Knowledge-Based Systems 60 (2014) 58–72

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Dependence clustering, a method revealing community structure with group dependence

Hyunwoo Park [a], Kichun Lee [b,*]

[a] Industrial and Systems Engineering and Tennenbaum Institute, Georgia Institute of Technology, Atlanta, USA
[b] Industrial Engineering, Hanyang University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 1 May 2013
Received in revised form 30 December 2013
Accepted 6 January 2014
Available online 23 January 2014

Keywords:
Group dependence
Clustering
Markovian
Community structure
Mutual information

ABSTRACT

We propose a clustering method maximizing a new measure called "group dependence." Group dependence quantifies how precise a certain division of a graph is in terms of dependence distance. Built upon statistical dependence measure between points driven by Markovian transitions, group dependence incorporates the geometric structure of input data. Besides capturing degrees of positive dependence and coherence for a group division, group dependence inherently supplies the proposed clustering method with a definite decision on the depth of division. We provide an optimality aspect of the method as theoretical justification in consideration of posterior transition probabilities of input data. Illustrating its procedure using data from a known structure, we demonstrate its performance in the clustering task of real-world data sets, Amazon, DBLP, and YouTube, in comparison with selected clustering algorithms. We show that the proposed method outperforms the selected methods in reasonable settings: in particular, the proposed method surpasses modularity clustering in terms of normalized mutual information. We also show that the proposed method reveals additional insights on community structure detection according to its connectivity scale parameter.

1. Introduction

Identifying community structure in networks has been a central issue in many fields including sociology, bio-informatics, physics, and applied mathematics to name only a few. Community structure detection is a branch of the broader class problem: cluster analysis. Cluster analysis is an unsupervised task of assigning a set of objects into homogeneous groups. Depending on the number of groups in demand, the nature of clustering tasks can be divided into the following two kinds of problems. In the first case, the number of clusters is known when clustering is carried out. *Graph partitioning* is one line of research that fits this type of clustering. One of the most well known examples of such tasks that arise in computer science is assigning to multiple processors a number of inter-dependent tasks represented as a graph. Since the number of processors is likely fixed and known, a clustering algorithm that cannot consider the predefined number of clusters is of little practical use in this context. *Community structure detection*, on the other hand, pursues a slightly different goal in the task of clustering. In this setting, the number of communities is unknown beforehand. Not only grouping nodes precisely but also determining the number of meaningful clusters latent in the graph structure is of great importance in this class of problems. Social network analysis falls into this category.

Indeed, many clustering algorithms have been proposed in the data mining society. Among them is hierarchical clustering, *k*-means clustering, distribution-based clustering, and so forth. They are commonly based on the similarities or closeness among nodes. Han and Kamber [1] and Gan et al. [2] provide a thorough survey on details of the algorithms. Conceptually, in view of the number of clusters in demand, two approaches are possible in revealing the community structure in networks: agglomerate and divisive methods. Agglomerate methods start from grouping nodes with the highest similarity and repeat the process with recalculated similarities among groups and nodes. The agglomerate approach is more intuitive than the divisive approach, so it was developed earlier and has been widely used. Agglomerate hierarchical clustering is a representative example in this approach when the true number of clusters is unknown. This approach needs to be followed by an additional critical step that involves a decision criterion for the optimal number of clusters [3]. *k*-means clustering groups nodes in a similar way with predefined number of centroids. In contrast, divisive methods—possessing inherent rules for the optimal number of clusters—repeat cutting the network successively until no subdivision of the network yields gain. One

* Corresponding author. Tel.: +82 222200478.
E-mail addresses: skylee@hanyang.ac.kr, skylee1020@gmail.com (K. Lee).

http://dx.doi.org/10.1016/j.knosys.2014.01.004
0950-7051/

of the methods in this avenue is modularity-based clustering proposed by Newman [4]. Modularity measures how precise a division of the network is against a graph with edges placed at random and has played an essential role in detecting community structures in networks. One drawback of this method is that the random modularity measure employs a fixed global model, which assumes that each node can be linked to any other nodes of the network whether they are large or small regardless of the geometric structure of the network. Thus, it cannot adjust the level of resolution or the scale on which the modularity measure relies. Therefore, it would be desirable to propose another clustering method which not only includes an inherent rule for the optimal number of clusters, but also possesses flexibility in adjusting the level of scale from which it takes into account the geometric structure of the network.

