Hierarchical Representatives Clustering with Hybrid Approach

Byung-Joo An and **Eun-Ju Kim** and **Yill-Byung Lee**
Department of Computer Science,
College of Engineering,
Yonsei University, Korea
{joo,outframe,yblee}@csai.yonsei.ac.kr

Abstract

Clustering is a discovering process of meaningful information by grouping similar data into compact clusters. Most of traditional clustering methods are in favor of small datasets and have difficulties handling very large datasets. They are not adequate clustering methods for partitioning huge datasets in data mining perspective. We propose a new clustering technique, HRC(hierarchical representatives clustering), that can be applied to large datasets and find clusters with good quality. HRC is a two phase algorithm that take advantage of a hybrid approach that combine SOM and hierarchical clustering. Experimental results show that HRC can discover better clusters efficiently in comparison to traditional clustering methods.

Introduction

Data clustering is an important technique for exploratory data analysis. It has been used practically in real world application of data classification, image processing, and information retrieval. Clustering is a basic technique to find useful information hidden in databases by grouping data with similar characteristics (Zhang, Ramakrishnan, & Miron June 1996; 1997). It is a process to maximize within-clusters similarity and minimize between-clusters similarity (Duda & Hart 1973). Discovered clusters explain distributions of dataset and give a foundation for further analysis (Berry & Linoff 1997). So clustering is a starting point in data mining process. Clustering technique is applicable to group customers according to buying pattern, categorize web documents by subject, and extract interesting spatial pattern in GIS databases. In recent, there have been many studies of data mining or knowledge discovery in databases that is defined as the discovery of interesting, implicit, and previously unknown knowledge from large databases (Fayyad, Piatetsky-Shapiro, & Smyth 1996). As data mining emerges, many researchers are interested in efficient and effective clustering algorithm that can reduce the number of scanning raw data and computational complexity.

In this paper, we propose a new clustering technique, HRC(hierarchical representatives clustering), that can be efficiently applied to large datasets and find clusters with good quality. Most existing clustering algorithms can find only hyper-spherical clusters with similar size (Guha, Rastogi, & Shim June 1998; Jain, Murty, & Flynn Sep 1999). Some can find clusters with various shape and size, but are limited because of quadratic computational complexity. Our proposed method gets over these difficulties and finds good clusters from very large datasets. HRC uses a hybrid approach which combine SOM(Self-Organizing Map) and hierarchical clustering. It is implemented with two phases. The first phase is sub-clustering by SOM and the second phase is merging stage by hierarchical approach with novel similarity measure based on *cohesiveness* and *closeness*.

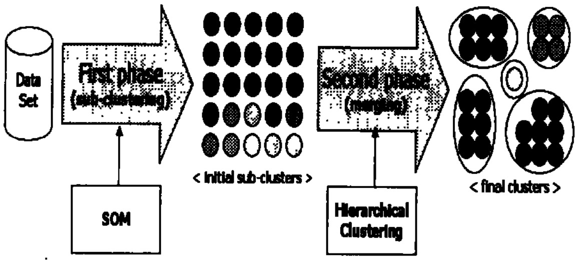

Figure 1: Overview of HRC

Related Works

Clustering has been studied actively in the statistics, database, and machine learning. There are two main categories of clustering algorithms. They are hierarchical clustering and partitional clustering (Duda & Hart 1973; Kaufman & Rousseeuw 1990).

Hierarchical clustering can be divided to divisive and agglomerative methods (Duda & Hart 1973; Murtagh 1983). Divisive hierarchical clustering starts with one big cluster and splits the farthest pair until individual objects to form clusters respectively. On the other hand, agglomerative hierarchical clustering starts with each data objects forming its own cluster and merges repeatedly the closest pair until all data objects gathered in one big cluster. Based on how to measure similarity between clusters, there are single linkage, complete linkage, average linkage, centroid linkage and ward's method (Duda & Hart 1973).

Partitional clustering optimizes objective function to partition datasets into k clusters. K-Means, K-Medoid, and PAM are belong to this category and K-Means is the most common algorithm because of its simplicity. In K-Means, we select random initial centroids and assign data objects to cluster whose centroid is nearest, then update cluster centroids. This process of assigning data objects to cluster and re-calculating centroids is continues until stopping condition is satisfied.

