AUTOMATIC DETERMINATION OF INTRINSIC CLUSTER NUMBER FAMILY IN SPECTRAL CLUSTERING USING RANDOM WALK ON GRAPH

Xin Zheng, and Xueyin Lin

Department of Computer Science, Tsinghua University, Beijing, P.R. China
zhengxin99@mails.tsinghua.edu.cn, lxy-dcs@mail.tsinghua.edu.cn

ABSTRACT

In spectral clustering algorithms, selecting the cluster number and determining the parameter of affinity function are two generally unsolved problems. In this paper we analyze in detail the influence of these two parameters on clustering results and show their close relationship. We further extend one of them, the cluster number, to *intrinsic cluster number family*, which is designed to achieve *stable* clustering hierarchy. Specifically, we use *random walk on graph* and *eigengap* to discover the intrinsic structure of the data. We proposed an algorithm to simultaneously determination the cluster number family and the parameter of affinity function. The experimental results on both simulated data clustering and natural image segmentation show that our proposed algorithm has many advantages.

1. INTRODUCTION

Spectral clustering methods have attracted a lot of attention in the last few years [5]. They showed impressive performances and many advantages in image and video segmentation [11][10][9]. However, like most of other clustering strategies, spectral clustering methods remain one problem unsolved: the predetermination of the cluster number k, which is a key and sensitive factor.

Some literature explored the cluster number determination problem. Fraley et.al [3] use Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) for cluster number selection. Figueirdo et.al [4] and Bischof et.al [2] integrated, respectively, the Minimum Message Length (MML) and Minimum Description Length (MDL) criterion into the EM algorithm. These methods share a common opinion: a single *golden* cluster number exists. However, this may not be always the case. We believe that sometimes hierarchical structures are hidden in data and it is more reasonable to represent such structure using an *intrinsic cluster number family*.

Based on spectral clustering algorithm, we present in this paper an algorithm to determine the *intrinsic cluster number family*. We start from a brief introduction for spectral clustering methods.

This paper is supported by National Key Basic Research and Development Program 2002CB312101.

2. SPECTRAL CLUSTERING

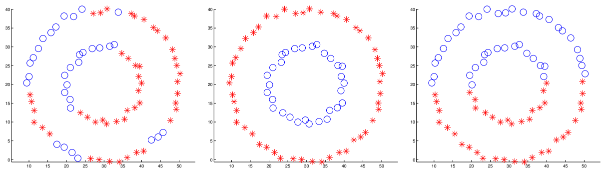
2.1. Principle

Spectral clustering is a method that groups data points with eigenvectors of pairwise affinity matrix of the data set. The method consists of three steps. Firstly, the pairwise affinity matrix A is calculated. Secondly, the top k (manually decided cluster number) eigenvectors of either A [10], or Laplacian matrix of A [11][9], is calculated. Data points are thus embedded into corresponding k-dimensional eigenspace. The final step is to cluster the data in the new space using a simple method such as K-means. We quote Ng et.al's algorithm [9] here for detailed analysis.

Algorithm 1 (Ng)
1. *Calculate distance matrix D and affinity matrix A: $A_{ij} = h(D_{ij}, \sigma)$ where $h(\cdot, \sigma)$ is the kernel function with parameter σ, say $h(d) = exp\left(-(d^2)/(2\sigma^2)\right)$.*
2. *Let $T = diag(A \cdot 1)$ be a diagonal matrix of row-sums of A and $L = T^{-1/2}AT^{-1/2}$ the Laplacian of A. Calculate top k eigenvectors $V \in \mathbf{R}^{n \times k}$ and corresponding eigenvalues $\{\lambda\}_k$ of L. Generate matrix Y by normalizing the rows of V.*
3. *Cluster the n row vectors of Y using K-means.*

