www.ijecs.in
International Journal Of Engineering And Computer Science ISSN: 2319-7242
Volume 4 Issue 8 Aug 2015, Page No. 14027-14032

Potential based similarity metrics for implementing hierarchical clustering

M.K.V.Anvesh[#1], Dr. B. Prajna[#2]
#1 Dept. of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, mkvanvesh@gmail.com
#2 Associate Professor, Dept. of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, prajna.mail@gmail.com

Abstract: - The main aim of the data mining process is to extract information from a large data set and transform it into an understandable form for further use. Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters). A prominent clustering is hierarchical clustering. Hierarchical clustering is a common method used to determine clusters of similar data points in multidimensional spaces. When performing hierarchical clustering, some metric must be used to determine the similarity between pairs of clusters. Traditional similarity metrics either can only deal with simple shapes or are very sensitive to outliers. Potential - based similarity metrics, Average potential energy similarity metric and Average maximal potential energy similarity metric have special features like strong anti-jamming capability and they are capable of finding clusters of complex irregular shapes.

Keywords: ***hierarchical clustering, traditional similarity metrics, potential based similarity metrics.***

I. Introduction

A cluster is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. a good clustering algorithm is able to identity clusters irrespective of their shapes. Other requirements of clustering algorithms are scalability, ability to deal with noisy data, insensitivity to the order of input records, etc. Clustering is very important in pattern recognition, machine learning and data mining.

In clustering, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. It is connectivity based clustering. This method starts with a set of distinct points, each of which is considered a separate cluster. The two clusters that are closest according to some metric are agglomerated. This is repeated until all of the points belong to one hierarchically constructed cluster. The final hierarchical cluster structure is called a **dendogram** (see fig.1) which is simply a tree that shows which clusters were agglomerated at each step.

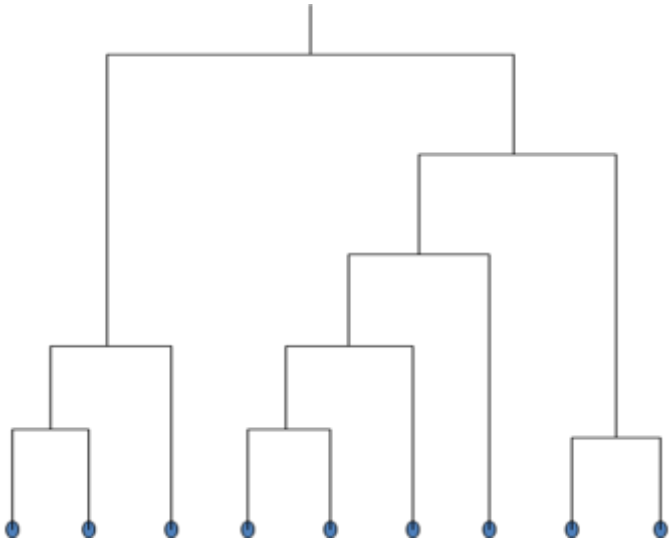

Figure 1.

A clustering of the data objects is obtained by cutting the dendogram at the desired level, then each connected component forms a cluster. Hierarchical clustering algorithm is a prominent one. The advantages with this algorithm

1. Do not have to assume any particular number of clusters (Any number of clusters can be obtained by 'cutting' the dendogram at the proper level),
2. Embedded flexibility regarding the level of granularity.

3. Ease of handling any forms of similarity or distance.
4. Applicability to any attributes type.

Strategies for hierarchical clustering generally fall in two types.

1. Agglomerative and 2. Divisive

Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In order to decide which clusters should be combined (for agglomerative), or where a cluster should be split (for divisive), a measure of dissimilarity between sets of observations is required. In most methods of hierarchical clustering, this is achieved by use of an appropriate metric (a measure of distance between pairs of observations). The state-of-the-art metrics include single link, (see ref.2) complete link, average link, centroid, median, and minimum variance. Ref.8 gives more information about these traditional metrics.