Some recent researches targeted on handling large datasets by using a summarized cluster representation or a special data structure. BIRCH(Balanced Iterative Reducing and Clustering using Hierarchies) (Zhang, Ramakrishnan, & Miron June 1996; 1997) defined a cluster feature that is a summarized cluster representation. A cluster feature consists of the number of data objects, the linear sum of data objects, and the square sum of data objects. This method utilizes a balanced tree of cluster features without dealing with all data objects. When a new data object is inserted to a cluster, the new cluster feature can be calculated from the previous cluster feature without requiring all data objects in the cluster. The incremental BIRCH algorithm is a fast clustering algorithm with limited resources, but can't find diverse shape of clusters because it is based on the centroid-based approach.

Chameleon (Karypis, Han, & Kumar August 1999) suggested a dynamic modeling of cluster similarity. Through two phases, chameleon constructs graph and partitions the graph, then merges these partitions. It yielded good results for finding highly variable clusters. Thus it is more applicable to spatial data mining.

Hierarchical Representatives Clustering

HRC is a hybrid method that can be applicable to very large datasets . Because it exploits representatives of cluster to reduce computational complexity, it is scalable and robust to outliers and noises.

HRC is a two phase algorithm that take advantage of a hybrid approach which combine SOM and hierarchical clustering. SOM has advantages in applying large datasets due to its on-line learning process. Hierarchical clustering is known to be better than SOM in point of clustering quality, although slower. HRC adopted good features of two methods, SOM's efficiency of processing large datasets and hierarchical clustering's cluster quality. In the first phase, HRC uses SOM to partition dataset into initial small sub-clusters that have two representatives. In the second phase, these sub-clusters found in the first phase are merged by agglomerative hierarchical manner. At this time, to measure between clusters similarity we use *cohesiveness* and *closeness*. Similarity based on *cohesiveness* and *closeness* can take account of contextual relation between clusters.

The first phase: sub-clustering

We use SOM to divide dataset into initial small sub-clusters. Neurons in the map form sub-clusters respectively. The two representatives of initial sub-cluster are determined as following. We compute eigenvector having the largest eigenvalue of data objects fallen on a sub-cluster and project data objects on the eigenvector. As in figure 3, two middle points on eigenvector are determined to represent sub-cluster.

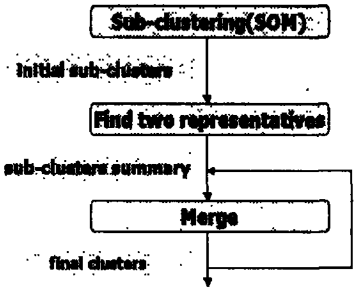

Figure 2: Flow of HRC algorithm

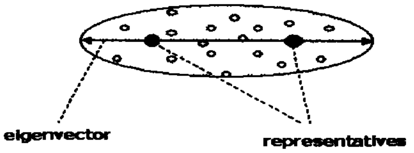

Figure 3: Two representatives of initial sub-cluster

After finding representatives of sub-clusters, in the second phase only these representatives are used to compare sub-clusters similarity. Therefore we get a data reduction effect, in addition to, reduce influence of outliers and noises.

The second phase: merging

In this phase, we merge repeatedly sub-clusters found in the first phase by way of agglomerative hierarchical approach and find final clusters. It is efficient to compare clusters similarity because it is required only sub-cluster representatives not raw data objects.

How to measure similarity of two clusters is critical to discovering good results (Karypis, Han, & Kumar August 1999). To find better clusters, we used a similarity measure based on *cohesiveness* and *closeness* that considers contextual relation between clusters. *cohesiveness* is measured by *cohesiveDistance* in equation (2) and *closeness* is by *closeDistance* in equation (3).

$$similarity_{ij} = \frac{1}{cohcsiveDistancc_{ij} \times closcDistance_{ij}{}^{2}} \qquad (1)$$

CohesiveDistance shows a degree of compactness for internal cluster and *closeDistance* shows a degree of external nearness for between-clusters.

$$cohesiveDistance_{ij} = \frac{link_{ij}}{\frac{link_{i}+link_{j}}{2}} \qquad (2)$$

clustering method	time complexity	space complexity
SOM	$O(nkl)$	$O(k+n)$
K-Means	$O(nkl)$	$O(k+n)$
hierarchical clustering	$O(n^2)$	$O(n^2)$
HRC	$O(nkl+m^2)$	$O(k+n+m^2)$