2.2. Selection of Kernel Parameter σ

Besides the cluster number k, there is another unspecified parameter σ in Algorithm 1. It seems that most of the literature regarded the selection of σ as a trivial practical problem. Some of the literature gave simple formulas such as $\sigma = std\{elements\, of\, D\}$ [11], or (for data set in Euclidean metric space) $\sigma = \sqrt{nc}$, where n equals the dimension of input space, c equals the average of the data's variance in each dimension [8]. Others simply gave numeric



(a) (b) (c)

Fig. 1. Different results of spectral clustering, in the cases of different kernel parameter σ-s: (a)small σ, (b)proper σ, (c)large σ

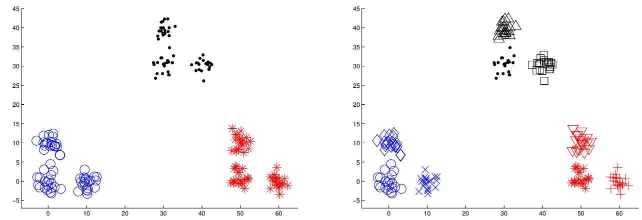
values or even ignored it. However, the following experiment and sequential analysis show that the choice of σ is important not only practically but theoretically.

Figure 1 shows three different clustering results of a data set which take the shape of two homocentric rings. Distance $d(\cdot)$ is set to be Euclidean metric, $k = 2$, $\sigma = coef \cdot std\{elements\, of\, D\}$, with $coef = 0.18, 2.87, 7.17$ in Figure 1(a),1(b),1(c) respectively. The reader can see that the results are remarkably different. Such a phenomenon can be explained as follows:

The σ in Figure 1(a) is so small that the affinity between any two points (even for the nearest pair) are ≈ 0. Since the entire graph is almost an empty graph, the costs (or cut, as commonly used in graph theory) of arbitrary bipartitions will all be approximately equal. This explains why such a strange and unstable partition appears. On the other side, the σ in Figure 1(c) is so large that the affinity between any two points (even for some pairs linked by *bridges* over two rings) are ≈ 1. Since separating two neighboring points is not much more expensive than separating a short bridge, it is not strange the linear partition in Figure 1(c), rather than the partition in Figure 1(b), minimizes the normalized cut [11]. This is because the partition in Figure 1(c) separates a few neighbors, while the partition in Figure 1(b) separates a bunch of short bridges. Of course, results in Figure 1(c) is also unstable due to the randomness of partition line direction.

Here we have used an intuitive term: *stable*. To be more precise, we say a spectral clustering result is highly stable, if a relatively big perturbation of the affinity matrix will not generate a different result. We learn from experiments that the *stability* of clustering should be a key criteria of the quality of clustering result, including cluster number and partitioning result. An *obvious* (for human, of course) cluster number with an inappropriate σ may produce fairly unstable result.

Moreover, we find that changing σ is equivalent to multiplying D by corresponding factor. Intuitively, that means changing the distance between your eyes and the plane where the points are located. It's natural that one is able to see different structures or cluster numbers from different distances.



(a) (b)

Fig. 2. Two stable clustering results with automatically determined cluster number k=3(a), 9(b) respectively

We claim that there is generally not a *single* cluster number. Instead, a family of *intrinsic cluster numbers* may exist with the change of kernel parameter, or visually, with the change of viewing distance. This family is what we call in this paper *intrinsic cluster numbers family*. The criteria for selecting the family is based on clustering stability.

2.3. Clustering Stability

As we know, spectral clustering methods embed data into a k-dimensional eigenspace, which is expanded by top k eigenvectors. One can show that if two different sets of k eigenvectors (induced from two different affinity matrices) expand the same eigenspace, the clustering results will be identical. Thus we can transform stability of discrete clustering result into stability of continuous subspace. We describe the latter mathematically in the following:

