few of points, all points of the two sub clusters are considered to evaluate the inter cluster similarity.
2) This measure makes the effect of outliers are much smaller than that of other points since outliers are much farther from the centroids of the two clusters.

$$\text{Chs}(Ci,Cj)=\frac{\Sigma_{p\,\epsilon\,Ci,Cj}\ \text{join}(p,Ci,Cj)}{|Ci|+|Cj|}$$

$C_{i,}$ C_j are clusters, p is a point on cluster.

Where |Ci| is the size of Cluster Ci and |Cj| is the size of cluster Cj.

3.3 K-Harmonic Means Algorithm:

KHM [1] is a center-based clustering algorithm which uses the Harmonic Averages of the distances from each data point to the centers as components to its performance function. It is demonstrated that K-

Harmonic Means is essentially insensitive to the initialization of the centers. In certain cases, K-Harmonic Means significantly improves the quality of clustering results comparing with both K-Means and *EM*, which are the two most popular clustering algorithms used in data exploration and data compression. KHM is essentially insensitive to the initialization of the centers than *KM* and *EM*. K-Harmonic Means (*KHM*), takes the sum over all data points of the Harmonic average of the squared distance from a data point to all the centers as its performance Function. It is different from the total with in cluster variance used by *KM*. Let *M = {ml | l=1… K}* be *K* centers and *S = {xi | i=1… N}* is *N* given data points.

K-Harmonic Means' performance function is

$$\mathrm{Perf}_{khm}\left(\{x_i\}^{n}_{i=1,}\ \{ml\}^{k}_{l=1}\right) = \sum_{i=1}^{N} \frac{K}{\sum_{l=1}^{K} \frac{1}{\|x_i - m_l\|}}$$

Harmonic average calculation:

$$\mathrm{HA}\left(\{a_i | i=1\ldots k\}\right) = \frac{K}{\sum_{i=1}^{K} \frac{1}{a_i}}$$

The harmonic average of *K* numbers is the reciprocal of the arithmetic average of the reciprocals of the numbers in the set.

3.4 CURE (Clustering Using Representatives) Algorithm:

CURE is a clustering algorithm [6] that uses a variety of different techniques to create an approach that can handle large data sets, outliers and clusters with non-spherical shapes and non uniform sizes. CURE represents a cluster by using multiple representative points from the cluster. These points will, in theory capture the geometry and shape of the cluster.

CURE Algorithm:

```
Input: A set of points S
Output: k clusters

Method:
1) For every cluster u (each input point), in u.mean and u.rep store the mean of the points in the cluster
   and a set of c representative points of the cluster (initially c = 1 since each cluster has one data point).
   Also u.closest stores the cluster closest to u.
2) All the input points are inserted into a k-d tree T
3) Treat each input point as separate cluster, compute u.closest for each u and then insert each cluster into
   the heap Q. (clusters are arranged in increasing order of distances between u and u.closest).
4) While size(Q) > k
5) Remove the top element of Q(say u) and merge it with its closest cluster u.closest (say v) and compute
   the new representative points for the merged cluster w. Also remove u and v from T and Q.
6) Also for all the clusters x in Q, update x.closest and relocate x
7) insert w into Q
8) repeat
```

IV. Datasets

4. Dataset:

Input: Large data set: Bank dataset, Adult dataset (Census Income) from UCI Repository datasets.

Type of input:

- Bank dataset is a categorical data we need to convert some of the fields in to numerical form.

Ex: sex is a column where it contains 2two options like male, and female. There are converted into numerical form as male as '0', female as'1'.

- Adult dataset is also called as census income. This is a multivariate type dataset, containing categorical integer data. This contains details of the persons. Like age, work class, education etc…Ex: if we take the education details bachelors, 11th, HS_grad,

Output:

Desired no of clusters, hierarchical form of k-mean and k-harmonic mean algorithms and outliers for each algorithm.

4.1 Introduction to Adult Data set:

Adult data set [8] which is based on census data, also known as 'Census Income' dataset. This is taken from the UCI Repository data. Data set is a multivariate data. There are 15 attributes and they are categorical, integer. Instances of 600.

