Un-normlized and Random Walk Hypergraph Laplacian Un-supervised Learning

Loc Hoang Tran[1(✉)], Linh Hoang Tran[2], and Hoang Trang[3]

[1] Computer Science Department/University of Minnesota, Minneapolis, USA
tran0398@umn.edu
[2] ECE Department/Portland State University, Portland, USA
linht@pdx.edu
[3] Ho Chi Minh City University of Technology-VNU HCM
Ho Chi Minh City, Vietnam
hoangtrang@hcmut.edu.vn

Abstract. Most network-based clustering methods are based on the assumption that the labels of two adjacent vertices in the network are likely to be the same. However, assuming the pairwise relationship between vertices is not complete. The information a group of vertices that show very similar patterns and tend to have similar labels is missed. The natural way overcoming the information loss of the above assumption is to represent the given data as the hypergraph. Thus, in this paper, the two un-normalized and random walk hypergraph Laplacian based un-supervised learning methods are introduced. Experiment results show that the accuracy performance measures of these two hypergraph Laplacian based un-supervised learning methods are greater than the accuracy performance measure of symmetric normalized graph Laplacian based un-supervised learning method (i.e. the baseline method of this paper) applied to simple graph created from the incident matrix of hypergraph.

Keywords: Hypergraph Laplacian · Clustering · Un-supervised learning

1 Introduction

In data mining problem sceneries, we usually assume the pairwise relationship among the objects to be investigated such as documents [1,2], or genes [3], or digits [1,2]. For example, if we group a set of points in Euclidean space and the pairwise relationships are symmetric, an un-directed graph may be employed. In this un-directed graph, a set of vertices represent objects and edges link the pairs of related objects. However, if the pairwise relationships are asymmetric, the object set will be modeled as the directed graph. Finally, a number of data mining methods for un-supervised learning [4] (i.e. clustering) and semi-supervised learning [5,6,7] (i.e. classification) can then be formulated in terms of operations on this graph.

However, in many real world applications, representing the set of objects as un-directed graph or directed graph is not complete. Approximating complex relationship as pairwise will lead to the loss of information. Let us consider classifying a set of

P.C. Vinh et al. (Eds.): ICTCC 2014, LNICST 144, pp. 254–263, 2015.
DOI: 10.1007/978-3-319-15392-6_25

genes into different gene functions. From [3], we may construct an un-directed graph in which the vertices represent the genes and two genes are connected by an edge if these two genes show a similar pattern of expression (i.e. the gene expression data is used as the datasets in [3]). Any two genes connected by an edge tend to have similar functions. However, assuming the pairwise relationship between genes is not complete, the information a group of genes that show very similar patterns of expression and tend to have similar functions [8] (i.e. the functional modules) is missed. The natural way overcoming the information loss of is to represent the gene expression data as the hypergraph [1,2]. A hypergraph is a graph in which an edge (i.e. a hyper-edge) can connect more than two vertices. However, the clustering methods for this hypergraph datasets have not been studied in depth. Moreover, the number of hyper-edges may be large. Hence this leads to the development of the clustering method that combine the dimensional reduction methods for the hypergraph dataset and the popular hard k-mean clustering method. Utilizing this idea, in [1,2], the symmetric normalized hypergraph Laplacian based un-supervised learning method have been developed and successfully applied to zoo dataset. To the best of our knowledge, the random walk and un-normalized hypergraph Laplacian based un-supervised learning methods have not yet been developed and applied to any practical applications. In this paper, we will develop the random walk and un-normalized hypergraph Laplacian based un-supervised learning methods and apply these two methods to the zoo dataset available from UCI repository.

We will organize the paper as follows: Section II will introduce the definition of hypergraph Laplacians and their properties. Section III will introduce the un-normalized, random walk, and symmetric normalized hypergraph Laplacian based un-supervised learning algorithms in detail. In section IV, we will apply the symmetric normalized graph Laplacian based un-supervised learning algorithm (i.e. the current state of art network based clustering method) to zoo dataset available from UCI repository and compare its accuracy performance measure to the two proposed hypergraph Laplacian based un-supervised learning algorithms' accuracy performance measures. Section V will conclude this paper and the future directions of research of these methods will be discussed.

2 Hypergraph Definitions

Given a hypergraph $G=(V,E)$, where V is the set of vertices and E is the set of hyper-edges. Each hyper-edge $e \in E$ is the subset of V. Please note that the cardinality of e is greater than or equal two. In the other words, $|e| \geq 2$, for every $e \in E$. Let $w(e)$ be the weight of the hyper-edge e. Then W will be the $R^{|E|*|E|}$ diagonal matrix containing the weights of all hyper-edges in its diagonal entries.

2.1 Definition of Incidence Matrix H of G

The incidence matrix H of G is a $R^{|V|*|E|}$ matrix that can be defined as follows

$$h(v,e) = \begin{cases} 1 \; if \; vertex \; v \; belongs \; to \; hyperedge \; e \\ 0 \; otherwise \end{cases}$$