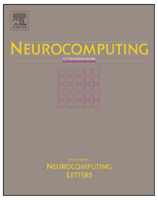
Neurocomputing 173 (2016) 3–8

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



A novel travel-time based similarity measure for hierarchical clustering

Yonggang Lu [a,*], Xiaoli Hou [a], Xurong Chen [b,c]

[a] School of Information Science and Engineering, Lanzhou University, Lanzhou, Gansu 730000, China
[b] Institute of Modern Physics, Chinese Academy of Sciences, Lanzhou, Gansu 730000, China
[c] Institute of Modern Physics of CAS and Lanzhou University, Lanzhou, Gansu 730000, China

ARTICLE INFO

Article history:
Received 2 June 2014
Received in revised form
31 December 2014
Accepted 28 January 2015
Available online 4 August 2015

Keywords:
Clustering
Similarity measure
Travel time

ABSTRACT

The similarity measure plays an important role in agglomerative hierarchical clustering. Following the idea of gravitational clustering which treats all the data points as mass points under a hypothetical gravitational force field, we propose a novel similarity measure for hierarchical clustering. The similarity measure is based on the estimated travel time between data points under the gravitational force field: the shorter the travel time from one point to another, the larger the similarity between the two data points. To simplify the computation, the travel time between a pair of data points is estimated using the potential field produced by all the data points. Based on the new similarity measure, we also propose a new hierarchical clustering method called Travel-Time based Hierarchical Clustering (TTHC). In the TTHC method, an edge-weighted tree of all the data points is first built using the travel-time based similarity measure, and then the clustering results are derived from the edge-weighted tree directly. To evaluate the proposed TTHC method, it is compared with four other hierarchical clustering methods on six real datasets and two synthetic dataset families composed of 200 datasets. The experiments show that using the travel-time based similarity measure can improve both the robustness and the quality of hierarchical clustering.

1. Introduction

As an important unsupervised learning method, clustering can be used to explore data structures of large and complex data. It has been applied to pattern recognition, data mining and image processing [1–3]. The clustering methods are usually divided into two different groups: partitional and hierarchical. Partitional clustering method produces a single partition, while hierarchical clustering method produces a result called dendrogram from which different and consistent partitions can be derived at different levels of abstraction [1–3]. Different partitions produced from a dendrogram are consistent because they form a totally ordered set under the refinement relation. So the hierarchical clustering can be used to analyze the structure of the data at different levels. Actually, it has been widely used in a lot of applications, such as document clustering [4], the analysis of gene expression data, regulatory networks and protein interaction networks [5–7]. There are two different hierarchical clustering approaches: agglomerative and divisive. The agglomerative method follows a bottom-up approach: initially each data point is in a different cluster, and then the two most similar clusters are merged at each step until a single cluster is produced. The divisive method follows a top-down approach: initially all the data points are in a single cluster and the selected cluster at each step is divided into two clusters until no more division can be made. Four traditional agglomerative methods are Single Linkage, Complete Linkage, Average Linkage and Ward's method [1,8]. Although a lot of progresses have been made recently in hierarchical clustering, challenges remain on how to improve the efficiency and the quality of the method to address many important problems [8].

Gravitational clustering [9] is an interesting and effective method which performs clustering by simulating a natural process: movements under gravitation. The data points in the feature space are all treated as mass points which can move following the Newton's Law of gravitation. The data points which move close enough to each other are grouped into a same cluster. This way, the clusters can be found naturally without specifying the number of clusters. Although the idea is proposed a long time ago [9], it has attracted lots of attentions recently [10–14]. The gravitational clustering is shown to be more adaptive and robust than other methods when dealing with arbitrarily-shaped clusters and clusters containing noise data [10,12,13]. Because it is very difficult to simulate the movement of mass points using molecular dynamics,

[*] Corresponding author.
E-mail addresses: ylu@lzu.edu.cn (Y. Lu), houxl12@lzu.edu.cn (X. Hou), xchen@impcas.ac.cn (X. Chen).

http://dx.doi.org/10.1016/j.neucom.2015.01.090
0925-2312/© 2015 Elsevier B.V. All rights reserved.

many approximations have to be made [10,13,14]. To avoid the complexity in the simulation, potential-based methods have also been proposed [14–16]. In the potential-based methods, the clustering results can be derived from the computed gravitational potential field without simulating the data point movements. We have proposed a novel potential-based hierarchical clustering method called PHA in one of our previous papers [16]. In the PHA method, the computed potential field and the distance matrix are used to produce an edge-weighted tree of the data points, from which the clustering results are produced efficiently. It is shown that the PHA method usually runs much faster and can produce more satisfying results compared to other hierarchical clustering methods [16]. In this work, the travel time instead of the distance between a pair of data points is used to build the edge-weighted tree and to derive the final clustering results. Travel time is a better choice than distance in computing the similarity in the potential-based method, because clusters are formed by the data point movements in the gravitational clustering [6], different levels of the clustering are mainly determined by different travel time needed for the data points to meet each other. The experiments also show the superiority of the travel-time based similarity measure over the distance-based similarity measure. We have reported the initial results of the method in a conference paper [17]. In this paper, more experimental results on high dimensional data as well as the analysis of the travel-time based similarity measure are included.