Number of attributes: 6 continuous data, 8 nominal attributes.

Attribute information:

Age: continuous.
Work class: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
Fnlwgt:continuous.
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th,12th,Masters,1st-4th,10th,Doctorate,5th-6th,Preschool.
Education-num:Continuous.
Marital-status: Married-city-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
Occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
Relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
Race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
Sex: Female, Male.
Capital-gain: continuous.
Capital-loss: continuous.
Hours-per-week: continuous.
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holland-Netherlands.

4.2 Bank Dataset:

Bank dataset contains details of all the employees which contains 6 attributes and 30 instances

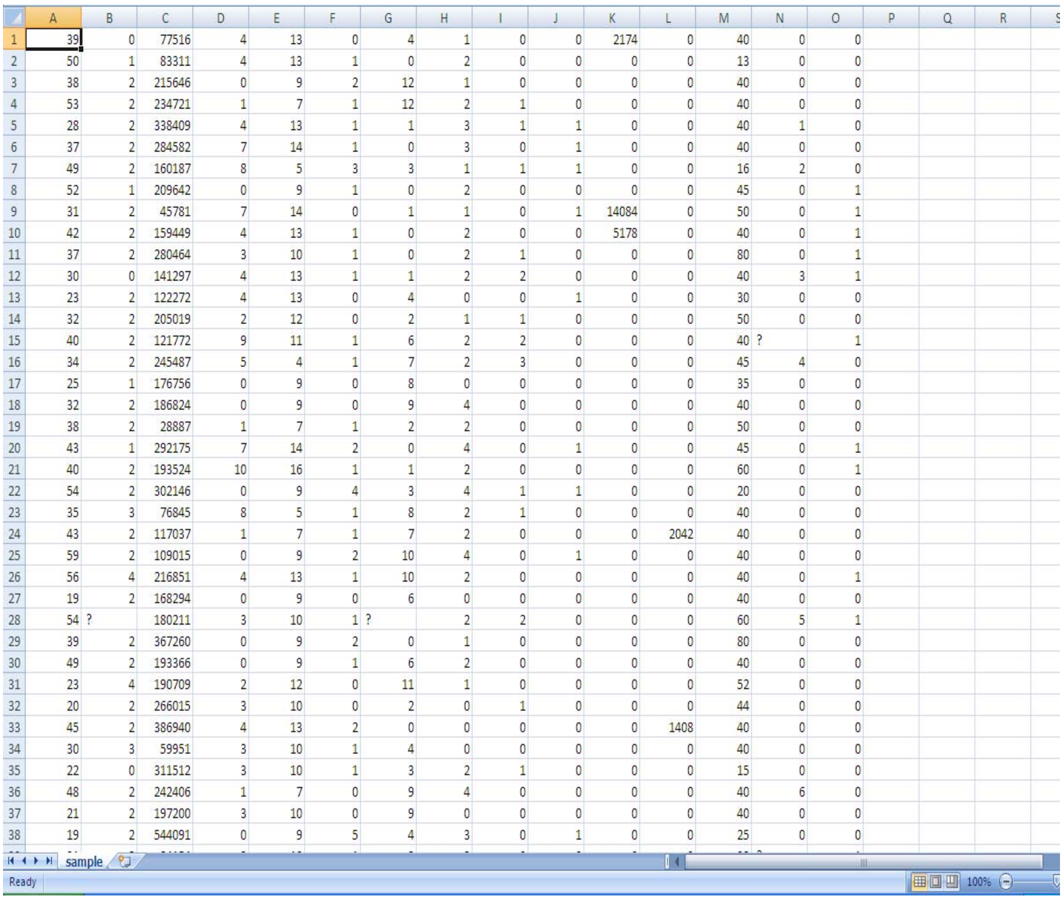
Attribute Information:
Empno: continuous
Ename: continuous
Designation: clerk, officer, cashier, assistance, probationary officer, manager.
Salary: continuous
Work-class: temporary, hourly based, daily based.
Date of joining: continuous

4.3 Introduction to (Comma Separated Value) .CSV fie

- CSV, comma separated values, files are commonly used to transport large amounts of tabular data between either companies or applications that are not directly connected. The files are easily editable using common spreadsheet applications like Microsoft Excel.

