Chinese Journal of Electronics
Vol.18, No.1, Jan. 2009

An Optimal Spectral Clustering Approach Based on Cauchy-Schwarz Divergence*

XU Haixia[1] and TIAN Zheng[2,3]

(1.*School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China*)

(2.*School of Science, Northwestern Ploytechnical University, Xi'an 710072, China*)

(3.*State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing 100101, China*)

Abstract — A new global criterion, the Cauchy Schwarz (CS) cut, for spectral clustering is presented based on the Cauchy-Schwarz divergence. It is proved that when the sum of intra-cluster and inter-cluster similarity is fixed, optimizing the CS cut criterion can ensure intra-cluster similarity maximized and inter-cluster similarity minimized simultaneously. An efficient computational technique is developed based on eigenvalue problem. Experimental results on artificial data sets and natural images show that the proposed approach is very encouraging.

Key words — Spectral clustering, Graph, Cauchy-Schwarz divergence.

I. Introduction

Clustering is one of the most widely used techniques for exploratory data analysis, with applications ranging from statistics, computer science, to social sciences[1,2]. The objective of clustering is to partition a given data set into non-empty subsets such that similar data are grouped together and data in different subsets are dissimilar. Traditional clustering algorithms such as K-means or EM-based algorithms tend to be computationally efficient since they only require one to compare the data points to a small set of cluster prototypes. However, they suffer from many drawbacks. First, Gaussian cluster distributions are implicitly assumed. It minimizes a sum-of-squares cost function, equivalent to variance minimization, and thus fails if the cluster distributions are not spherical. Second, convergence is not guaranteed and the clustering result is sensitive to the initialization. Finally, the rate of data changing class between iterations may be relatively high, even after several iterations.

Recently, a promising alternative that has emerged is known as the spectral clustering which inspired by the concept of graph partitioning. Its ability to identify non convex clusters makes it ideal for a number of applications, including computer vision and pattern recognition[3,4]. As pointed in Refs.[5] and [6], spectral clustering has the advantages that parametric assumptions about data distributions do not have to be made and allows propagation of similarity from point to point. By the flexibility in the definition of similarity of the data points, this method can recovery of clusters that take on more complicated manifold structures.

A graph can be bi-partitioned into two disjoint subsets simply by removing edges between the two parts, one measure of similarity between these two pieces can be computed as the total weight of the edges that have been removed. This quantity is called the cut. The minimum cut criterion by Wu and Leahy[7] has been used in spectral clustering in Ref.[8]. However, the cut cost favors cutting small sets of isolated vertices in the graph. To compensate for this fact, a number of rather heuristically motivated improvements to the cut cost have been proposed, such as the normalized cut by Shi and Malik[9], the min-max cut due to Ding *et al.*[10] and Hagen's ratio cut[11]. Since these cut criteria normalize the cut cost by dividing it by either the cardinal numbers of the clusters or a measure derived from the similarity between vertices within the clusters, optimizing these criteria can indeed prevent skew cut efficiently. However, it is also found that these criteria cannot always ensure the high intra-cluster similarity especially when the overlap between clusters is large. In this paper, we propose a new cut criterion based on the Cauchy-Schwarz divergence. It is important to stress that clusters obtained by optimizing this new cut criterion have high intra-cluster similarity and low inter-cluster similarity simultaneously.

This paper is organized as follows. In Section II, we propose the CS cut criterion. A spectral clustering solution is justified to optimize the new criterion in Section III. In Section IV, we describe the clustering algorithm for data sets and further develop it to accommodate image segmentation based on the Nyström method. Section V contains the experimental results with conclusion in Section VI.

*Manuscript Received Oct. 2007; Accepted June 2008. This work is supported by the National Natural Science Foundation of China (No.60375003), the Aeronautics and Astronautics Basal Science Foundation of China (No.03I53059) and the Science and Technology Innovation Foundation of Northwestern Polytechnical University (No.2007KJ01033).

II. Cauchy Schwarz Cut

Given a set of n data points, we constructed a weighted undirected graph $G(V,E)$, where the vertex set V is the data set and the edge set E is the set $V^{(2)}$ of unordered pairs of V. The weight on each edge, $w(i,j)$, representing the similarity between every pair of vertices and the weight matrix is defined as $W=(w(i,j))_{n\times n}$. If G can be partitioned into two disjoint subsets, $A\neq\emptyset$, $B\neq\emptyset$, $A\cup B=V$, $A\cap B=\emptyset$, then the cut is $cut(A,B)=\sum_{i\in A,j\in B}w(i,j)$, which can be viewed as the inter-cluster similarity. In the same way, the intra-cluster similarity can be defined as the sum of all edge weights within a cluster, that is $assoc(A,A)=\sum_{i,j\in A}w(i,j)$ and $assoc(B,B)=\sum_{i,j\in B}w(i,j)$. Although the optimal bipartition of a graph is the one that minimizes the cut value, in clustering, we wish that not only the cut value is minimized, but also the intra-cluster similarity is maximized. Since the sum of $cut(A,B)$, $assoc(A,A)$ and $assoc(B,B)$ is always $cut(V,V)$, this implies that a good clustering must satisfy the following requirement