In this paper, we propose a clustering method maximizing a new measure called "group dependence." Based on mutual information as well as posterior probabilities of network connections, group dependence provides flexibility in adjusting the scale on which the whole graph is viewed and the level of connectedness upon which a division of the network is evaluated in terms of dependence. Lee et al. [5] demonstrated the efficacy of the dependence concept and dependence distance in the context of dimensionality reduction. The statistical dependence measure between nodes which the proposed clustering method relies on is initially motivated by Markovian transitions among nodes and extends the concept of mutual information into a point-wise fashion. The dependence measure is also a lift measure between nodes that is widely used to capture the level of association in association rule learning. A graph can be viewed as being connected via a Markov chain, which means that the neighborhood of a node evolves through Markovian transitions. The adjacency matrix of the graph and the transition steps in the neighborhood of each node determine the neighborhood structure of the graph and the scale on which the whole graph is viewed. The degree of relative dependence for a group division against random division from the neighborhood structure, called as group dependence, is assessed as a coherence measure for the division, naturally leading to a clear answer on the division depth and a group configuration with maximized group dependence. Furthermore, the level of connectedness for subdivision is adjustable in a straightforward manner. We will describe the detailed machinery and performance of the algorithm in the following sections.

This paper proceeds as follows. In Section 2, we start with defining group dependence as an extension of dependence distance. We then explain the machinery of the proposed dependence clustering and illustrate its use by clustering a simple data set. Section 3 compares performance of the dependence clustering with that of popular clustering methods such as hierarchical, spectral, and modularity clustering. We first run comparison on simulated data, and then use three real-world data sets: the karate club factions data by Zachary [6], tags co-occurrence network from Groupon, and large social and information network data from Amazon, DBLP,[1] and YouTube. We conclude with discussion and future research issues in Section 4.

2. The dependence clustering

In this section, we first briefly summarize the concept of statistical dependence. We then introduce the concept of group dependence to measure coherence for a group division in terms of dependence, followed by a new clustering method based on group dependence. We also provide an optimality aspect of the proposed method and further details of the algorithm.

[1] A computer science bibliography website; http://www.informatik.uni-trier.de/ley/db/.

2.1. Dependence

Suppose we have n data points in $\mathbb{R}^b$. Each data point $x_1, \ldots, x_n$ represents a node in an undirected graph. Denote the set of nodes by $\Omega = \{x_1, \ldots, x_n\}$. We view the graph as a Markov chain assuming that the whole chain is ergodic and all transitions follow the Markovian property. We can then define the neighborhood of a node as the nodes that can be reached through Markovian transitions from the focal node. This neighborhood structure provides a foundation to calculate the distance in a new measure between nodes in a graph. To illustrate the concept of neighborhood transitions, we provide Fig. 1, in which the node 3 is closer to the node 1 than the node 2 is to the node 1 in Euclidean distance using the arrows. However, in consideration of the edge structure representing Markovian transition between the nodes, the graph geometry suggests that the distance between nodes 1 and 2 ought to be smaller than that between nodes 1 and 3 in terms of neighborhood-transition steps: node 2 is four transition steps away from node 1 while node 3 is thirteen.

Lee et al. [5] proposed "dependence distance" between two nodes for a Markov chain in the t-step-wide neighborhood evolution in Ω, where t is an exogenously given parameter. They demonstrated its use in their proposed dimensionality reduction algorithm, while we devise a new community structure detection algorithm based on it. So we summarize the concept briefly in this section and propose a new measure on dependence suitable for community structure detection in the next section. We assume that any two nodes in Ω can be connected via Markovian transitions, although the probability of a transition decays as the number of steps between the two nodes increases. Let us define X_t as a random walk that represents a node (or state) at tth transition. We define the statistical dependence between a node in the initial state (X_0) and another node at step t (X_t) as follows:

Definition 2.1. Dependence between $x_m, x_i \in \Omega$, denoted by $Dep(X_0 = m, X_t = i)$, is

$$Dep(X_0 = m, X_t = i) = \frac{Pr(X_t = i, X_0 = m)}{Pr(X_t = i)Pr(X_0 = m)}. \tag{2.1}$$

By definition, dependence is closely linked to the point-wise mutual information. The point-wise mutual information is widely used in information theory and statistics as a measure of association. Since mutual information $I(X_0, X_t)$ between two random variables X_0 and X_t is the expectation of the point-wise mutual information for all realizations of X_0 and X_t, we can express it in terms of dependence as follows:

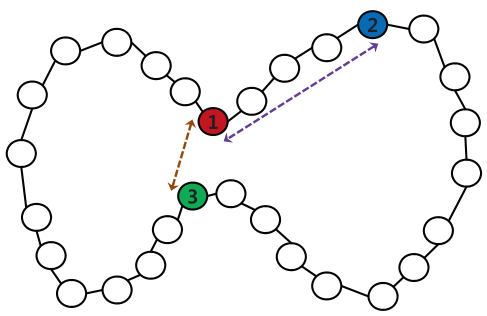

Fig. 1. The concept of neighborhood transitions is illustrated. The edge represents one step transition. Although node 3 (in green) is closer to node 1 (in red) in Euclidean distance denoted by arrows than node 2 (in blue) is to node 1, node 2 is closer to node 1 than node 3 is to node 1 in terms of neighborhood-transition steps denoted by edges. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)