- Fields are separated by commas.

- Records are separated with system end of line characters, CRLF (ASCII 13 Dec or 0D Hex and ASCII 10 Dec or 0A Hex respectively) for Windows, LF for Unix, and CR for Mac.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	39	0	77516	4	13	0	4	1	0	0	2174	0	40	0	0			
2	50	1	83311	4	13	1	0	2	0	0	0	0	13	0	0			
3	38	2	215646	0	9	2	12	1	0	0	0	0	40	0	0			
4	53	2	234721	1	7	1	12	2	1	0	0	0	40	0	0			
5	28	2	338409	4	13	1	1	3	1	1	0	0	40	1	0			
6	37	2	284582	7	14	1	0	3	0	1	0	0	40	0	0			
7	49	2	160187	8	5	3	3	1	1	1	0	0	16	2	0			
8	52	1	209642	0	9	1	0	2	0	0	0	0	45	0	1			
9	31	2	45781	7	14	0	1	1	0	1	14084	0	50	0	1			
10	42	2	159449	4	13	1	0	2	0	0	5178	0	40	0	1			
11	37	2	280464	3	10	1	0	2	1	0	0	0	80	0	1			
12	30	0	141297	4	13	1	1	2	2	0	0	0	40	3	1			
13	23	2	122272	4	13	0	4	0	0	1	0	0	30	0	0			
14	32	2	205019	2	12	0	2	1	1	0	0	0	50	0	0			
15	40	2	121772	9	11	1	6	2	2	0	0	0	40	?	1			
16	34	2	245487	5	4	1	7	2	3	0	0	0	45	4	0			
17	25	1	176756	0	9	0	8	0	0	0	0	0	35	0	0			
18	32	2	186824	0	9	0	9	4	0	0	0	0	40	0	0			
19	38	2	28887	1	7	1	2	2	0	0	0	0	50	0	0			
20	43	1	292175	7	14	2	0	4	0	1	0	0	45	0	1			
21	40	2	193524	10	16	1	1	2	0	0	0	0	60	0	1			
22	54	2	302146	0	9	4	3	4	1	1	0	0	20	0	0			
23	35	3	76845	8	5	1	8	2	1	0	0	0	40	0	0			
24	43	2	117037	1	7	1	7	2	0	0	0	2042	40	0	0			
25	59	2	109015	0	9	2	10	4	0	1	0	0	40	0	0			
26	56	4	216851	4	13	1	10	2	0	0	0	0	40	0	1			
27	19	2	168294	0	9	0	6	0	0	0	0	0	40	0	0			
28	54	?	180211	3	10	1	?	2	2	0	0	0	60	5	1			
29	39	2	367260	0	9	2	0	1	0	0	0	0	80	0	0			
30	49	2	193366	0	9	1	6	2	0	0	0	0	40	0	0			
31	23	4	190709	2	12	0	11	1	0	0	0	0	52	0	0			
32	20	2	266015	3	10	0	2	0	1	0	0	0	44	0	0			
33	45	2	386940	4	13	2	0	0	0	0	0	1408	40	0	0			
34	30	3	59951	3	10	1	4	0	0	0	0	0	40	0	0			
35	22	0	311512	3	10	1	3	2	1	0	0	0	15	0	0			
36	48	2	242406	1	7	0	9	4	0	0	0	0	40	6	0			
37	21	2	197200	3	10	0	9	0	0	0	0	0	40	0	0			
38	19	2	544091	0	9	5	4	3	0	1	0	0	25	0	0			