$$\begin{cases}\min\ cut\,(A,B),\max\ assoc(A,A)\ \text{and}\ \max\ assoc\ (B,B)\\ s.t.cut(A,B)+assoc\,(A,A)+assoc(B,B)=cut(V,V)\end{cases}$$

Recall that W is a symmetry matrix, so all of its eigenvalues are real, there exists an orthonormal basis eigenvectors $(\phi_1,\cdots,\phi_n)$ satisfies

$$\Phi^TW\Phi=\Lambda \tag{1}$$

where the columns of Φ contain the orthonormal eigenvectors, and $\Lambda=diag(\lambda_1,\lambda_2,\cdots,\lambda_n)$ is a diagonal matrix that contains the corresponding eigenvalues. Define the matrix $C=\Phi\Lambda^{\frac{1}{2}}\Phi^T=C^T$, and let $u=Cx$, $v=Cy$, based on the Cauchy-Schwarz inequality, $(u^Tv)^2\leq u^Tu\cdot v^Tv$, we can obtain the following expression

$$(x^TWy)^2\leq x^TWx\cdot y^TWy \tag{2}$$

Now, we define the CS cut as

$$\begin{aligned}\mathrm{CS}\,cut(A,B)=&\frac{(x^TWy)^2}{x^TWx\cdot y^TWy}\\ =&\frac{(cut(A,B))^2}{assoc(A,A)assoc(B,B)}\end{aligned} \tag{3}$$

where x is the indicator vector of subset A, such that $x_i=1$ if vertex i is in A and $x_i=0$ otherwise, y is defined for B analogously.

III. Optimizing the Cauchy Schwarz Cut

Theorem 1 Minimizing the CS cut is NP-hard.

Proof According to Papadimitrou's method[9], minimizing Eq.(3) on grid graph is NP-hard. Hence minimizing the CS cut is NP-hard.

Although minimizing the CS cut is NP-hard, we will show that, when we embed the cut problem in the real value domain, an approximate discrete solution can be found efficiently.

Theorem 2 If $d_i=\sum_{j=1}^n w(i,j)$ is the total connection from vertex i to all other vertices, D is a $n\times n$ diagonal matrix with $d_1,d_2,\cdots,d_n$ on its diagonal, and $z^{(2)}$ is an eigenvector corresponding to the second smallest eigenvalue of normalized Laplacian matrix $L=D^{-1/2}(D-W)D^{-1/2}$, then $(D^{-1/2}z^{(2)})^2$ is the real valued solution to minimize CS cut, here the square is with respect to the element of vector.

Proof Let m be a n dimensional indicator vector such that $m_i=1$ if vertex i is in A and -1, if it is in B. With the definition of m, Eq.(3) can be rewritten as

$$\begin{aligned}\mathrm{CS}\,cut(A,B)=&\frac{(1+m)^TW(1-m)}{(1+m)^TW(1+m)}\cdot\frac{(1+m)^TW(1-m)}{(1-m)^TW(1-m)}\\ =&\frac{(1+m)^T(D-W)(1+m)}{(1+m)^TW(1+m)}\\ &\cdot\frac{(1-m)^T(D-W)(1-m)}{(1-m)^TW(1-m)}\end{aligned} \tag{4}$$

Thus, Minimizing the CS cut is equal to minimize $\frac{(1+m)^T(D-W)(1+m)}{(1+m)^TW(1+m)}$ and $\frac{(1-m)^T(D-W)(1-m)}{(1-m)^TW(1-m)}$ simultaneously. Fortunately, the two optimization problems can be achieved under the same condition. Now, we consider $\frac{(1+m)^T(D-W)(1+m)}{(1+m)^TW(1+m)}$ only. According to the theorem 2 of Ref.[12], this can be minimized by solving the eigenvalue problem

$$Lz=\lambda z \tag{5}$$

where $z=D^{1/2}m$. Recall a simple fact about the Rayleigh quotient[13], $D^{-1/2}z^{(2)}$ is the real valued solution to $\min\frac{(1+m)^T(D-W)(1+m)}{(1+m)^TW(1+m)}$, hence $m=(D^{-1/2}z^{(2)})^2$ is the real valued solution to minimize CS cut.