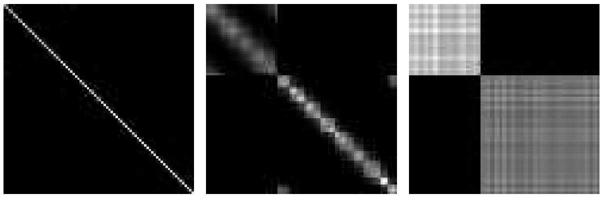
Denote $EV(A,k) = [v_1, v_2, \cdots, v_k]$ the matrix of top k eigenvectors of matrix A, and $ES(A,k) = Rang(EV(A,k))$ the eigenspace expanded by $EV(A,k)$. Given $U = EV(A,k)$, $U_e = EV(A+E,k)$, where E is an additional perturbation of A. Set eigenspace $X = ES(A,k)$ and its perturbation $X_e = ES(A+E,k)$. The different between eigenspace and its perturbation, can be measured, according to *matrix perturbation theory* [12], as

$$\|\sin\Theta(X, X_e)\|_2 = \|\sin(arccos[(U^T U_e U_e^T U)^{1/2}])\|_2.$$

Matrix perturbation theory provide some upper limits on eigenspace perturbation, in terms of eigenvalues. Most of them have similar forms, such as: $\|\sin\Theta(X, X_e)\|_2 < \frac{f(A,E)}{\lambda_k - \lambda_{k+1}}$. That means a large eigengap $(\lambda_k - \lambda_{k+1})$ indicates a relatively stable eigenspace, thus a stable cluster number k and corresponding k-way clustering results.

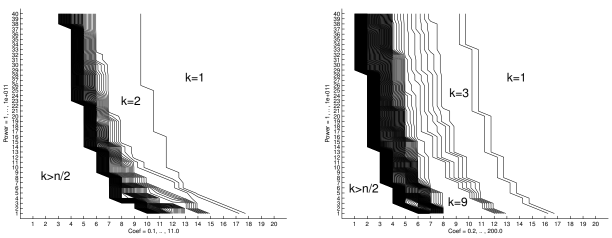
A straightforward method to discover intrinsic cluster numbers is to find for each σ a single optimal k (noted by k_{opt}) by maximizing $(\lambda_k - \lambda_{k+1})$, then drawing a graph of $k_{opt} - \sigma$ and finding out continuously appearing k_{opt}s.

This method showed perfect results for data like Figure2, but couldn't be worse for data in Figure 1(not a single σ pro-



(a) (b) (c)

Fig. 3. Random walk's effect on affinity matrix of data in Fig. 1. Points on the inside ring are placed before outside, for a clearer visualization. (a)Original (normalized) affinity matrix. (b)Affinity matrix after 100-step random walk. Intra-ring affinities are higher than (a)(check the left-top block and right-bottom block). (c)Affinity matrix after 10000-step random walk. It's now much easier to see the cluster number.



(a) (b)

Fig. 4. k-contour of data in (a)Fig. 1 and (b)Fig. 2. *Kernel parameter* σ varies logarithmically along x-axis, *step of random walk* increases logarithmically along y-axis. White areas indicate stable cluster numbers

duces $k_{opt} = 2$). The reason is that points in the same ring form a manifold and show more *transitive similarity* than *direct similarity*, i.e. points in the same ring may not be near to each other, but are linked to each other through at least one compact path. Unlike ISOMAP or LLE, which explicitly discover the lower dimensional structure, the method we used here is to perform *random walk on graph*. This method can take advantage of spectral clustering, and model the manifold data as well as clustering data much faster than ISOMAP or LLE.

2.4. Random Walk-enhanced Affinity

Given a weighted complete graph $G(V, V \times V, A)$, we can induce a Markov model $M(V, \pi_0, P = T^{-1}A)$, where $T = diag(A \cdot 1)$ and π_0 is not important here. Meila et.al [7] stated that minimizing normalized cut of G is equivalent to finding the (right) eigenvectors of transition probability matrix P.

Define the n-step random walk on graph G as another Markov model $M^t(V, \pi_0, P^n)$. We see in Figure 3 that although some pairs belonging to the same ring might be far from each other (Figure3(a)), their n-step transition probability may be significantly > 0 (Figure 3).