Figure1: UCI Repository Adult Dataset

V. Results

1. Upload the input data file.

2. Enter the number of clusters K.(ex: k=7)

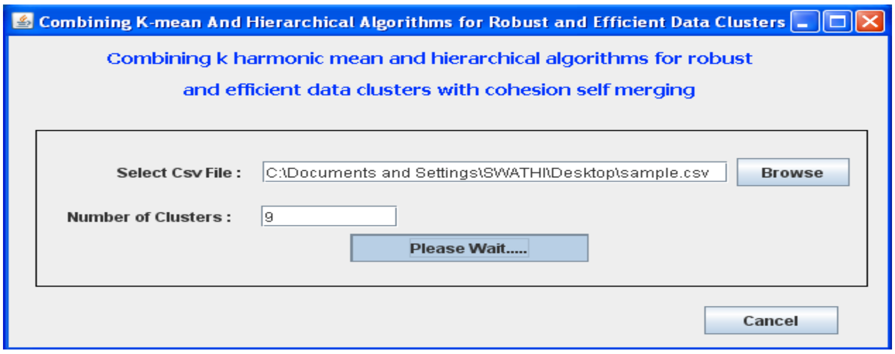
3. Input data is clustered into K sub clusters using k mean and k harmonic mean algorithms.

That is (c0, c1, c2, c3)

4. Obtained clusters are arranged into hierarchical format using CURE algorithm with the similarity measure cohesion.

5. Hierarchical form of k mean, k harmonic mean clusters are shown clusters are arranged in increasing order of distances between input point (u) and closest to u (u.closest).

6. Outliers are removed and shown.

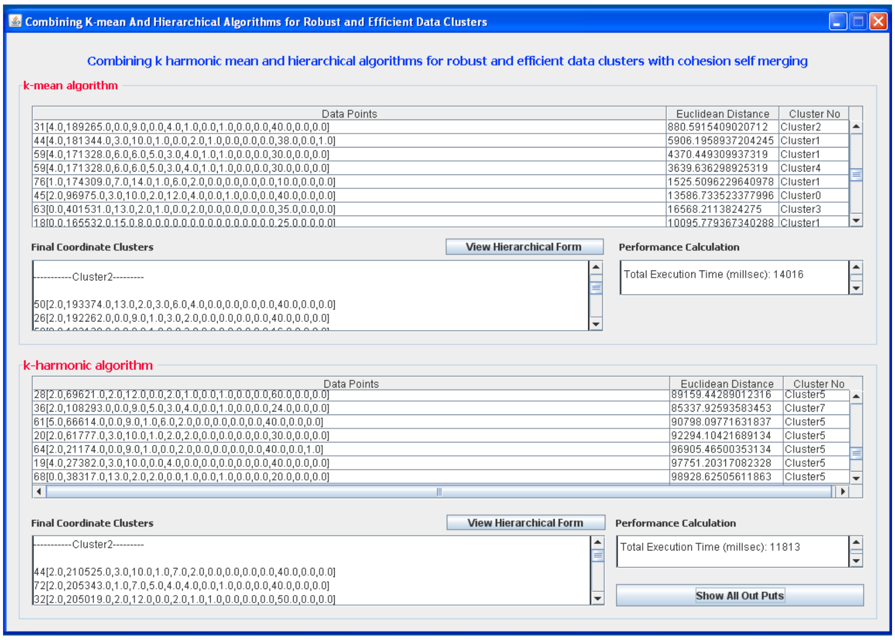

Combining K-mean And Hierarchical Algorithms for Robust and Efficient Data Clusters
Browse
Please Wait.....
Cancel

Combining k harmonic mean and hierarchical algorithms for robust and efficient data clusters with cohesion self merging

Select Csv File : C:\Documents and Settings\SWATHI\Desktop\sample.csv

Number of Clusters : 9

Figure2: uploading .CSV file and number of desired clusters


Combining K-mean And Hierarchical Algorithms for Robust and Efficient Data Clusters
View Hierarchical Form
View Hierarchical Form
Show All Out Puts