Each of the above-mentioned traditional metrics has its own well-known disadvantages. The main shortcoming of the single link metric is the 'chaining' effect. (see ref.2 and 9) And it is sensitive to noise and outliers. The Limitations of complete linkage metric is tends to break large clusters and biased towards globular clusters. Remainder metrics are only capable of finding clusters of spherical shapes.

Many papers and documentations in the literature focus on improving algorithms of hierarchical clustering. These include reducing the computational complexity and memory requirement, (see ref. 2, 3 and 9) proposing parallel or on-line algorithms, (see ref. 1 and 8) and presenting efficient clustering algorithms for large databases. (see ref.4)

The main contribution of this paper is reducing the time and space complexities by using potential based similarity metrics. These metrics have strong anti-jamming capability and can find clusters of arbitrary irregular shapes.

The most important step in hierarchical clustering is to find a pair of clusters with the highest degree of similarity and to agglomerate them as a single cluster. Different clustering results can be acquired using different similarity metrics. Many clustering methods use distance measures to determine the similarity or dissimilarity between any pair of objects. It is useful to denote the distance between two instances x_i and x_j as: d (x_i, x_j). A valid distance measure should be symmetric and obtains its minimum value (usually zero) in case of identical vectors. The different distance measures are

1. Euclidean distance
2. Manhattan distance
3. Mahalanobis distance
4. Cosine similarity

Euclidean distance is commonly used.

The rest of the paper is as follows: Section 2 gives the definition of metric functions. Section 3 describes existing system (different traditional similarity metrics). Section 4 devoted to the analysis of potential-based similarity metrics. Experimental results are shown in section 5. Section 6 is the conclusion and 7 is future work.

II. (a) Similarity metric function

Let Π be the set of all *d*-dimensional finite vector sets, and let Π_s be the set of all d-dimensional vector singletons:

$\Pi = \{\{P_1, P_2, \ldots P_k\} : 1 \le k < \infty, p \in R^d \forall\ 1 \le i \le k\}$

$\Pi_s = \{\{P\} : P \in R^d\}$

Apparently the following formula holds $\Pi_s \subset \Pi$

Function ψ defined as follows is called a similarity metric function on Π.

$$(a).\ \psi : \Pi \times \Pi \mapsto R^+ \cup \{0\}$$

$$(b).\ \psi(\Xi_1, \Xi_2) = \psi(\Xi_2, \Xi_1), \quad \forall \Xi_1, \Xi_2 \in \Pi$$

(b) Distance metric function

δ can be defined similarly except that

$$(c).\ d(P_1, P_2) > d(P_3, P_4) \Leftrightarrow \delta(\{P_1\}, \{P_2\}) > \delta(\{P_3\}, \{P_4\})$$

Function ψ gives some kinds of similarity between any pair of elements in Π, and function δ gives some kinds of distance between them.

III. Existing system

(Classical distance metrics between vector sets.**)**

DOI: 10.18535/ijecs/v4i8.71

Let $\Xi 1$, $\Xi 2$ are two d-dimensional vector sets with m, n elements respectively

$$\Xi_1 = \{P_{11}, P_{12}, \ldots, P_{1m}\},\ m>0$$

$$\Xi_2 = \{P_{21}, P_{22}, \ldots, P_{2n}\},\ n>0$$

So $\Xi_1 \in \Pi$, and $\Xi_2 \in \Pi$.

The classical distance metric functions can be defined formally as follows.
The **single link** distance between two clusters is given by the minimum distance between points in the two clusters, which is formally defined as follows,.

$$\delta_{sl}(\Xi_1, \Xi_2) = \min\{d(P_{1i}, P_{2j})\},\ 1 \le i \le m, 1 \le j \le n$$

The **complete link** distance between two clusters is given by maximum distance between points in the two clusters, which is formally defined as follows.

$$\delta_{cl}(\Xi_1, \Xi_2) = \max\{d(P_{1i}, P_{2j})\},\quad 1 \le i \le m, 1 \le j \le n$$

The **average link** distance between two clusters is given by mean distance between points in the two clusters, which is formally defined as follows.

$$\delta_{al}(\Xi_1, \Xi_2) = \text{avg}\{d(P_{1i}, P_{2j})\},\quad 1 \le i \le m, 1 \le j \le n$$

Here, d is the distance between the points, and the Euclidean distance is commonly used. Other traditional distance metrics can be defined similarly.