We find that it is not necessary to calculate P^n If $Px = \lambda x$, we have $P^n x = \lambda^n x$. This fact avoid power operation on matrix P because we care for $\|\lambda\|$ only. We then propose the following algorithm for automatic selection k_{opt} family.

Algorithm 2 *1. Set $\tilde{K}$ larger than any possible cluster number. Calculate distance matrix D. Generate a geometric sequence of $\{\sigma_i\}$ and a geometric sequence of $\{pow_j\}$, $pow_0 = 1$.*

2. For each σ_i, Calculate $A_{ij}^{(\sigma_i)} = h(D_{ij}, \sigma_i)$. $P^{(\sigma_i)} = (T^{(\sigma_i)})^{-1} A^{(\sigma_i)}$, where $T^{(\sigma_i)} = diag(A^{(\sigma_i)} \cdot 1)$. Calculate top $\tilde{K}$ eigenvectors $V(\sigma_i)$ and corresponding eifenvalues $\lambda_t(\sigma_i, pow_0), t = 1, 2, \cdots, \tilde{K}$.

3. Obtain $\lambda_t(\sigma_i, pow_j) = |\lambda_t(\sigma_i, pow_0)|^{pow_j}$, $k_{opt}(\sigma_i, pow_j) = \arg\max_k(\lambda_k - \lambda_{k+1})$.

4. Select a family of cluster number $\{k_{opt}^{(i)}\}$ which continuously appear in the $k_{opt} - power - \sigma$ graph, and corresponding family of $\{\sigma^{(i)}\}$ and $\{pow^{(i)}\}$ s.t. $k_{opt}(\sigma^{(i)}, pow^{(i)}) = k_{opt}^{(i)}$.

5. Cluster the data with parameter triples $(k_{opt}^{(i)}, \sigma^{(i)}, pow^{(i)})$. Build a dendrogram with the clustering results.

Figure 4(b) shows the contour lines of $k_{opt} - power - \sigma$ graph wrt. data in Figure 2. We can find 4 consistent areas(bands) that are marked with k_{opt}'s value. Figure 2 are the clustering results given the parameter triples $(k_{opt}^{(i)}, \sigma^{(i)}, pow^{(i)})$ in Area $k = 3$ and $k = 9$ respectively. Figure 4(a) shows k-contour of data in Figure 1. Only one non-trivial and significant band ($k = 2$) exists. It is interesting to note that Area $k = 2$ disappears at the bottom of the graph, i.e. there is not any $k = 2$ appearing without random walk. That's why we couldn't find the most reasonable cluster number(k=2) without a new dimension random walk provided.

3. APPLICATION ON IMAGE SEGMENTATION

We also apply the above algorithm on natural image segmentation. Since computational complexity is the bottleneck of spectral methods for large database, Nyström approximation was introduced to spectral algorithm [1]. What we do is to introduc random walk into Belongie et.al [1]'s algorithm to generate segment numbers and segmentations automatically. Detailed algorithm and theorems supporting the algorithm are beyond the scope of this paper.

The data we used are provided by Martin et.al [6]. Color images are quantized to 8-color indexed images with 75%

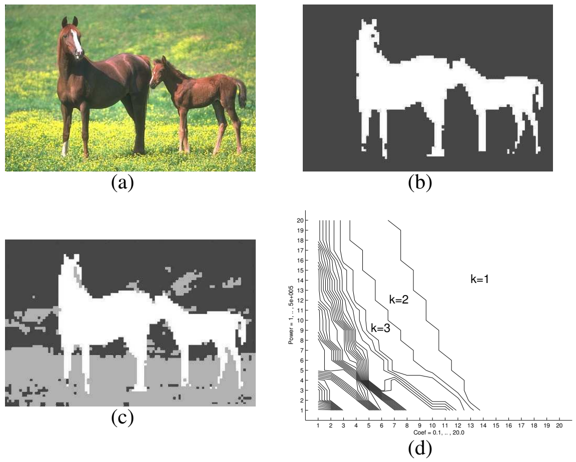


Fig. 5. Results on horses image. (a)original image, (b)result for k=2,(c)result for k=3, (d)k-contour.