The rest of the paper is organized as follows. In Section 2, we introduce a simple physics model for estimating the travel time. In Section 3, we introduce the modified PHA clustering method. In Section 4, experimental results are shown. Finally, we conclude the paper in Section 5.

2. Estimation of the travel time

The travel time between two data points is defined as the time needed for a hypothetical mass point to travel from one point to another under the potential field. The potential field produced by all the data points is computed similarly as in [16]. The total potential at point i is

$$\Phi_i = \sum_{j=1..N} \Phi_{i,j}(r_{i,j}) \tag{1}$$

where $\Phi_{i,j}$ is the potential between points i and j, which is given by

$$\Phi_{i,j}(r_{i,j}) = \begin{cases} -\frac{1}{r_{i,j}} & \text{if} \quad r_{i,j} \geq \delta \\ -\frac{1}{\delta} & \text{if} \quad r_{i,j} < \delta \end{cases} \tag{2}$$

where $r_{i,j}$ is the Euclidean squared distance between points i and j, and δ is a distance parameter used to avoid singularity when the distance approaches zero. The parameter δ is determined by

$$\delta = \frac{\text{mean}\left(\min\limits_{j=1..N,\ r_{i,j}\neq 0}(r_{i,j})\right)}{C} \tag{3}$$

where C is a scale parameter.

After the potential field is computed, two approximations are used to simplify the estimation of the travel time: (a) when computing the travel time of the mass point between two data points, the path of the movement is assumed to be on a straight line; (b) the gradient of the potential field along the straight line is assumed to be constant, so that the acceleration is constant along the path.

Using the assumptions and Newton's Law of movement, the attractive force on the mass point is

$$F_{i,j} = \frac{|\Phi_i - \Phi_j|}{r_{i,j}} \tag{4}$$

and the acceleration of the mass point is

$$a_{i,j} = \frac{F_{i,j}}{m} = \frac{|\Phi_i - \Phi_j|}{m r_{i,j}} \tag{5}$$

where m is the mass of the mass point. Thus, the travel time of the mass point between point i and point j is

$$t_{i,j} = \sqrt{\frac{2r_{i,j}}{a_{i,j}}} = \sqrt{\frac{2mr_{i,j}^2}{|\Phi_i - \Phi_j|}} \propto \frac{r_{i,j}}{\sqrt{|\Phi_i - \Phi_j|}} \tag{6}$$

Based on the travel time given above, the similarity between points i and j is defined as

$$S_{i,j} = \begin{cases} 1 + \frac{|\Phi_i - \Phi_j|}{r_{i,j}^2} & \text{if} \quad r_{i,j} \geq \delta \\ 1 + \frac{|\Phi_i - \Phi_j|}{\delta^2} & \text{if} \quad r_{i,j} < \delta \end{cases} \tag{7}$$

If the distance between two data points is larger than δ, the similarity value given by (7) is one plus the part proportional to the inverse of the travel time squared; otherwise, δ is used as the distance in the computation, which is consistent with the computation of the potential field.

3. The TTHC clustering method

Given the similarity between two data points defined by (7), we can define the similarity between two clusters. First, an edge-weighted tree is constructed using the following two definitions:

Definition 1. For a data point i, another data point which is most similar to i within the data points having potential values lower than or equal to that of i is called the parent point of i, which is represented as

$$p(i) = \arg\max_k \left(S_{i,k} \,|\, \Phi_k \leq \Phi_i \quad \text{AND} \quad k \neq i\right) \tag{8}$$

Definition 2. For an edge E_i connecting points i and $p(i)$, the weight of the edge is defined as

$$\omega(E_i) = S_{i,p(i)} \tag{9}$$

It can be seen from Definition 1 that, except the root point which has the lowest potential value, each of the other points has exactly one parent point. Definition 2 gives the weight for every edge connecting a point and its parent point. This way an edge-weighted tree T can be built using all the data points as the tree nodes. Based on the edge-weighted tree T, a new similarity metric is defined as follows:

Definition 3. The similarity between cluster C_1 and cluster C_2 is

$$S(C_1, C_2) = \begin{cases} S_{i,j} & \begin{array}{l} if(\exists i \in C_1 \quad \text{AND} \quad \exists j \in C_2) \\ \text{AND} \\ (p(i) = j \quad \text{OR} \quad p(j) = i) \end{array} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $p(i)$ is the parent node of point i in the edge-weighted tree T.

It can be seen from Definition 3 that the similarity between two clusters is not zero only if there exists a tree edge connecting the two clusters. It has been shown that each cluster produced this way is a subtree of the edge-weighted tree T [16]. So there is at most one edge connecting any two clusters. This proves that the similarity metric given by Definition 3 is well-defined.