Combining k harmonic mean and hierarchical algorithms for robust and efficient data clusters with cohesion self merging

k-mean algorithm

Data Points	Euclidean Distance	Cluster No
31[4.0,189265.0,0.0,9.0,0.0,4.0,1.0,0.0,1.0,0.0,0.0,0.0,40.0,0.0,0.0,0.0]	880.5915409020712	Cluster2
44[4.0,181344.0,3.0,10.0,1.0,0.0,2.0,1.0,0.0,0.0,0.0,0.0,38.0,0.0,1.0]	5906.1958937204245	Cluster1
59[4.0,171328.0,6.0,6.0,5.0,3.0,4.0,1.0,1.0,0.0,0.0,0.0,30.0,0.0,0.0,0.0]	4370.449309937319	Cluster1
59[4.0,171328.0,6.0,6.0,5.0,3.0,4.0,1.0,1.0,0.0,0.0,0.0,30.0,0.0,0.0,0.0]	3639.636298925319	Cluster4
76[1.0,174309.0,7.0,14.0,1.0,6.0,2.0,0.0,0.0,0.0,0.0,0.0,10.0,0.0,0.0,0.0]	1525.5096229640978	Cluster1
45[2.0,96975.0,3.0,10.0,2.0,12.0,4.0,0.0,1.0,0.0,0.0,0.0,40.0,0.0,0.0,0.0]	13586.733523377996	Cluster0
63[0.0,401531.0,13.0,2.0,1.0,0.0,2.0,0.0,0.0,0.0,0.0,0.0,35.0,0.0,0.0,0.0]	16568.2113824275	Cluster3
18[0.0,165532.0,15.0,8.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,25.0,0.0,0.0,0.0]	10095.779367340288	Cluster1

Final Coordinate Clusters

```
-----------Cluster2---------

50[2.0,193374.0,13.0,2.0,3.0,6.0,4.0,0.0,0.0,0.0,0.0,0.0,40.0,0.0,0.0,0.0]
26[2.0,192262.0,0.0,9.0,1.0,3.0,2.0,0.0,0.0,0.0,0.0,0.0,40.0,0.0,0.0,0.0]
```

Performance Calculation

Total Execution Time (millsec): 14016

k-harmonic algorithm

Data Points	Euclidean Distance	Cluster No
28[2.0,69621.0,2.0,12.0,0.0,2.0,1.0,0.0,1.0,0.0,0.0,0.0,60.0,0.0,0.0,0.0]	89159.44289012316	Cluster5
36[2.0,108293.0,0.0,9.0,5.0,3.0,4.0,0.0,1.0,0.0,0.0,0.0,24.0,0.0,0.0,0.0]	85337.92593583453	Cluster7
61[5.0,66614.0,0.0,9.0,1.0,6.0,2.0,0.0,0.0,0.0,0.0,0.0,40.0,0.0,0.0,0.0]	90798.09771631837	Cluster5
20[2.0,61777.0,3.0,10.0,1.0,2.0,2.0,0.0,0.0,0.0,0.0,0.0,30.0,0.0,0.0,0.0]	92294.10421689134	Cluster5
64[2.0,21174.0,0.0,9.0,1.0,0.0,2.0,0.0,0.0,0.0,0.0,0.0,40.0,0.0,1.0]	96905.46500353134	Cluster5
19[4.0,27382.0,3.0,10.0,0.0,4.0,0.0,0.0,0.0,0.0,0.0,0.0,40.0,0.0,0.0,0.0]	97751.20317082328	Cluster5
68[0.0,38317.0,13.0,2.0,2.0,0.0,1.0,0.0,1.0,0.0,0.0,0.0,20.0,0.0,0.0,0.0]	98928.62505611863	Cluster5