IV. Proposed system

(Potential-based similarity metrics between vector sets)

(a). Electrical potential, defined as the work done in carrying a unit positive charge from infinity to that point.
Electric potential can be depicted intuitively by means of isopotential contours (see Figure 2).

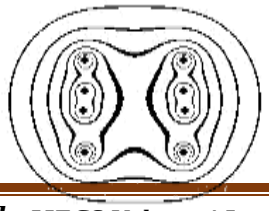

Fig. 2. Isopotential contours of two sets of charges

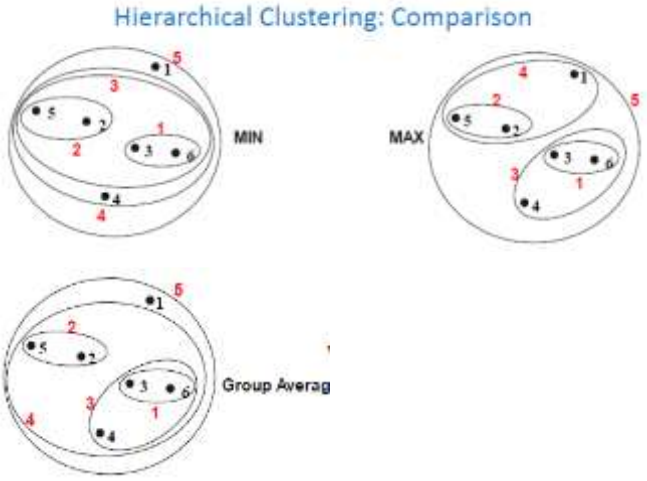

Fig.3 Hierarchical clusters for different metrics. Where MIN is single linkage, MAX is complete linkage and group average is average linkage.

From the two figures 2 and 3, we can immediately see the wonderful similarity and correlation between Isopotential contours and hierarchical clustering. Using this, we hope to find out a better hierarchical similarity metrics that can overcome the defects of classical similarity metrics. To achieve this, we will use two potential-based similarity metrics: Average potential energy similarity metric and Average maximal potential energy similarity metric.
The electric potential due to a system of point charges is equal to the sum of the point charges' individual potentials. This fact simplifies calculations significantly, since addition of potential (scalar) fields is much easier than addition of the electric (vector) fields.

(b). Potential based similarity metrics
(i). Average potential energy similarity Metric.
The average potential energy similarity between two clusters is given by the average of all of the potential energies between points in the two clusters:

DOI: 10.18535/ijecs/v4i8.71

(Exponential function)

$$\psi_{apes}(\Xi_1, \Xi_2) = \frac{\sum_{1\le i\le m, 1\le j\le n} V(P_{1i}, P_{2j}) + \sum_{1\le i\le m, 1\le j\le n} V(P_{2j}, P_{1i})}{2mn}$$

$$= \frac{\sum_{1\le i\le m, 1\le j\le n} V(P_{1i}, P_{2j})}{mn}$$

Here, $V(P_{1i}, P_{2j})$ is the potential energy between P_{1i} and P_{2j}.

(ii). Average maximal potential energy similarity Metric. The average maximal potential energy similarity between two clusters Ξ_1 and Ξ_2 defined as follows:

$$\psi_{amapes}(\Xi_1, \Xi_2) = \frac{1}{2}\frac{\sum_{1\le i\le m} \max_{1\le j\le n} V(P_{2j}, P_{1i})}{m} + \frac{1}{2}\frac{\sum_{1\le j\le n} \max_{1\le i\le m} V(P_{1i}, P_{2j})}{n}$$

Now we discuss the potential function V (p1, p2) Let P1 and P2 are unit charges, then potential function V(p1,p2) can be defined as follows,

$$V(p1, p2) = \frac{K}{d(P_1, P_2)}$$

(Inverse ratio potential function.)

