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/

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

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

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.

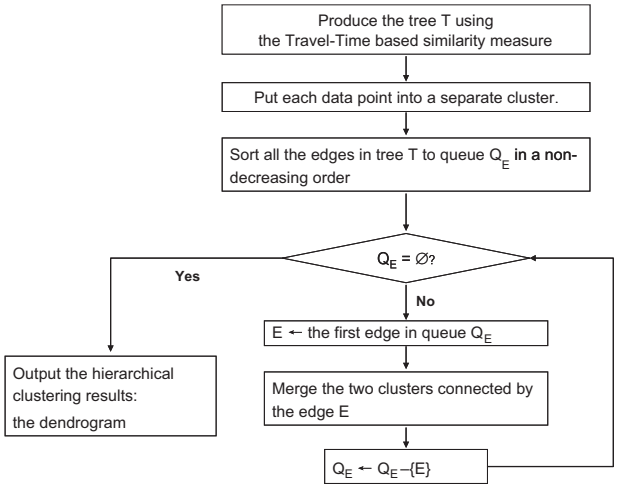


Fig. 1. The flowchart of the TTHC algorithm.

If there is a tree edge connecting two clusters, the similarity between the two clusters is the weight of the tree edge. So the two most similar clusters before each merging step can be identified easily from the sorted list of the tree edges. The flowchart of the proposed algorithm, called TTHC, is shown in Fig. 1. The proposed TTHC algorithm is easy to implement, and it can be shown that the time complexity of the algorithm is $O(n^2)$.

4. Experimental results

Matlab is used to implement all the codes in the experiments which are carried on a desktop computer with an Intel 3.06 GHz Dual-Core CPU and 3 GB of RAM. The proposed TTHC method is compared with the PHA method [16] and three traditional methods: Single Linkage, Complete Linkage and Ward's method. When computing the parameter δ using (3), the best values of the scale parameter C found in the experiments are used for each method, which are $C=10$ for PHA and $C=1$ for TTHC.

To evaluate the hierarchical clustering result, the dendrogram is cut horizontally to produce the same number of the clusters as in the benchmark. Then Fowlkes–Mallows index (FM-Index) [18] is used to compare the produced clusters with the benchmark. The range of the FM-Index is between 0 and 1. A larger FM-Index usually indicates a better match between the clustering result and the benchmark.

4.1. Experiments with two synthetic dataset families

Two 2D synthetic data, *Dataset Family A* and *Dataset Family B*, are used in our experiments. Each dataset family contains 100 randomly produced datasets of the same type. Each of the dataset in *Dataset Family A* has 400 data points from 2 bivariate normal distributions with parameters $\sigma_x=1$, $\sigma_y=5$ and $\mathrm{cov}_{xy}=0$ centered at $\mu_1=(0, 0)$ and $\mu_2=(5, 0)$ respectively. Each of the dataset in *Dataset Family B* has 800 data points from 4 normal distributions of different sizes, which are produced by the following parameters: $\sigma_1=2$, $\mu_1=(0, 0)$, $\sigma_2=3$, $\mu_2=(6, 13)$, $\sigma_3=4$, $\mu_3=(12, 0)$, $\sigma_4=2$ and $\mu_4=(16, 11)$. For the synthetic data, each normal distribution is considered as a benchmark cluster when computing the FM-Index.

For each dataset family, there are 100 different datasets. The maximum and the average FM-Index of the 100 datasets and the total running time in seconds of all the datasets are recorded. Table 1 shows the results for *Dataset Family A*. The maximum FM-Index produced by TTHC and PHA is 1.000 which is a perfect result, while the maximum FM-Index produced by the other three methods is only 0.7045. TTHC has also produced the best average FM-Index which is 0.8335. Table 2 shows the results for *Dataset Family B*. TTHC has produced the best average FM-Index and the best maximum FM-Index for the dataset family too. It can also be noted that Ward's method, PHA and TTHC have produced better results than the other two methods. For both dataset families, PHA and TTHC run faster than the other three methods. Compared to PHA, TTHC has produced higher average FM-Indices and very similar maximum FM-Indices. This shows that TTHC is more robust than PHA for the synthetic datasets.

Table 1
Experimental results for *Dataset Family A*.

Method	Time	FM-Index	
		Max	Avg
Single	10.43	0.7045	0.7036
Complete	9.917	0.6975	0.5627
Ward's	11.03	0.6308	0.5413
PHA	2.329	1.0000	0.8126
TTHC	2.797	1.0000	0.8335

Table 2
Experimental results for *Dataset Family B*.