Final Coordinate Clusters

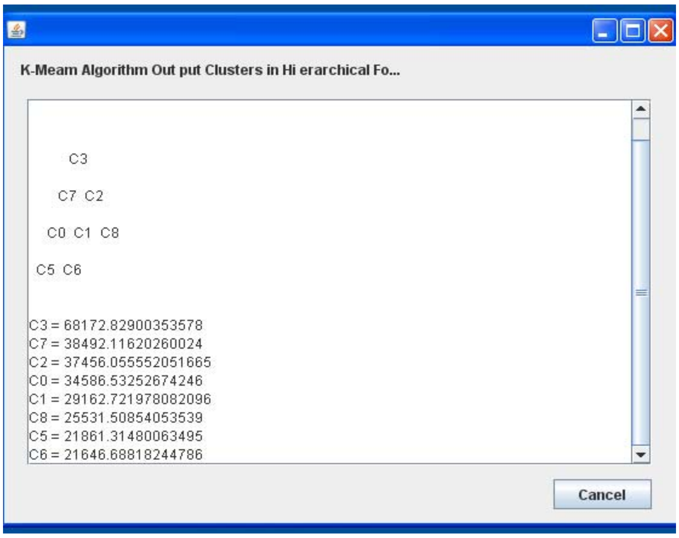
```
-----------Cluster2---------

44[2.0,210525.0,3.0,10.0,1.0,7.0,2.0,0.0,0.0,0.0,0.0,0.0,40.0,0.0,0.0,0.0]
72[2.0,205343.0,1.0,7.0,5.0,4.0,4.0,0.0,1.0,0.0,0.0,0.0,40.0,0.0,0.0,0.0]
32[2.0,205019.0,2.0,12.0,0.0,2.0,1.0,1.0,0.0,0.0,0.0,0.0,50.0,0.0,0.0,0.0]
```

Performance Calculation

Total Execution Time (millsec): 11813

Figure3: Data is clustered into K no of clusters using K-mean and K-Harmonic means algorithms

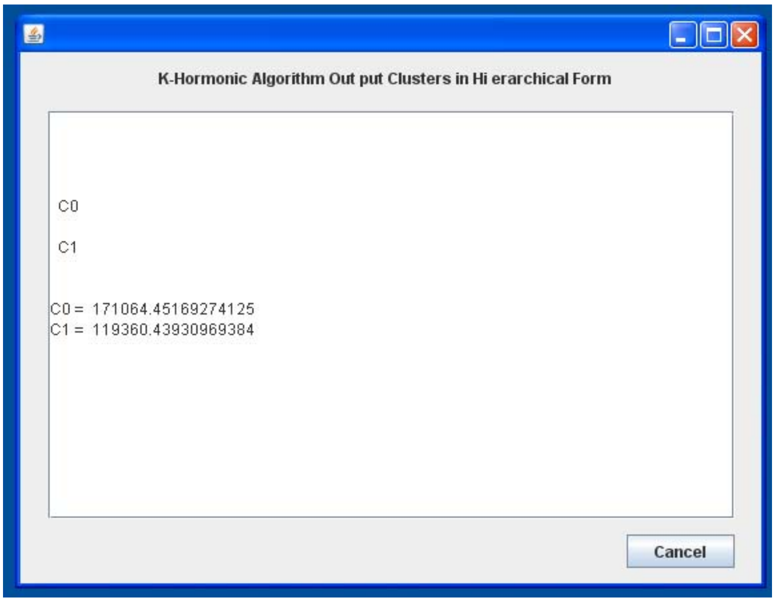


K-Meam Algorithm Out put Clusters in Hi erarchical Fo...

```
    C3
   C7  C2
  C0  C1  C8
 C5  C6

C3 = 68172.82900353578
C7 = 38492.11620260024
C2 = 37456.055552051665
C0 = 34586.53252674246
C1 = 29162.721978082096
C8 = 25531.50854053539
C5 = 21861.31480063495
C6 = 21646.68818244786
```

