Spectral Methods for Data Clustering

Wenyuan Li
Nanyang Technological University, Singapore

INTRODUCTION

With the rapid growth of the World Wide Web and the capacity of digital data storage, tremendous amount of data are generated daily from business and engineering to the Internet and science. The Internet, financial real-time data, hyperspectral imagery, and DNA microarrays are just a few of the common sources that feed torrential streams of data into scientific and business databases worldwide. Compared to statistical data sets with small size and low dimensionality, traditional clustering techniques are challenged by such unprecedented high volume, high dimensionality complex data. To meet these challenges, many new clustering algorithms have been proposed in the area of data mining (Han & Kambr, 2001).

Spectral techniques have proven useful and effective in a variety of data mining and information retrieval applications where massive amount of real-life data is available (Deerwester et al., 1990; Kleinberg, 1998; Lawrence et al., 1999; Azar et al., 2001). In recent years, a class of promising and increasingly popular approaches — spectral methods — has been proposed in the context of clustering task (Shi & Malik, 2000; Kannan et al., 2000; Meila & Shi, 2001; Ng et al., 2001). Spectral methods have the following reasons to be an attractive approach to clustering problem:

- Spectral approaches to the clustering problem offer the potential for dramatic improvements in efficiency and accuracy relative to traditional iterative or greedy algorithms. They do not intrinsically suffer from the problem of local optima.
- Numerical methods for spectral computations are extremely mature and well understood, allowing clustering algorithms to benefit from a long history of implementation efficiencies in other fields (Golub & Loan, 1996).
- Components in spectral methods have the naturally close relationship with graphs (Chung, 1997). This characteristic provides an intuitive and semantic understanding of elements in spectral methods. It is important when the data is graph-based, such as links of WWW, or can be converted to graphs.

In this paper, we systematically discuss applications of spectral methods to data clustering.

BACKGROUND

To begin with the introduction of spectral methods, we first present the basic foundations that are necessary to understand spectral methods.

Mathematical Foundations

Data is typically represented as a set of vectors in a high-dimensional space. It is often referred as the matrix representation of the data. Two widely used spectral operations are defined on the matrix.

- EIG(A) operation: Given a real symmetric matrix $A_{n\times n}$, if there is a vector $x\in \mathrm{R}^n \neq 0$ such that $Ax=\lambda x$ for some scalar λ, then λ is called the eigenvalue of A with corresponding (right) eigenvector x. EIG(A) is an operation to compute all eigenvalues and corresponding eigenvectors of A. All eigenvalues and eigenvectors are real, that is, guaranteed by Theorem of real schur decomposition (Golub & Loan, 1996).
- SVD(A) operation: Given a real matrix $A_{m\times n}$, similarly, there always exists two orthogonal matrices $U\in \mathrm{R}^{m\times m}$ and $V\in \mathrm{R}^{n\times n}$ ($U^{\mathrm{T}}U= I$ and $V^{\mathrm{T}}V=I$) to decompose A to the form $A=USV^{\mathrm{T}}$, where $S=\mathrm{diag}(\sigma_1,\cdots,\sigma_r)\in \mathrm{R}^{r\times r}$, $r=\mathrm{rank}(A)$ and $\sigma_1\geq\sigma_2\cdots\geq\sigma_r=\cdots=\sigma_n=0$. Here, the σ_i are the singular values of A and the first r columns of U and V are the left and right (respectively) singular vectors of A. SVD(A) is called Singular Value Decomposition of A (Golub & Loan, 1996).

Typically, the set of eigenvalues (or singular values) is called the spectrum of A. Besides, eigenvectors (or singular vectors) are the other important components of spectral methods. These two spectral components have

been widely used in various disciplines and adopted to analyze the key encoding information of a complex system. Therefore, they are also the principle objects in spectral methods for data clustering.

Transformations

As observed by researchers, two key components of spectral methods — eigenvalues and eigenvectors — scale with different matrices. Therefore, before the analysis and application of them, some transformations, or more exactly, normalizations of two spectral components are needed. Although this might look a little complicated at first, this way to use them is more consistent with spectral geometry and stochastic processes. Moreover, another advantage of normalized components is due to its better relationship with graph invariants while the raw components may fail to do. There are three typical transformations often used in spectral methods.