Table 1: Complexity of clustering methods

$$closeDistance_{ij} = \frac{\sum_a^{n_i}\sum_b^{n_j} w_{r_a}\times w_{r_b}\times ||r_a - r_b||^2}{n_i \times n_j} \tag{3}$$

We defined the degree of linking. It is the sum of weighted distances that are between pairs of representatives. The degree of linking between cluster i and j is the sum of weighted distances between representatives in cluster i and representatives cluster j. The degree of linking in cluster i is the sum of weighted distances which are between representatives in cluster i. In equation (3) (4) (5), r_a and r_b are representative vectors, w_{r_a} is the number of raw data objects that representative r_a represent, n_i is the number of representatives of sub-cluster i.

$$link_{ij} = \sum_a^{n_i}\sum_b^{n_j} w_{r_a}\times w_{r_b}\times ||r_a - r_b||^2 \tag{4}$$

$$link_i = \frac{\sum_a^{n_i}\sum_b^{n_i} w_{r_a}\times w_{r_b}\times ||r_a - r_b||^2}{2} \tag{5}$$

$cohesiveDistance_{ij}$ is represented by $link_{ij}$ normalized by average of $link_i$ and $link_j$. This measures compactness,if cluster i and cluster j were merged, whether the merged cluster is more compact than average compactness of cluster i and j. $closeDistance_{ij}$ is the average weighted distance between representatives in cluster i and representatives in cluster j.

Complexity Analysis

When the number of data objects is n, the number of cluster is k, the iteration frequency is l, and the number of sub-clusters is m, time and space complexity of SOM, K-MEANS, hierarchical clustering and HRC are in table 1.

Time and space complexity of SOM and K-Means are linearly proportional to the number of data objects, but hierarchical clustering is quadratic. HRC's complexity is linearly proportional to n, although it needs more computation than SOM and K-Means for the second phase. Because the number of sub-clusters, m, is very small compared to n, it is negligible. Thus HRC is more efficient than hierarchical clustering of quadratic complexity and have good clustering results with linear complexity.

Experiments

In experiments, we compared our method with K-Means, hierarchical clustering, and a simple hybrid method. The simple hybrid method is a mere combination of SOM and hierarchical clustering with single, complete and average linkage.

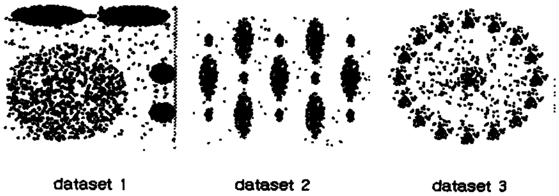

Figure 4: Two dimensional spatial datasets

We tested with three spatial datasets and seven UCI machine learning repository datasets(australian, diabetes, heart, iris, soybean, wine and zoo) (UCI 2000). To confirm clustering results visually, spatial datasets are expressed as points in two-dimensional euclidean space.

$$Q = \sum_{i=1}^{k} \frac{D_i}{k} \tag{6}$$

$$D_i = \frac{\sum_{a=1}^{n_i}\sum_{b=1}^{n_i} ||x_a - x_b||}{n_i(n_i - 1)} \tag{7}$$

Goodness of clustering results is evaluated by Q in equation(6). Equation(7) shows that D_i is the average pairwise distance within cluster i. Q is the average of D_is.

Experimental Results

Experimental results with three spatial datasets are shown and compared by clustering quality(Q) in table 2. In dataset 1, We have observed that K-Means can find two elliptical clusters with similar size on the upper side, but failed to one big circular cluster and two adjacent small clusters on the middle. The big circular cluster is divided and two small clusters are merged. Because K-Means algorithm implicitly assumes that clusters are hyper-ellipsoidal and of similar size, it can't find clusters varied in size as dataset 1. Hierarchical clustering shows different results according to how to measure cluster similarity. Single linkage and complete linkage merge most different clusters because they are very sensitive to noise and outlier. Average linkage, although it is less sensitive to noise, splits big cluster and merges small clusters due to its centroid-based nature. The simple hybrid method has similar results without regard to simple, complete, or average linkage in merging phase because it reduces influence of noise by sub-clustering. But it failed to find correct clusters. HRC finds two elliptical clusters that are connected by line and of similar size on the upper side, and separates one big circular cluster and two adjacent small clusters on the middle. Results with datasets 2 and dataset 3 are also similar to the result of dataset 1. In table 2, clustering quality of HRC is better than other methods for dataset 1, dataset 2, and dataset 3. From these experimental results we confirmed that HRC could find variable sized clusters with noises.