IV. The Spectral Clustering Algorithm

Our spectral clustering algorithm consists of the following steps:

1. Given a set of data points, set up a weighted undirected graph $G(V,E)$ and set the weight on the edge connecting two vertices to be a measure of the similarity between them;
2. Solve for eigenvector $z^{(2)}$ corresponding to the second smallest eigenvalue of the matrix L, then let $m=(D^{-1/2}z^{(2)})^2$;
3. Bi-partite the graph G based on m;
4. Decide if the current partition should be subdivided and recursively repartition the sub-graph if necessary.

The above steps are designed for clustering sets of data points. Due to computational cost, we cannon spectrally cluster all pixels in an image. So we use random sampling to extract a sample $\{I_1,\cdots,I_N\}$ of size N. If the remaining out-of-sample pixels can be mapped into eigenspace, the complete image can be classified using the above algorithm. This is achived by an approximation known as Nyström method[5], which is a technique to find numerical approximations for eigenvalue problems. Here we only need the eigenvector $z^{(2)}$ correspond to the second smallest eigenvalue λ_2, so for an arbitrary out-of-sample pixel I_t

$$z_t^{(2)}\approx\frac{1}{N\lambda_2}\sum_{i=1}^N z_i^{(2)}w(I_t,I_i) \tag{6}$$

V. Experiments and Comparisons

In this section, we present some experimental results on artificial spatial point sets and images from Berkeley image dataset (http://www.cs.berkeley.edu/projects/vision/grouping /segbench/) to demonstrate the effectiveness of our approach. We compare the results of the CS cut criterion to them obtained by normalized cut algorithm (http://www.cis.upenn.edu/∼!jshi/software/). For all experiments, we use an exponential weight function of the form

$$w(i,j) = \exp\left\{ -\frac{\|X(i)-X(j)\|_2}{\sigma_1} - \frac{\|F(i)-F(j)\|_2}{\sigma_2} \right\} \quad (7)$$

where $X(i)$ is the spatial location of vertex i, $F(i)$ is a feature vector based on intensity or color at that vertex defined as

$F(i) = 1$, in the case of clustering point sets;

$F(i) = I(i)$, the intensity value, for segmenting brightness images;

$F(i) = [R(i), G(i), B(i)]$, the color value, for segmenting color images.

where the scaling parameter σ_1 and σ_2 control how rapidly the weight w falls off with the spatial distance and feature distance between two vertices.

Fig.1 shows two point sets and the clustering results of CS cut. These point sets are good examples to illustrate the novel features of the proposed method, as they consist of non convex clusters which cannot be separated directly using K-means or similar clustering algorithms. (a) and (c) show the original point sets, where (a) consists of four groups: a circular ring with a cloud of points inside and two clouds outside, and (c) consists three circles with the outer two jointed. (b) and (d) are the associated clustering results of CS cut. It shows that our algorithm is indeed able to partition the point sets in a desirable way.

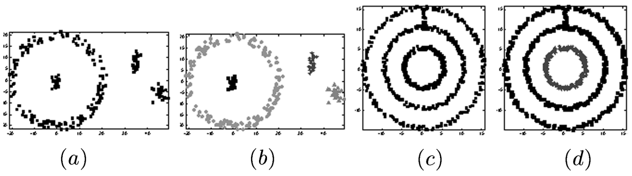

Fig. 1. (a) and (c) are the original spatial point sets; (b) and (d) are the clustering results of CS cut

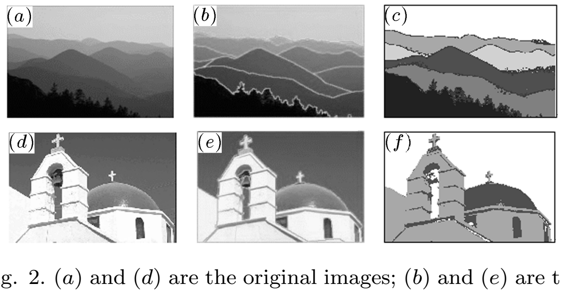

Fig. 2. (a) and (d) are the original images; (b) and (e) are the human segmentations; (c) and (f) are the segmentation results of CS cut

Fig.2 shows two images (one is brightness image and the other is color image) and the segmenting results of CS cut. (a) and (d) are the original images. (b) and (e) are the human segmentations from Berkeley image dataset. (c) and (f) are the segmentation results of CS cut. We can see that even with a simple feature extraction, our algorithm manages to produce somewhat sensible segmentations.

Fig.3 and Fig.4 show the comparisons of our approach with the normalized cut algorithm. Though the two algorithms use the eigenvector corresponding to the second smallest eigenvalue of normalized Laplacian matrix to group the point set, our algorithm does better. Fig.3(a) and Fig.4(a) are the original point set and color image. Fig.3(b) and Fig.4(b) are the grouping results based on the normalized cut criterion. Fig.3(c) and Fig.4(c) are the grouping results based on our criterion. The grouping results show that the algorithms based on the normalized cut criterion has trouble deciding on where to cut, while our algorithm performs well in the two cases and produces a clear solution.

