Applied Mechanics and Materials
ISSN: 1662-7482, Vols. 556-562, pp 3654-3657
doi:10.4028/www.scientific.net/AMM.556-562.3654

Online: 2014-05-23

Similarity Metric Based on Resistance Distance and Its Applications to Data Clustering*

Gengqi Guo[1,3,a], Wenjun Xiao[2,b], Bin Lu[1,c]

[1]School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, P. R. China

[2]School of Software Engineering, South China University of Technology, Guangzhou 510006, P. R. China

[3]GuangDong Communication Polytechnic, Guangzhou 510650, P. R. China

[a]email:953578286@qq.com, [b]email: wjxiao@scut.edu.cn, [c]email: 2259975946@qq.com
*Project supported by the National Natural Science Foundation of China (Grant Nos. 61370003, 61170313, 61103037)

Keywords: Clustering; High-performance comparison; Network similairty; Resistance distance; Similarity metric.

Abstract. We propose a new clustering method that uses a similarlity metric deived from electrical resistance networks. The proposed metric allows us to quantify the mutual relevancy between data objects or nodes. We show how to derive the metric from the data collection and how to apply it to various application contexts. Our theoretical analyses and experiments show the excellent potential of the method to identifying clusters of networks and to improving data clustering performance on a number of data sets.

Introduction and related work

A fundamental challenge in data clustering is to find clusters that are relevant for data objects of interest [1-3]. More specifically, data objects in the same cluster exhibit higher similarity and data objects in the different clusters are less similar. Data clustering may find the properties and features of data in data sets. Therefore, it has broad applications in domains such as image segmentation, text analysis, and chemical data clustering. Clustering methods that have been proposed include *k*-means, hierarchical, and spectral. Among these, spectral clustering, which is based on similarity matrix of data objects and applies Laplacian matrix of graphs, offers many benefits [1]. The similarity matrix forms an important focal point in such data clustering schemes [2], but so far no general rules have been developed for dealing with similarity matrices.

In this study, we tackle the data clustering problem by defining a new similarity metric based on the resistance of electrical networks [2]. The metric is derived from the effective resistance distances between pairs of vertices in an undirected, weighted graph, wherein nodes represent data objects that are linked together based on some relationships. Edge weights represent the strength of relationships and are interpreted as electrical resistances. Based on the resistance distances between pairs of objects, we demonstrate how to assess the similarity between objects by deriving the metric from the data collection and applying it to clustering data for various application domains. Our theoretical analyses and experiments confirm that the method finds good clusters of networks and improves the data clustering performance on several datasets.

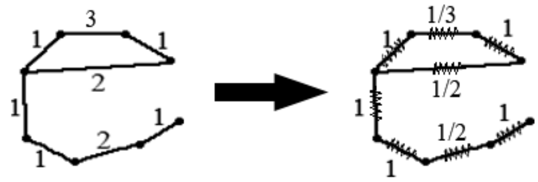
We cover the needed background and introduce the basis of our method, that is, computing resistance distances, in Section 2. In Section 3, we discuss and analyze the application of resistance distance in clustering. In Section 4, we evaluate the performance of our method using a number of data sets. Finally we conclude the paper and suggest directions for future work in Section 5.

Resistance distance in networks

Our objective is to define a measure in the data space that allows us to judge the similarity of data objects [2]. To accomplish this, our model consists of an undirected weighted graph $G = (V, E,$

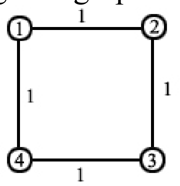
w), where the node set V corresponds to the set of data objects, the edge set E designates the existence of similarity between objects, and weight w of an edge defines the strength of the similarity associated with a particular edge. We begin by building a similarity graph (Figure 1a), where greater similarity between two objects is represented by a higher weight.

An edge represents direct similarity between two objects. To measure similarity between any pair of objects, we view the weighted graph as an electrical network [3-6], as depicted in Figure 1b. In the electrical version of the graph, the weight w_{ij} of the edge (i, j) corresponds to the conductance c_{ij}, which is the reciprocal of its electrical resistance, that is $r_{ij} = 1/c_{ij} = 1/w_{ij}$. We can use the electrical network thus obtained to compute the effective electrical resistance between any two nodes, which forms the basis of our similarity measure (see the example in Figure 2). This measure, called "resistance distance," has an intutitive random-walk interpretation [1].



(a) (b)

Fig. 1. A weighted graph and its associated electrical resistance network.



$$R=\begin{pmatrix} 0 & 0.7500 & 1.0000 & 0.7500 \\ 0.7500 & 0 & 0.7500 & 1.0000 \\ 1.0000 & 0.7500 & 0 & 0.7500 \\ 0.7500 & 1.0000 & 0.7500 & 0 \end{pmatrix}$$

Fig. 2. Simple example of computing resistance distances in a 4-node graph.

