Pattern Recognition 46 (2013) 3056–3065

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Agglomerative clustering via maximum incremental path integral

Wei Zhang [a], Deli Zhao [a], Xiaogang Wang [b,*]

[a] Department of Information Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong
[b] Department of Electronic Engineering, The Chinese University of Hong Kong, Rm. 415, Ho Sin Hang Engineering Building, Shatin, Hong Kong

ARTICLE INFO

Article history:
Received 18 July 2012
Received in revised form
3 April 2013
Accepted 10 April 2013
Available online 24 April 2013

Keywords:
Agglomerative clustering
Path integral
Graph algorithms
Random walk

ABSTRACT

Agglomerative clustering, which iteratively merges small clusters, is commonly used for clustering because it is conceptually simple and produces a hierarchy of clusters. In this paper, we propose a novel graph-structural agglomerative clustering algorithm, where the graph encodes local structures of data. The idea is to define a structural descriptor of clusters on the graph and to assume that two clusters have large affinity if their structural descriptors undergo substantial change when merging them into one cluster. A key insight of this paper to treat a cluster as a dynamical system and its samples as states. Based on that, *Path Integral*, which has been introduced in statistical mechanics and quantum mechanics, is utilized to measure the stability of a dynamical system. It is proposed as the structural descriptor, and the affinity between two clusters is defined as *Incremental Path Integral*, which can be computed in a closed-form exact solution, with linear time complexity with respect to the maximum size of clusters. A probabilistic justification of the algorithm based on absorbing random walk is provided. Experimental comparison on toy data and imagery data shows that it achieves considerable improvement over the state-of-the-art clustering algorithms.

1. Introduction

Clustering is a classical machine learning topic with wide applications in diverse fields. It includes two major categories [1,2]: partitional clustering, which determines all clusters at once, and hierarchical clustering, which creates a hierarchy of clusters in a bottom-up (or agglomerative) process by merging small clusters or in a top-down (or divisive) process by dividing large clusters into small ones. Numerous algorithms have been proposed, such as *k*-means [2], spectral clustering [3–8] and affinity propagation [9], and achieved great success.

This work stands on the success of agglomerative clustering, which is commonly used because it is conceptually simple and produces a hierarchy of clusters. Beginning with a large number of initial small clusters, the agglomerative clustering algorithms iteratively select two clusters with the largest affinity under certain measures to merge, until some stopping condition is reached. Therefore, the affinity measure of clusters is critically important. Linkages, e.g., single linkage, complete linkage and average linkage [2], define the affinity based on inter-cluster pairwise distances. Since pairwise distances do not well capture the global structures of data, complete linkage and average linkage fail on clustering data with manifold structures. Although single linkage performs better in this case, it is very sensitive to noisy distances. Examples can be found in Fig. 3. Lossy coding theory of multivariate mixed data [10] characterizes the affinity of two clusters with the variational coding length of coding the merged cluster against coding the two clusters separately. It exhibits exceptional performance for clustering multivariate mixed Gaussian or subspace data, but is not suitable for data from other distributions. There are also approaches based on probabilistic models, such as Bayesian hierarchical clustering [11]. They all assume the forms of underlying data distributions, which are unknown in many applications.

In this paper, we propose a novel graph-structural agglomerative clustering algorithm. Although the power of graphs has been extensively exploited in clustering [3,5,12,13], semi-supervised learning [14,15], and manifold learning [16], they have received little attention in agglomerative clustering. In our algorithm the pairwise distances are only used to build a neighborhood graph, since studies [16] show the effectiveness of using local neighborhood graphs to model data lying on a low-dimensional manifold embedded in a high-dimensional space. Then a structural descriptor is defined to characterize the global structure of a cluster from the local information encoded by the graph. It is assumed that two clusters have large affinity if their structural descriptors undergo substantial change when merging them into one cluster.