Potential function is not necessarily inverse ratio form from the mathematic point of view. Any decrease function with respect to d (P_1, P_2) is potential function in

$$V(p1, p2) = \frac{K}{d^2(P_1, P_2)}$$

(inverse square function)

$$V(p1, p2) = K \times \exp\left(-\frac{d^2(P_1, P_2)}{2\sigma^2}\right)$$

(Gauss function)

$$V(p1, p2) = K \times \exp\left(-\frac{d(P_1, P_2)}{\sigma}\right)$$

(C) Algorithm analysis

Given a set of N items to be clustered, and an N*N distance (or similarity) matrix, the basic process of hierarchical clustering is

1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.
4. Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)

In this paper Potential-based similarity metrics bear an analogy with the classical distance-based metrics. Existing sequential, parallel, distributed and online algorithms [2, 3, 9, 1, 8] can be used in the new metrics with little modification.

We briefly analyze the time and space complexity of potential-based hierarchical clustering algorithms now.

The reducibility property [9, 8] for a distance metric requires that when we agglomerate clusters i and j, the new cluster i + j cannot be closer to any cluster than both clusters i and j were. Formally, if the following distance constraints hold:

$$\delta(i, j) < d;\ \delta(i, k) > d;\ \delta(j, k) > d$$

then we must have for the agglomerated cluster i + j:

$$\delta(i + j, k) > d.$$

Similarly, the reducibility property for a similarity metric can be defined:

$$\psi(i, j) > s;\ \psi(i, k) < s;\ \psi(j, k) < s \Rightarrow \psi(i + j, k) < s$$

For distance or similarity metrics that satisfy the reducibility property we can perform clustering in $o(n^2)$ time by computing nearest neighbor chains. Ref. [8] (or [9]) gives the relevant algorithm that works by following a nearest neighbor chain until a pair of mutual nearest neighbors are found and then agglomerating them.

Average potential energy similarity metric and Average maximal potential energy similarity

DOI: 10.18535/ijecs/v4i8.71

metric both satisfy the reducibility property. So algorithms based on them can be performed in $O(n^2)$ time. The space requirement of them is $O(n^2)$ by keeping an array of the inter-cluster distances.

V. Experimental results

In this section the performance of the potential based similarity metrics is tested and evaluated using some test data sets. The data sets used in these experiments are synthetic data. We first compare different metrics (include traditional and potential based similarity metrics) with respect to the quality of clustering, and then compare the clustering quality of different potential functions. Finally, we focus on analyzing the sensitivity of some potential functions to their parameters.

We experimented with synthetic data set of size 600×20 containing numerical points.

Data sets with different shapes and outliers have different difficulty level for clustering.

The data set is large data set containing 600 rows and 20 columns of numerical data. We apply the above procedure (hierarchical cluster analysis) over this data set. Table1 show us the clustering quality of the different similarity metrics over this dataset.

Quality of clustering

We use data set to compare different metrics with respect to the quality of clustering. Metrics that participate in the comparisons are: SLD, CLD, ALD, APES and AMAPES. Where SLD is Single Linkage, CLD is Complete Linkage, ALD is Average Linkage, APES is Average Potential Energy Similarity metric and AMAPES is Average Maximum Potential Energy Similarity metric.

Table1: Comparision of different metrics with respect to clustering quality. Assume σ = 10000

		SLD	CLD	ALD	APES	AMAPES
Data set	Correct rate	10%	10%	10%	95%	100%
	No. of steps required	599	599	598	18	14

Because of the 'chaining' effect, SLD merges the neighboring clusters with the chain outliers linked, while splitting the big circle. SLD can give the clustering result of the data set which only contains randomly scattered outliers.