Method	Time	FM-Index	
		Max	Avg
Single	153.4	0.5852	0.5820
Complete	154.2	0.8857	0.6398
Ward's	154.4	0.9340	0.8433
PHA	8.374	0.9347	0.8855
TTHC	9.664	0.9348	0.8947

The results of two datasets randomly selected from two dataset families are also shown in Figs. 2 and 3. The trees shown in Figs. 2 and 3 represent the details of the clustering results, which correspond to the structure of the data at different levels. It can be seen that the tree structures produced by TTHC is different from these produced by PHA. Compared to the results of PHA, the directions of the tree edges produced by TTHC are oriented more towards the cluster centers in both cases.

To further explore the differences between the travel-time based similarity measure given by (7) and the distance based similarity measure which is 1/(1+distance), the similarities given by the two measures are compared. For a dataset with N data points, because the distributions of N^2 similarity values given by the two measures are very different, the similarity values are first converted to N^2 ranks, and then the rank differences are used to compare the two similarity measures. In our experiments, the ranks of the similarities between each data point and the corresponding cluster centroid in each dataset are compared. The comparison result for the first dataset from *Dataset Family A* is shown in Fig. 4, and the comparison result for the first dataset from *Dataset Family B* is shown in Fig. 5. From Figs. 4(a) and 5(a), it can be seen that the travel-time based similarity values have a higher rank than the distance based ones except the data points very close to the centroids. From Figs. 4(b) and 5(b), it can be seen that for the data points far away from the centroids, the ranks of the travel-time based similarity values are larger than 10,000 plus the corresponding distance based ones. So, compared to the distance based similarity, the travel-time based similarity gives larger relative similarity values between the centroids and the data points further away from them, while giving smaller relative similarity values between the centroids and the data points close to them, which results in a more narrow similarity distribution within a cluster and thus is beneficial to the clustering process.

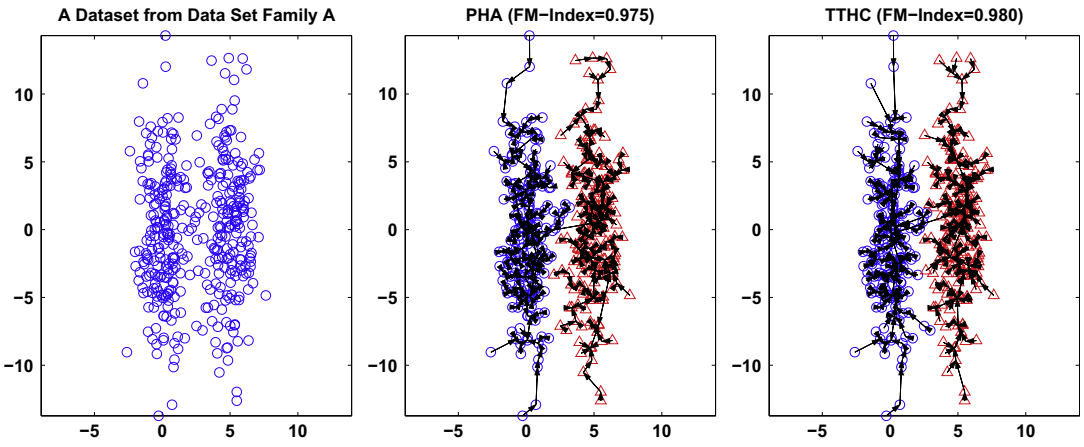


Fig. 2. The clustering results and the tree structures produced by PHA and TTHC for a dataset from *Dataset Family A*. The arrows are used to indicate the parent point, and different symbols are used to show the data points belong to different clusters.