Figure4: Hierarchical representation of clusters using CURE algorithm



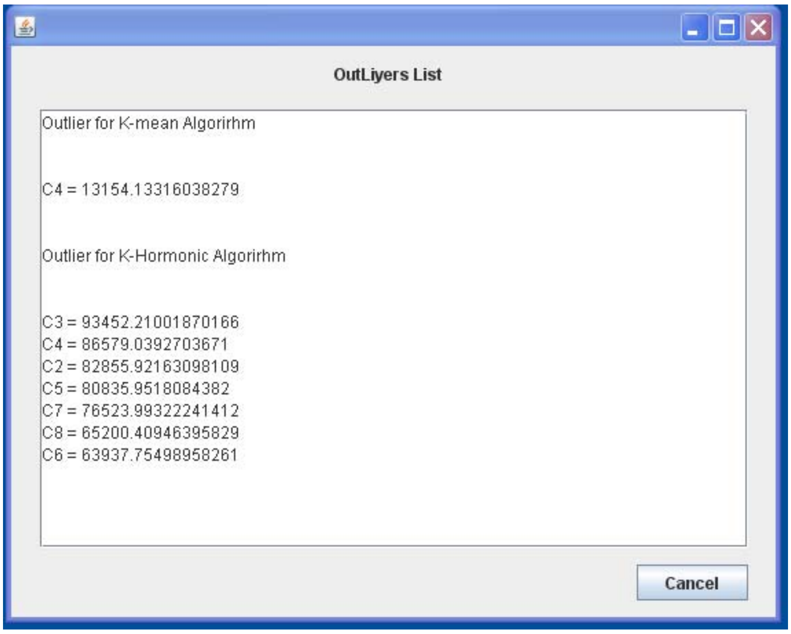
K-Hormonic Algorithm Out put Clusters in Hi erarchical Form

```
 C0

 C1


C0 =  171064.45169274125
C1 =  119360.43930969384
```

Figure5: Hierarchical representation of clusters obtained from the K-Harmonic mean algorithm.



OutLiyers List

```
Outlier for K-mean Algorirhm


C4 = 13154.13316038279


Outlier for K-Hormonic Algorirhm


C3 = 93452.21001870166
C4 = 86579.0392703671
C2 = 82855.92163098109
C5 = 80835.9518084382
C7 = 76523.99322241412
C8 = 65200.40946395829
C6 = 63937.75498958261
```

Figure6: Outlier detection

Summarized Results:

Algorithm	**No of elements**	**Choosing of center**	**Cluster formation**	**Time taken for clustering**	**Type of file format taken**
K-means	600 rows, 15 columns	Randomly selected	Instance 2,3,5,7 comes under cluster 1	14016ms	Comma separated value
K-Harmonic means	600 rows, 15 columns	Harmonic averages are taken	Instances 1,3,4,5,6 comes under cluster 5	11813ms	Comma separated value

VI. Conclusion

K-mean is a center based clustering algorithm. Centroid is taken randomly in k mean so it is sensitive to the initialization of cluster centers. Where as in k harmonic mean centers are taken based on the harmonic average .Performance in k mean depends on the initialization of the centers which is a major problem, but performance of KHM is a function of the harmonic averages of the distances from each data point. Execution time for k harmonic mean is less compared to k- mean algorithm. Outliers are removed efficiently in KHM than K-mean algorithm.

VII. Refernces

[1] Bin Zhang, Meichun Hsu, Umeshwar Dayal, "K-Harmonic mean data clustering algorithm" Software Technology Laboratory HP Laboratories Palo Alto, HPL-1999-24 October1999, pp1-26.

[2] Jiawei Han and Micheline Kamber, "Data Mining: Concepts and Techniques", 2nd edition.

[3] Arun .K. Pujari "Data mining techniques", Universities Press (India) Private Limited.

[4] Pang-ning tan, Micheal Steinbach, vipin kumar "Introduction to data mining techniques" Wesley publications.

[5] Cheng-Ru Lin Ming-Syan Chen, "Combining Partitional and Hierarchical Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging". IEEE Transactions on knowledge and data engineering, vol 17, no 2, february2005, pp. 145-159.

[6] Sudipto Guha , Rajeev Rastogi , Kyuseok Shim, "CURE: an efficient clustering algorithm for large databases, Proceedings of the 1998 ACM SIGMOD, June 01-04, 1998, p.73-84,

[7] K-mean clustering example -

[8] URL:http://www.faculty.uscupstate.edu/atzacheva/SHIM450/kmeanexample.doc

[9] URL:http://archive.ics.uci.edu/ml.

[10] URL:http://en.wikipedia.org/wiki/K-means_algorihtm.

[11] URL: http://en.wikipedia.org/wiki/K-medoids.