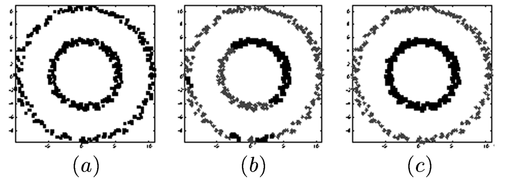

Fig. 3. (a) The point set; (b) The clustering result of normalized cut; (c) The clustering result of CS cut

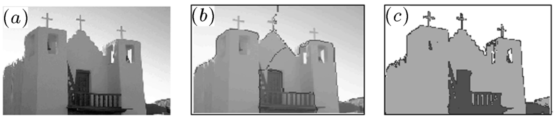

Fig. 4. (a) The original image; (b) The segmentation result of normalized cut; (c) The segmentation result of CS cut

VI. Conclusion

In this paper, the CS cut criterion for spectral clustering is proposed based on Cauchy-Schwarz divergence. Clusters obtained by optimizing the new criterion have high intra-cluster similarity and low inter-cluster similarity simultaneously. Though optimization of the CS cut is NP-hard, an efficient computational technique based on eigenvalue problem is developed. Experiments on spatial data sets, brightness and color images illustrate that the CS cut can get sensible clustering and segmentation results with simple features such as spatial location, intensity or color.

References

[1] K. Jain, R.C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, USA, 1988.

[2] R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York, USA, 1973.

[3] C. Alpert, A. Kahng, S. Yao, "Spectral partitioning with multiple eigenvectors", *Discrete Applied Mathematics*, Vol.90, No.1-3, pp.3–26, 1999.

[4] Y. Weiss, "Segmentation using eigenvectors: A unifying view", *Proc. of Int. Conf. on Computer Vision*, Corfu, Greece,

pp.975–982, 1999.

[5] C. Fowlkes, S. Belngie, F. Chung, J. Malik, "Spectral grouping using the Nystr?m method", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.26, No.2, pp.214–225, 2004.

[6] A.Y. Ng, M.I. Jordan, Y. Weiss, "On spectral clustering: Analysis and an algorithm", *Proc. of Adv. in Neural Info. Processing Systems*, Vancouver, Canada, pp.849–856, 2001.

[7] Z. Wu, R. Leahy, "An optimal graph theoretic approach to data clustering: theory and its application to image segmentation", *IEEE Transaction on Pattern Analysis and Machine Intelligence*, Vol.15, No.11, pp.1101–1113, 1993.

[8] N. Cristianini, J. Shawetaylor, J. Kandola, "Spectral Kernel Methods for Clustering", *Proc. of Adv. in Neural Info. Processing Systems*, Vancouver, Canada, pp.649–655, 2001.

[9] J. Shi, J. Malik, "Normalized cuts and image segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.22, No.8, pp.888–905, 2000.

[10] C. Ding, X. He, H. Zha, M. Gu, H. Simon, "A min-max cut algorithm for graph partitioning and data clustering", *Proc. of IEEE International Conference on Data Mining*, San Jose, California, USA, pp.107–114, 2001.

[11] L. Hagen, A.B. Kahng, "New spectral methods for ratio cut partitioning and clustering", *IEEE Trans. on Computed Aided Des.*, Vol.11, No.9, pp.1074–1085, 1992.

[12] X. Li, Z. Tian, "Optimum cut-based clustering", *Signal Processing*, Vol.87, No.11, pp.2491–2502, 2007.

[13] G.H. Golub, C.F. Van Loan, *Matrix Computations*, John Hopkings Press, Baltimore, 1989.

XU Haixia was born in Tianjin, China in 1980. She received B.S. and M.S. degrees from the Department of Applied Mathematics, Northwestern Polytechnical University (NWPU), in 2003 and 2006, respectively. From 2005, she has been pursuing the Ph.D. degree at the Department of Computer Science and Engineering of NWPU. Her research interests include image processing and pattern recognition.

TIAN Zheng received M.S. degree from the Department of Applied Mathematics, Northwestern Polytechnical University (NWPU), in 1985. During 1993–1994, she was invited to do the Postdoctoral research work at the Department of Statistics, University of Dortmund, Germany. In 2002, she was invited to do cooperation research work by the Department of Statistics Actuary Science, University of Hong Kong. She is currently a Professor and Ph.D. supervisor of the Department of Applied Mathematics and the Department of Computer Science and Engineering of NWPU. Her research interests include nonlinear time series, multiscale random models, spectral graph theory, and their applications in image processing.