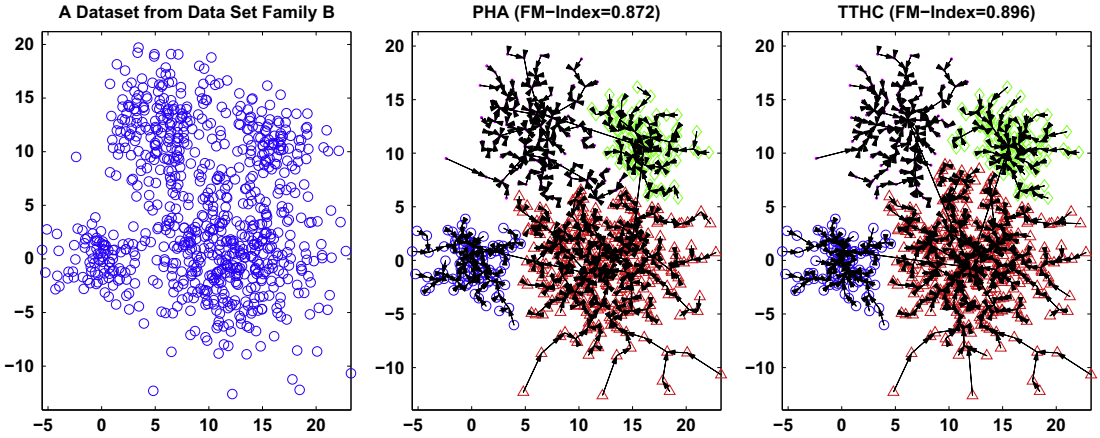


Fig. 3. The clustering results and the tree structures produced by PHA and TTHC for a dataset from *Dataset Family B*. The arrows are used to indicate the parent point, and different symbols are used to show the data points belong to different clusters.

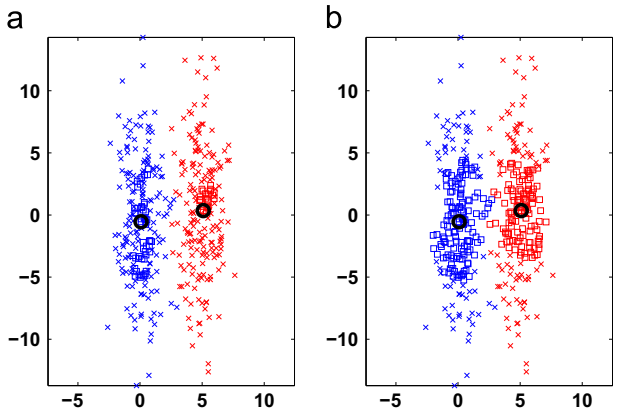


Fig. 4. Comparison results of two similarity measures for the first dataset in *Dataset Family A*. The circles indicate the centroids of the clusters given by the benchmark. (a) The x mark indicates that the rank of the travel-time based similarity value is larger than the corresponding rank of the distance based similarity value. (b) The x mark indicates that the rank of the travel-time based similarity value is larger than 10,000 plus the corresponding rank of the distance based similarity value.

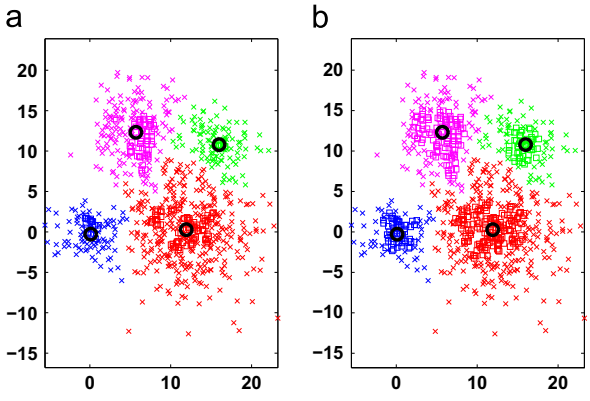


Fig. 5. Comparison results of two similarity measures for the first dataset in *Dataset Family B*. The circles indicate the centroids of the clusters given by the benchmark. (a) The x mark indicates that the rank of the travel-time based similarity value is larger than the corresponding rank of the distance based similarity value. (b) The x mark indicates that the rank of the travel-time based similarity value is larger than 10,000 plus the corresponding rank of the distance based similarity value.

4.2. Experiments with four real datasets

Four popular real datasets, Iris, Wine Quality (White), Wine Quality (Red), and Yeast from UCI Machine Learning Repository [19] are also selected to evaluate the clustering methods. They are all labeled datasets, so the benchmarks are available. The FM-indices and the running time in seconds for the four datasets are shown in Table 3. TTHC has produced the best results for all the datasets. It is also noted that TTHC and PHA run much faster than the other traditional methods when applied to large datasets such as Wine Quality and Yeast. For the Iris dataset with 4 attributes,

Table 3
Experimental results for four real datasets.