ALD and CLD have strong anti-jamming capacity, but can only deal with spherical or near-spherical shapes and similar sizes. They can't deal with the complex shapes.

APES and AMAPES can deal with any type data sets in spite of the outliers and the various sizes.

The data set which doesn't contain chain outliers and which can be dealt with SLD.As SLD only works for the datasets that doesn't contain any outliers. Only APES and AMAPES can recognize this complex shapes correctly.

Comparison of different potential functions

Here we consider four potential functions: Inverse Ratio, Inverse Square, Exponential and Gauss. Then we compare the clustering quality of this different potential functions. We use the APES and the AMAPES metrics separately for our study and parameters of these potential functions are adjusted for optimum clustering. Where nos is number of steps required shown in Table 2.

Table2: comparision of different potential functions. Assume σ = 40000.

	Inverse ratio	Inverse square	Exponential	Gauss
APES(number of steps)	492	99	226	5
AMAPES (number of steps)	517	110	241	4

We can learn from table 2 that dramatically different results can be acquired using different potential functions. Thus, choosing an appropriate potential function is very important to clustering. The Gauss function and the inverse square function behave better than the other two potential functions. In our experiments, we also find that

the parameter of the exponential function is more difficult to adjust than that of the Gauss. Thus, the Gauss potential function is recommended.

In this section, we perform a sensitivity analysis for some potential functions with respect to their parameters.

VI. Conclusion

A total of five similarity metrics have been evaluated in this paper. The Potential based similarity metrics were found to be more robust to outliers than the other metrics. The study also found that the gauss function performed better than the other potential functions. Although the results from analyzing the synthetic data set with the potential based similarity metrics shows that it is preferable for large data sets.

From the experimental results and complexity analysis the following points can be concluded.

1. The potential based similarity metrics can be used efficiently for hierarchical clustering.
2. APES and AMAPES both satisfy the reducibility property. So algorithms based on them can be performed in $O(n^2)$ time.
3. When σ is small, APES is very fragile in the presence of chain outliers, behaving similar to SLD. As the value of σ grows, its anti-jamming capability becomes stronger and stronger, and the quality of clustering better and better.
4. For large data sets The Gauss potential function is considered to be more efficient.

VII. Future work

An important direction for further study is how to adjust parameters of potential functions automatically for optimal results.

VIII. References

[1] Yasser El-Sonbaty, M.A. Ismail. On-line hierarchical clustering. *Pattern Recognition Letters* 19 (1998) 1285-1291

[2] R. Sibson. SLINK: An optimally efficient algorithm for the single link cluster method, *Computer Journal*, 16:30-34, 1973

.

[3]M. Dash and H. Liu. Efficient Hierarchical Clustering Algorithms using Partially Overlapping Partitions.

[4]S. Guha, R. Rastogi, and K. Shim. Cure: An efficient clustering algorithm for large databases. *Information Systems* Volume: 26, Issue: 1, March, 2001, pp. 35-58

[5]min tang li, fang zhao, yifei wu, albert gan: evaluation of agglomerative hierarchical clustering methods . Presentation and Publication at the 82nd Annual Meeting of the Transportation Research Board Washington, D.C.

[6]M.S.Yang," A Survey of hierarchical clustering" Mathl. Comput. Modelling Vol. 18, No. 11, pp. 1-16, 1993.

[7] George Karypis, Eui-Hong (Sam) Han, and Vipin Kumar: multilevel refinement for hierarchical clustering

[8]C. F. Olson. Parallel algorithms for hierarchical clustering. *Parallel Computing*. 21:1313-1325, 1995.

[9]F. Murtagh. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26:354-359, 1983

[10]M. dash and h.liu :efficient hierarchical clustering algorithms using partially overlapping partitions. *The 5th Pacific Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2001*, HongKong

[11]D. Fasulo, "An analysis of recent work on clustering algorithms," Department of Computer Science and Engineering,
University of Washington, Tech. Rep. # 01-03-02, 1999. [Online]. Available:
citeseer.nj.nec.com/fasulo99analysi.html