- Laplacian: Given a symmetric matrix $A=(a_{ij})_{n\times n}$ with $a_{ij}\geq 0$, we define Laplacian $L_A=(l_{ij})_{n\times n}$ of A as

$$l_{ij}=\begin{cases}1-\frac{a_{ij}}{d_i}. & \text{if } i=j\\ -\frac{a_{ij}}{\sqrt{d_i d_j}}. & \text{if } a_{ij}\neq 0\\ 0. & \text{otherwise}\end{cases}$$

  The spectral graph theory takes this transformation (Chung, 1997).
- Variant of Laplacian: Given a symmetric matrix $A=(a_{ij})_{n\times n}$ with $a_{ij}\geq 0$, we define the variant of Laplacian $T_A=(t_{ij})_{n\times n}$ of A to be $T_A=D^{-1/2}(S-I)D^{-1/2}$. It can be easily proved that $L_A+T_A=2I$. This transformation of the matrix is often used (Li et al., 2004; Ng et al., 2001).
- Transition (or Stochastic) Matrix: Given a symmetric matrix $A=(a_{ij})_{n\times n}$ with $a_{ij}\geq 0$, we define the transition matrix $P_A=(p_{ij})_{n\times n}$ of A satisfying $p_{ij}=a_{ij}/d_i$ so that the sum of each row is 1. Apparently, P is a stochastic matrix, in the sense that it describes the transition probabilities of a Markov chain in the natural way.

In the definitions of these three matrices, $d_i=\sum_j a_{ij}$ is the sum of the i-th row vector and D=diag(d_1, …, d_n). These three matrices have real eigenvalues and eigenvectors. Moreover, the eigenvalues of Laplacian and the transition matrix lie in [0, 2] and [-1, 1], respectively. We can easily deduce from the relationship between L_A and T_A to obtain $\text{SPECTRUM}(T_A)=\{1-\lambda \mid \lambda\in \text{SPECTRUM}(L_A)\}$, where SPECTRUM(•) represents the set of eigenvalues of a matrix. Hence, the eigenvalues of T_A lie in [-1, 1] and all the conclusions and properties of L_A are also applicable to T_A. Moreover, L_A and T_A have the same eigenvectors.

Relations to Graph

As a graph is represented by its adjacency matrix, there is a close relationship between the graph and the spectral components of its adjacency matrix. It is a long history to explore the fundamental properties of a graph from the view of the spectral components of this graph's adjacency matrix in the area of mathematics. Especially, eigenvalues are closely related to almost all major invariants of graphs and thus, play a central role in the fundamental understanding of graphs. Based on this perspective, spectral graph theory has emerged and rapidly grown in recent years (Chung, 1997). Hence, many characteristics of spectral components of a matrix can be intuitively explained in terms of graphs and meanwhile graphs also can be analyzed from its spectral components. A notable case is the authority and hub vertices of the Web graph that is important to Web search as shown in HITS algorithm (Kleinberg, 1998). Another example is that the spectrum of the adjacency matrix of a graph can be analyzed to deduce its principal properties and structure, including the optimization information about cutting a graph. This view has been applied to discover and predict the clustering behavior of a similarity matrix before the actual clustering is performed (Li et al., 2004).

MAIN THRUST

Spectral Analysis for Preprocessing of Data Clustering

In clustering, one common preprocessing step is to capture or predict the characteristic of target data set before the clustering algorithm is performed. Here, the spectral analysis of a data set is introduced to predict the clustering behavior before the actual data clustering. Investigating the clustering process as shown in Jain et al. (1999), an assumption is concluded that the *feature set*, and *similarity measure* embody intrinsic knowledge of the clustering domain. Data clustering algorithms are greatly dependent on the similarity matrix. Therefore, the similarity matrix can be the principal object to be considered for decoding clustering information of the data set.

Given the similarity matrix $S=(s_{ij})_{n\times n}$, we define G(S)=<V, E, S> as its associated graph where V is the set of n vertices and E is the set of weighted edges. In this graph, each vertex v_i corresponds to the i-th column (or row) and the weight of each edge (v_i,v_j) corresponds to the non-diago-