Methods	Iris		Wine Quality (White)		Wine Quality (Red)		Yeast	
	Time	FM-Index	Time	FM-Index	Time	FM-Index	Time	FM-Index
Single	0.020	0.7635	414.5	0.5696	15.36	0.5960	11.72	0.4700
Complete	0.023	0.7686	417.7	0.3961	14.92	0.3776	11.91	0.3160
Ward's	0.021	0.8222	418.8	0.2839	15.13	0.3145	12.11	0.2689
PHA	0.008	0.8670	3.623	0.5693	0.367	0.5018	0.321	0.4694
TTHC	0.008	0.9234	3.691	0.5697	0.367	0.5967	0.320	0.4731

Table 4
Experimental results for two large high-dimensional real datasets.

Methods	Spambase		USPS	
	Time	FM-Index	Time	FM-Index
Single	344.1	0.7227	5191	0.3460
Complete	346.2	0.7225	4982	0.2752
Ward's	346.8	0.6711	4996	0.4165
PHA	3.629	0.7226	51.73	0.3460
TTHC	3.719	0.7227	52.63	0.3461

TTHC produces a FM-Index of 0.9234. It is found that 144 out of 150 points in Iris are assigned correctly by TTHC. However, for the two Wine Quality datasets with 11 attributes and the Yeast dataset with 8 attributes, although TTHC has produced the best results, the FM-Indices produced by TTHC are all lower than 0.6, and the results of TTHC are slightly better than these of Single Linkage.

4.3. *Experiments with two large high-dimensional real datasets*

To further evaluate the proposed method, it is compared with other methods on two large high dimensional datasets: Spambase from UCI Machine Learning Repository [19] and the USPS handwritten digits [20], which are both labeled datasets. The Spambase dataset has 4601 data points with 57 attributes, while the USPS dataset has 11,000 data points with 256 attributes. The recorded FM-indices and the running time in seconds for the two datasets are shown in Table 4. For the Spambase dataset, TTHC and the Single Linkage have produced the best results. For the USPS dataset, the Ward's method has produced the best result, and TTHC has produced the second best result measured by the FM-indices. The results also indicate the limitations of the proposed methods on high dimensional datasets: compared with the results on low dimensional datasets such as Iris and two synthetic dataset families, TTHC can no longer produce obviously better results than the other methods. But TTHC still produces slightly better results than PHA for the two high dimensional datasets. It is also noticed that for the two datasets, TTHC and PHA both can run about 100 times faster than the three traditional methods, which shows that the efficiency is a big advantage when applying the proposed method on large datasets.

5. Conclusion and discussion

We have proposed a novel similarity measure for hierarchical clustering by introducing a travel time between data points under a hypothetical potential field. Compared with four other hierarchical clustering methods, the proposed TTHC method produces competitive and promising results when applied to different datasets. When applied to two large high-dimensional datasets, the proposed method can run 100 times faster than the traditional methods while still producing competitive results. For two data points, if the difference between the potential values of them becomes larger, the travel time between them becomes shorter, and the similarity between them becomes larger. For the points close to the border areas and far away from the centroids of the clusters, the potential differences between the data points and the centroids are usually larger than the potential differences between the centroids and the data points close to them, so the relative similarity values between the centroids and the border points are increased using the travel-time based similarity measure, which results in a narrow similarity distribution within a cluster as proved by our experiments. This may explain why using the travel time instead of the distance between data points can improve the quality of clustering. Another big advantage of the using the proposed TTHC method on large datasets is its high efficiency: it can run 100 times faster than the traditional methods. However, we also noticed the limitations of the TTHC method when applied to high dimensional datasets. Although TTHC uses the travel-time instead of the distance, the travel-time still needs to be computed from a function of the Euclidean squared distance as shown in (6). We have tried to use the Mahalanobis distance in computing the travel time in (6), but still cannot improve the results of the two large high-dimensional datasets. The reason may be because in the high dimensional space the Euclidean distances between different data points tend to become similar to each other, and using the Mahalanobis distance also cannot change the situation much. In future work, we plan to explore different ways for further improving the TTHC method on high dimensional data.

Acknowledgments

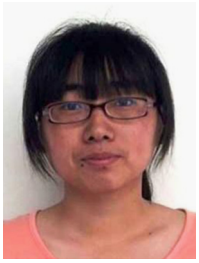
This work is supported by the National Natural Science Foundation of China (Grant no. 61272213). The authors would also like to thank anonymous reviewers for their valuable comments.