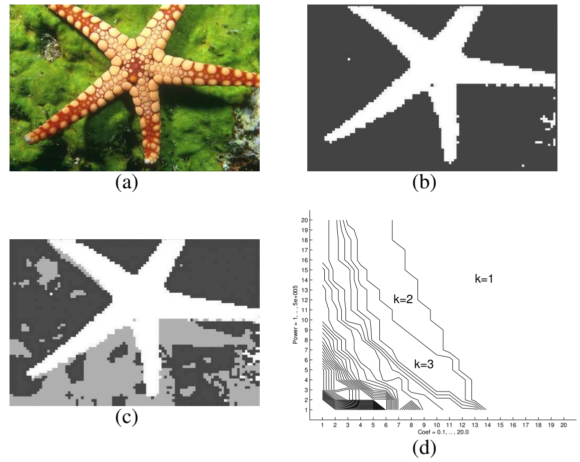


Fig. 6. Results on starfish image. (a)original image, (b)result for k=2,(c)result for k=3, (d)k-contour.

dithering using Photoshop, and 7*7 neighborhood 8-bin histogram is used as pixel feature. The kernel function is defined by Gaussian χ^2 similarity [1]. The results are shown in Figure 5,6. The reader can find that large white areas in k-contour, marked by k's value, produce reasonable and stable results.

4. CONCLUSION

We show in this paper that, under the criteria of clustering stability, there is a close correlation between two sensitive factors, kernel parameter σ and cluster number k. In order to reveal the correlation clearer, we use the technique of *random walk on graph*. The combination of random walk and choice of σ provide a two-dimension view to explore intrinsic data structure. Since one is able to see different structures in different parameter configurations, we introduce the concept of *intrinsic cluster number family*. And we build an automatic algorithm achieving simultaneous generation of the family and clustering dendrogram. The automatic algorithm shows reasonable clustering hierarchies in both simulated data and real image segmentation.

5. REFERENCES

[1] Belongie, S., Fowlkes, C., Chung, F., Malik, J.: *Spectral Partitioning with Indefinite Kernels Using the Nyström Extension*, Proc. ECCV 2002.

[2] Bischof H, Leonardis A, Sleb A.: MDL principle for robust vector quantization. Pattern Analysis and Applications, 1999.

[3] Fraley, C., Raftery, A.E.: *How Many Clusters? Which Clustering Method? Answers via Model-based Cluster Analysis*, Computer journal, 1998.

[4] Figueiredo, M.A.T and Jain, A.K.: Unsupervised learning of finite mixture models, IEEE Trans. PAMI, 2002.

[5] Kamvar, S.D., Klein, D., and Manning, C.D.: *Spectral Learning*, Proc. of the 18th IJCAI, August 2003

[6] Martin, D., Fowlkes, C., Tal, D., Malik, J.: *A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics*, Proc. ICCV 2001.

[7] Meila M., Shi, J.: *A Random Walks View of Spectral Segmentation*, Proc. International Workshop on AI and Statistics(AISTATS), 2001.

[8] Mika, S., et.al., *Kernel PCA and De–Noising in Feature Spaces*, Advances in Neural Information Processing Systems 11, 1999.

[9] Ng, A.Y., Jordan, M.I., Weiss, Y.: *On Spectral Clustering: Analysis and an Algorithm*, Advances in Neural Information Processing Systems 14, 2002.

[10] Scott, G. and Longuet-Higgins, H.: Feature Grouping by relocalisation of eigenvectores of proximity matrix, In Proc. British Machine Vision Conference, 1990.

[11] Shi, J., Malik, J.: *Normalized Cuts and Image Segmentation*, IEEE Trans. PAMI, 2000.

[12] Sun, J.G.: *Matrix Perturbation Analysis(2nd Ed.)*, Science Press, Beijing, China, 2001.