In general, the computation of the resistance between any two nodes is fairly complex, requiring us to consider all the paths between them. We compute the resistance distance by means of graph Laplacian $L = D - A$ [4, 5, 7], where A represents the adjacency matrix of the resistance network and D is the degree matrix of the graph (D is diagonal, with d_{ii} being the degree of node i). For $i = 1, 2, \ldots, n$, the (i, i) diagonal entry of L is defined to make the ith row sum equal to zero. For a connected graph, effective resistance is computed as:

$$r_{ij} = L^{+}_{ii} + L^{+}_{jj} - 2L^{+}_{ij}$$

where L^{+} represents the pseudoinverse of L and $L^{+} = (L + (1/n)J)^{-1} - (1/n)J$, with J being the matrix all of whose elements are 1. From [5], we have $X = (L + (1/n)J)^{-1}$ and:

$$r_{ij} = x_{ii} + x_{jj} - 2x_{ij}$$

Alternatively, the following formula [7]:

$$r_{ij} = |L(i, j)| / |L(i)|$$

where $L(i)$ and $L(i, j)$ are the matrices obtained from L by deleting its ith row and ith column or by deleting both its ith and jth rows and columns, respectively, is also available to us.

Data clustering based on resistance distance

We now introduce our clustering method, using r_{ij} to be the similarity measure among n nodes or data objects $\{1, 2, \ldots, n\}$. The terms r_{ij} are elements in the similarity matrix $R = (r_{ij})$, computed by one of the preceding two formulas. We can construct a sequence of vectors g_1, $g_2, \ldots, g_n$ such that:

$$\| g_i - g_j \|^2 = r_{ij}$$

where $\| z \|$ denotes the norm of the vector z. This end can be met by taking g_i to be the ith column of $X^{1/2}$, that is:

$$[g_1\ g_2\ \ldots\ g_n] = X^{1/2}$$

Expanding $\| g_i - g_j \|^2$, we get $x_{ii} + x_{jj} - 2x_{ij}$, which equals r_{ij} by one of our previous formulas.

According to Theorem 2 of [5], the eigenvalues of X are $1/\mu_1, 1/\mu_2, \ldots, 1/\mu_{n-1}, 1$, and there exists an orthogonal matrix W such that:

$$X = W^{t}\, diag[1/\mu_1, 1/\mu_2, \ldots, 1/\mu_{n-1}, 1]W$$

Thus:

$$X^{1/2} = W^t\, diag[1/\mu_1{}^{1/2}, 1/\mu_2{}^{1/2}, \ldots , 1/\mu_{n-1}{}^{1/2}, 1]W$$

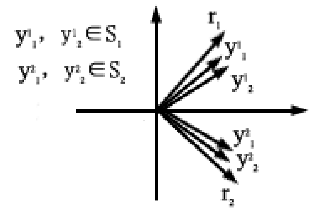


Fig. 3. Example for explaining the theorem.

Let $L_1 = D_1 - R$, where R is the similarity matrix and D_1 is the degree diagonal matrix. Recall that, for $i = 1, 2, \ldots , n$, the (i, i) entry of L_1 is defined to make the ith row sum equal to 0. Compute k unit eigenvectors of the equation $D_1{}^{-1}L_1u = \lambda u$. Let clustering number be k, $Y = [u_1\ u_2 \ldots u_k]$, and cluster the k-dimensional row vectors of Y. The clustering results are $S_1, S_2, \ldots , S_k$ and the corresponding row vectors are y_j^i, $i = 1, 2, \ldots , k$, $j = 1, 2, \ldots , n_i$. Let $r_1, r_2, \ldots , r_k$ be k-dimensional normal orthogonal vectors. Then, we have the following result.

Theorem: $\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left\|y_j^i - r_i\right\|^2 \le (1+\frac{k}{n}+2\sqrt{k/n}).$

Proof: Omitted.

In way of further explanation, consider that according to the theorem, members of cluster S_i are closer to the k-dimensional normal orthogonal vector r_i. Thus, in Figure 3, based on the computed similarity matrix R of Figure 2 and the row vectors of Y computed in our theorem, we obtain $y_1{}^1, y_2{}^1 \in S_1$ and $y_1{}^2, y_2{}^2 \in S_2$.

In comparison with the results of [1], our clustering algorithm leads to better performance. If the network is connected, we proceed as follows:

(1) Input the clustering number k and compute $R = (r_{ij})$ by one of two formulas;

(2) Make n k-dimensional row vectors of $Y = [u_1\ u_2 \ldots u_k]$ as data object set;

(3) Cluster the row vectors of Y by the algorithm of k-means;

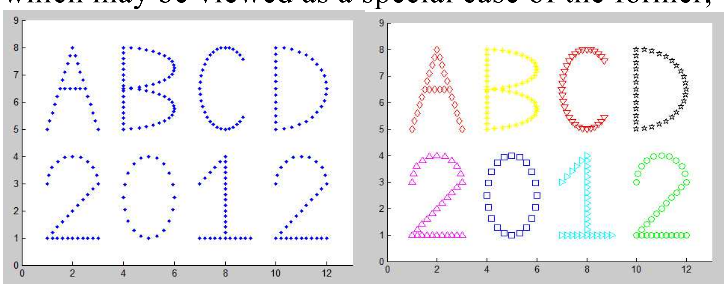
(4) The original object i belongs to S_j iff the ith row of Y belongs to S_j.