We propose *path integral* as the structural descriptor of clusters. Paths are a fundamental concept of graph theory, and are used in many graph-based algorithms. The description of paths gives rich information about the data. There has been a lot of research work

* Corresponding author. Tel.: +852 39438283; fax: +852 26035558.
E-mail addresses: wzhang009@gmail.com (W. Zhang), zhaodeli@gmail.com (D. Zhao), xgwang@ee.cuhk.edu.hk (X. Wang).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.04.013

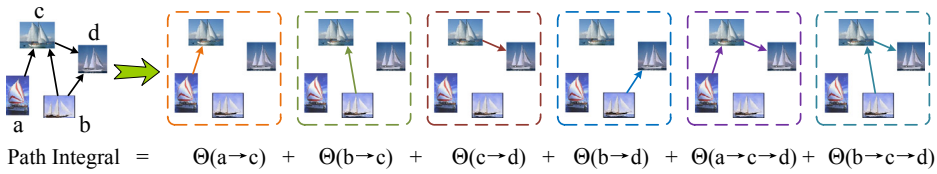

Fig. 1. A toy example on the path integral description of a cluster. There are four length-1 paths and two length-2 paths in the cluster. The path integral is computed as the sum of contributions of these paths. How to obtain each path's contribution is described in Section 3. For clarity, the vertices outside the cluster and the outer links are not shown.

on studying various aspects of paths on graphs, such as finding the shortest paths between nodes [17,18] or computing the similarity between two nodes over paths [19,20]. For example, Saerens et al. [18] proposed the randomized shortest path problem, which allows a route agent to follow different paths according to some probability distributions instead of only following the shortest path connecting a source and a destination. Their proposed model could be used to measure the dissimilarity between two nodes accounting for multiple paths. However, the purpose of this paper is to develop a structural descriptor of clusters, instead of finding the shortest path between nodes or computing the pairwise similarities between samples. The concept of path integral was first introduced in statistical mechanics and quantum mechanics [21–23], where it summed up the contributions of all possible paths to the evolution of a dynamical system. In this work, we provide our own formulation of path integral and its probabilistic interpretation based on absorbing random walk. If we treat a cluster as a dynamical system, with vertices as states and edge weights as transition probabilities between states, then the path integral measures the stability of the dynamical system, i.e. randomly starting with any state of the dynamical system, the probability of remaining within the same system after certain steps of transitions. An example is shown in Fig. 1. The affinity of two clusters is defined as the *incremental path integral* after merging them. An intuitive explanation is that if two clusters are closely connected, their stability will greatly increase after merging them. We show that the incremental path integral can be computed in a closed-form exact solution, with linear time complexity with respect to the maximum size of clusters. Experimental comparisons on toy data and imagery data show the excellent performance of the proposed algorithm and its robustness to parameter settings.

Our algorithm has several advantages compared with existing methods. First, since it measures the affinity of clusters based on the neighborhood graph instead of directly on pairwise distances between any pairs of samples, it can better cluster data on manifolds and is more robust to noisy distances compared with linkage algorithms [2] widely used in agglomerative clustering. Second, different from spectral clustering [3,5] and clustering on the manifold embedding results, it does not use any relaxation or approximation. The graph structural merging strategy also makes our algorithm more robust to noisy links than spectral clustering, because our structural descriptor involves solving a linear system, while the spectral clustering utilizes eigen-decomposition. Solving eigen-vectors is more sensitive to noise than solving a linear system [24,12]. Examples in the bottom row of Fig. 3 show that our algorithm can handle for multi-scale data, i.e., a dataset that contains structures with different densities and sizes, which is the limitation of spectral clustering [25,26]. Third, it only requires the pairwise distances or similarities of samples without any assumptions on the underlying data distributions. This is useful in the case when the vector representations of data are not available. Therefore, it has better flexibility and generalization than other agglomerative clustering methods such as lossy coding theory [10] and Bayesian hierarchical clustering [11].

The paper is organized as follows. For ease of reading, the overall clustering algorithm is first outlined in Section 2. Then, the theoretical framework of path integral and incremental path integral is presented in Section 3. Section 4 provides a probabilistic interpretation of our algorithm based on absorbing random walk. Experimental validations and conclusion are given in Sections 5 and 6, respectively.

2. Graph-structural agglomerative clustering

Our algorithm iteratively merges two clusters with maximum affinity on a directed graph.