References

[1] M.G. Omran, A.P. Engelbrecht, A. Salman, An overview of clustering methods, Intell. Data Anal. 11 (2007) 583–605.
[2] M. Filippone, F. Camastra, F. Masulli, S. Rovetta, A survey of kernel and spectral methods for clustering, Pattern Recognit. 41 (2008) 176–190.
[3] A.K. Jain, Data clustering: 50 years beyond *K*-means, Pattern Recognit. Lett. 31 (2010) 651–666.
[4] R. Gil-García, A. Pons-Porrata, Dynamic hierarchical algorithms for document clustering, Pattern Recognit. Lett. 31 (2010) 469–477.
[5] I. Assent, Clustering high dimensional data, WIREs Data Min. Knowl. Discov. 2 (2012) 340–350.
[6] H. Yu, M. Gerstein, Genomic analysis of the hierarchical structure of regulatory networks, Proc. Natl. Acad. Sci. USA 103 (2006) 14724–14731.
[7] J. Wang, M. Li, J. Chen, Y. Pan, A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks, IEEE Trans. Comput. Biol. Bioinform. 8 (2011) 607–620.
[8] F. Murtagh, P. Contreras, Algorithms for hierarchical clustering: an overview, WIREs Data Min. Knowl. Discov. 2 (2012) 86–97.
[9] W.E. Wright, Gravitational clustering, Pattern Recognit. 9 (1977) 151–166.
[10] J. Gómez, D. Dasgupta, O. Nasraoui, A new gravitational clustering algorithm, in: Proceedings of the 3rd SIAM International Conference on Data Mining, San Francisco, CA, USA, May 1–3, 2003, pp. 83–94.

[11] Y. Endo, H. Iwata, Dynamic clustering based on universal gravitation model, modeling decisions for artificial intelligence, Lect. Notes Comput. Sci. 3558 (2005) 183–193.
[12] L. Peng, B. Yang, Y. Chen, A. Abraham, Data gravitation based classification, Inf. Sci. 179 (2009) 809–819.
[13] J. Li, H. Fu, Molecular dynamics-like data clustering approach, Pattern Recognit. 44 (2011) 1721–1737.
[14] Y. Lu, Y. Wan, Clustering by sorting potential values (CSPV): a novel potential-based clustering method, Pattern Recognit. 45 (2012) 3512–3522.
[15] S. Shi, G. Yang, D. Wang, W. Zheng, Potential-based hierarchical clustering, in: Proceedings of the 16th International Conference on Pattern Recognition, Quebec, Canada, August 11–15, 2002, pp. 272–275.
[16] Y. Lu, Y. Wan, PHA: a fast potential-based hierarchical agglomerative clustering method, Pattern Recognit. 46 (2013) 1227–1239.
[17] Y. Lu, X. Hou, X. Chen, Measuring cluster similarity by the travel time between data points, in: Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods, Angers, Loire Valley, France, March 6–8, 2014, pp. 14–20.
[18] E.B. Fowlkes, C.L. Mallows, A method for comparing two hierarchical clusterings, J. Am. Stat. Assoc. 78 (1983) 553–569.
[19] A. Frank, A. Asuncion, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml], University of California, School of Information and Computer Science, Irvine, CA, 2010.
[20] The USPS dataset is available at ⟨http://www.cs.nyu.edu/~roweis/data.html⟩.

Yonggang Lu received both the B.S. and M.S. Degrees in Physics from Lanzhou University, Lanzhou, China in 1996 and 1999 respectively. Later he received the M.S. and Ph.D. Degrees in Computer Science from New Mexico State University, Las Cruces, NM, USA in 2004 and 2007 respectively. He finished some of the Ph.D. work at Los Alamos National Lab, NM, USA. He is now an Associate Professor in the School of Information Science and Engineering, Lanzhou University, Lanzhou, China. His main research interests include pattern recognition, image processing, neural networks, and bioinformatics.



Xiaoli Hou is a graduate student in the School of Information Science and Engineering, Lanzhou University, majored in computer science. Her research interests include pattern recognition and dimension reduction.

Xurong Chen received both the B.S. and M.S. Degrees in Physics from Lanzhou University, Lanzhou, China in 1996 and 1999 respectively. Later he received the M.S. and Ph.D. Degrees in Experimental Particle Physics from University of South Carolina, Columbia, SC, USA in 2004 and 2008 respectively. He is now the leader of high energy nuclear physics group in the Institute of Modern Physics, Chinese Academy of Sciences. His main research interests include experimental high energy physics, data mining, and statistical data analysis.