For a disconnected graph, we can handle each connected component by separately forming its vector of object set and then adding some constant vector to each of the vectors thus derived. In this way, the Euclidean distances between connected components become larger.

Experimental results

To verify the clustering performance of our algorithm, we simulate a series of experiments using different data sets, and then compare our results with those of the k-means algorithm and the NJW algorithm [1].

For the specific clustering problems considered, we describe the relations of data objects by means of graphs, which come in two varieties of disconnected and connected. We begin our experiments with disconnected graphs and then proceed to examples involving connected graph, which may be viewed as a special case of the former, with just one connected component.



(a) (b)

Fig. 4. Example of clustering in a disconnected graph.

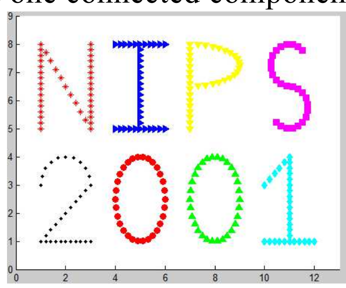


Fig. 5. Example of data clustering by the NJW algorithm of [1].

For points in the 2-dimensional space, we use the Euclidean distance metric to decide whether any two points are adjacent. Consider, for example, the 2-dimensional point set ‘ABCD2012’ of Figure 4a. We proceed by deriving the coordinates of each point, computing the Euclidean distance between them, and generating their Euclidean distance matrix.

To deal with this point set as graph, we set a threshold according to practical considerations, and deem two points connected iff the Euclidean distance between them is less than the threshold. Thus, the point set forms a disconnected graph with its corresponding adjacency matrix. We then compute the connected components of the graph and form the adjacency matrix of each component. We next compute the resistance distance between pairs of nodes by either of the two available formulas to generate the resistance distance matrix for each component. Next, we find the degree diagonal matrix of resistance matrix, and continue with calculating the Laplacian matrix L of each component. Finally, we calculate the eigenvalues and eigenvectors of the matrix $D_1^{-1}L_1$ and cluster these k eigenvectors by the algorithm of k-means, leading to the result shown in Figure 4b, where the success of our algorithm in separating the components is indicated by using different colors and tags.

To compare the performance of our algorithm with those of the k-means algorithm and the NJW algorithm of [1], we performed additional experiments. In the first of our experiments, we repeated the clustering of the data presented in Figure 1a of [1], with the results presented in Figure 5. Success in this experiment required an adjustment in the parameter σ, a detail that required much effort, because it was not introduced in [1].

Conclusion and further research

We have proposed a new similarity measure and applied it to different datasets for clustering. The theoretical analysis and experimental results show the method to be capable of finding good clusters of networks, while improving the data clustering performance compared with the other methods on several datasets.

We plan to seek potential improvements in our algorithm and to make more extensive experimental evaluations. In our experiments, we will seek worst-case conditions that expose weaknesses of our algorithm, in an effort to make modifications to counteract the problems. We will also assess the performance for various data-set categories so as to identify the best application areas and those to be avoided or marked as candidates with evaluation with multiple methods. Also, applications of our method to network clustering research may be pursued.

References

[1] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On Spectral Clustering: Analysis and an Algorithm,” Advances in Neural Information Processing Systems 14, pp. 849–856, MIT Press, 2002.

[2] A. K. Jain, M. N. Murty, and P. J. Flynn, “Data Clustering: A Review, ACM Computing Surveys, Vol. 31, No. 3, pp. 264-323, 1999.

[3] S. G. Wang and M. Hauskrecht, “Effective Query Expansion with the Resistance Distance Based Term Similarity Metric,” Proc. 33rd ACM SIGIR Conf., pp. 715-716, July 2010.

[4] D. J. Klein and M. Randic, “Resistance Distance,” J. Mathematical Chemistry, Vol. 12, pp. 81-95, 1993.

[5] W. J. Xiao and I. Gutman, “Resistance Distance and Laplacian Spectrum,” Theoretical Chemistry Accounts, Vol. 110, pp. 284-289, 2003.

[6] P. G. Doyle and J. L. Snell, Random Walks and Electrical Networks, The Mathematical Association of America, 1984.

[7] R. B. Bapat, I. Gutman, and W. J. Xiao, “A Simple Method for Computing Resistance Distance,” Zeitschrift Fur Naturforschung, Vol. 58a, pp. 494-498, 2003.