Building the digraph. Given a set of sample vectors $\mathcal{X}=\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, we build a directed graph $G=(V,E)$, where V is the set of vertices corresponding to the samples in $\mathcal{X}$, and E is the set of edges connecting vertices. The graph is associated with a weighted adjacency matrix $\mathbf{W}=[w_{ij}]$, where w_{ij} is the pairwise similarity between $\mathbf{x}_i$ and $\mathbf{x}_j$ defined as

$$w_{ij}=\begin{cases}\exp\left(-\dfrac{\mathrm{dist}(i,j)^2}{\sigma^2}\right), & \text{if } \mathbf{x}_j\in\mathcal{N}_i^K,\\ 0, & \text{otherwise.}\end{cases} \tag{1}$$

$\mathrm{dist}(i,j)$ is the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\mathcal{N}_i^K$ is the set of K-nearest neighbors of $\mathbf{x}_i$. If $\mathbf{x}_j\in\mathcal{N}_i^K$, there is an edge pointing from $\mathbf{x}_i$ to $\mathbf{x}_j$ with weight w_{ij}. σ^2 is estimated by $\sigma^2=[\sum_{i=1}^n\sum_{\mathbf{x}_j\in\mathcal{N}_i^3}\mathrm{dist}(i,j)^2]/[3n(-\ln a)]$.[1] K and a are free parameters to be set.

We define a random walk model on this directed graph. Denote the transition probability matrix as $\mathbf{P}$, whose element p_{ij} is the one-step transition probability from vertex i to vertex j. $\mathbf{P}$ is calculated as

$$\mathbf{P}=\mathbf{D}^{-1}\mathbf{W}, \tag{2}$$

where $\mathbf{D}$ is a diagonal matrix whose diagonal element $d_{ii}=\sum_{j=1}^n w_{ij}$, such that $\sum_{j=1}^n p_{ij}=1$. The path integral of a cluster is computed by summing the paths within the cluster on the directed graph weighted by transition probabilities.

Affinity measure of clusters. Given two clusters $\mathcal{C}_a$ and $\mathcal{C}_b$, their structural affinity is measured as the amount of incremental path integral $\mathcal{A}_{\mathcal{C}_a,\mathcal{C}_b}$ when merging them, i.e.,

$$\mathcal{A}_{\mathcal{C}_a,\mathcal{C}_b}=(\mathcal{S}_{\mathcal{C}_a|\mathcal{C}_a\cup\mathcal{C}_b}-\mathcal{S}_{\mathcal{C}_a})+(\mathcal{S}_{\mathcal{C}_b|\mathcal{C}_a\cup\mathcal{C}_b}-\mathcal{S}_{\mathcal{C}_b}). \tag{3}$$

$\mathcal{S}_{\mathcal{C}_a}$ is the path integral descriptor of $\mathcal{C}_a$ and sums up all the paths in $\mathcal{C}_a$. $\mathcal{S}_{\mathcal{C}_a|\mathcal{C}_a\cup\mathcal{C}_b}$ is the conditional path integral descriptor. All the paths to be counted lie in $\mathcal{C}_a\cup\mathcal{C}_b$. However, their starting and ending vertices must be within $\mathcal{C}_a$. If the vertices in $\mathcal{C}_a$ and $\mathcal{C}_b$ are strongly connected, merging them will create many new paths for the pairs of vertices in $\mathcal{C}_a$, and therefore $\mathcal{S}_{\mathcal{C}_a|\mathcal{C}_a\cup\mathcal{C}_b}$ will be much larger than $\mathcal{S}_{\mathcal{C}_a}$. Section 4 will show that $\mathcal{S}_{\mathcal{C}_a}$ measures the cluster's stability, if $\mathcal{C}_a$ is treated as a dynamical system. An example for illustration is shown in Fig. 2. The closed-form expressions of $\mathcal{S}_{\mathcal{C}_a}$

[1] It is equivalent to setting the geometric mean of weights associated with edges pointing to 3-nearest-neighbors as a, i.e., $(\prod_{i=1}^n\prod_{\mathbf{x}_j\in\mathcal{N}_i^3}w_{ij})^{1/(3n)}=a$.