To verify with real world datasets, seven UCI machine learning repository datasets (australian, diabetes, heart, iris, soybean, wine, and zoo) are tested and evaluated by Q, the

	K-Means	hierarchical clustering		
		single	complete	average
dataset 1	5.609	8.063	6.012	5.673
dataset 2	6.217	9.700	8.042	5.854
dataset 3	2.708	2.495	2.620	2.577

	HRC	simple hybrid		
		single	complete	average
dataset 1	4.369	5.917	5.699	5.602
dataset 2	5.274	5.884	6.523	5.646
dataset 3	1.960	2.613	2.517	2.616

Table 2: Comparison of clustering quality(Q) with spatial datasets

	K-Means	hierarchical clustering		
		single	complete	average
australian	1.322	1.538	1.334	1.538
diabetes	0.540	0.604	0.625	0.604
heart	1.475	1.673	1.624	1.378
iris	0.918	0.844	0.890	0.917
soybean	3.399	3.095	3.095	3.095
wine	0.715	0.988	0.719	0.988
zoo	1.185	1.354	1.361	1.306

	HRC	simple hybrid		
		single	complete	average
australian	1.535	1.048	1.322	1.322
diabetes	0.521	0.534	0.534	0.534
heart	1.284	1.358	1.488	1.487
iris	0.786	0.938	0.956	0.956
soybean	2.933	3.127	3.081	3.095
wine	0.663	0.839	0.732	0.707
zoo	1.090	1.223	1.317	1.380

Table 3: Comparison of clustering quality(Q) with UCI datasets

measure of clustering quality, in table 3. In table 3, we have obtained the fact that HRC is superior to other methods with six datasets except for one, australian dataset.

In the experimental results with spatial datasets and real world datasets, we confirm that HRC can find good clusters and partition efficiently large datasets.

Conclusions

In this paper we presented HRC, the new clustering algorithm, which can be applicable to very large datasets. Because it exploits representatives of cluster to reduce computational complexity, it is scalable and robust to outliers and noises. HRC is a two phases algorithm that take advantage of a hybrid approach that combine SOM and hierarchical clustering. HRC adopted good features of two methods, SOM's efficiency of processing large datasets and hierarchical clustering's cluster quality. In addition, to measure between clusters similarity we use a similarity, which can take account of contextual relation between clusters, based on *cohesiveness* and *closeness*.

By experimental results we confirm that HRC can discover better clusters efficiently in comparison to traditional clustering methods and the simple hybrid method.

References

Berry, M. J. A., and Linoff, G. 1997. *Data Mining Techniques for Marketing, Sales, and Customer Support.* Jone Wiley & Sons.

Duda, R. O., and Hart, P. E. 1973. *Pattern Classification and Scene Analysis.* A Wiley-Interscience Publication, New York.

Fayyad; Piatetsky-Shapiro; and Smyth. 1996. *Advances in knowledge discovery and data mining.* AAAI Press/The MIT Press.

Guha, S.; Rastogi, R.; and Shim, K. June 1998. Cure: An efficient clustering algorithm for large databases. *the ACM SIGMOD Conference on Management of Data, Seattle, Washington.*

Jain, A. K.; Murty, M. N.; and Flynn, P. J. Sep. 1999. Data clustering: a review. *the ACM Comput. Surv.* 31:264 – 323.

Karypis, G.; Han, E.-H.; and Kumar, V. August 1999. Chameleon: Hierarchical clustering using dynamic modeling. *IEEE Computer* 32.

Kaufman, L., and Rousseuw, P. J. 1990. *Finding Groups in Data - An Introduction to Cluster Analysis.* Wiley Series in Probability and Mathematical Statistics.

Murtagh, F. 1983. A survey of rescent advances in hierarchical clustering algorithms. *The Computer Journal.*

UCI. 2000. Maching Learning Repository, http://www.ics,uci.edu/ mlearn.

Zhang, T.; Ramakrishnan, R.; and Miron. 1997. Birch: A new data clustering algorithm and its applications. *Data Minning and Knowledge Discovery* 1:141–182.

Zhang, T.; Ramakrishnan, R.; and Miron. June 1996. Birch: An efficient data clustering method for very large databases. *the ACM SIGMOD Conference on Management of Data, Montreal, Canada.*