Statistical Test for Similarity Metrics and Clustering

Sayan Mukherjee

Artificial Intelligence Laboratory and
The Center for Biological and Computational Learning
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: Many algorithms for clustering and judging similarity metrics are being and have been developed [2, 5, 6]. It is important to be able to judge if these clusters are in some sense real or could have likely occurred by chance. We would like to develop a statistical test to determine based upon a similarity matrix a) whether there really are two clusters or different distributions generating the data b) if the data was labeled, whether a similarity metric is picking up structure or not. The test will be based upon the eigenvalues of the similarity matrix.

Given examples $x_1, ..., x_\ell$ we can construct a similarity matrix K where for example $K_{ij} =< x_i, x_j >$. One can then assign putative labels $y_1, ..., y_\ell$ by trying to maximize the following functional

$$\max_y \frac{y^T K y}{\sqrt{||y^T y||_F ||K||_F}}, \tag{1}$$

where $|| \cdot ||_F$ is the Frobenius norm. This functional is related to the cut-cost functional which is often used in clustering

$$\min_y \frac{\sum_{y_i \neq y_j} K_{ij}}{\sqrt{||y^T y||_F ||K||_F}}. \tag{2}$$

The optimum of both functionals can be upper bounded by an eigenvalue problem on the kernel matrix K [1].

Motivation: In analyzing DNA microarray data one often wants to discover morphological structure in the data or one might want to decide that one set of features are more relevant than another in extracting information from the data. For either task one can question whether the information extracted is real or due to inherent randomness in the data. Again we seek to formulate a statistical test to address this question. We also seek to formulate this approach of correlating labels with a kernel matrix in the regularization framework.

Previous Work: Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests based upon minimal spanning trees were introduced in the late seventies by Freidman and Rafsky [4].

Currently segmentation and clustering based upon spectral methods have been used by Kannan and Vempala [5], Cristianini et al [2], and Meila and Shi [6].

Edelman has extensively studied random random matrices [3].

Approach: We would like to show for a wide variety of distributions the eigenvalues of the kernel or similarity matrix for data drawn from a single distribution is very stereotypical. We will the structure of these eigenvalues for a statistical test that tests against the null hypothesis that the data comes from two distribution to either score a similarity metric or perform clustering. We will also use a regularization argument based upon correlation structures of a function to motivate the use of the kernel or similarity matrix [7].

Impact: A general statistical test to measure the "validity" of a similarity metric or cluster assignment will be of great use in gene selection for DNA microarray data as well as unsupervised taxonomy exploration.

Research Support: This report describes research done within the Center for Biological & Computational Learning in the Department of Brain & Cognitive Sciences and in the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology.

This research was sponsored by grants from: Office of Naval Research (DARPA) under contract No. N00014-00-1-0907, National Science Foundation (ITR) under contract No. IIS-0085836, National Science Foundation (KDI) under contract No. DMS-9872936, and National Science Foundation under contract No. IIS-9800032

Additional support was provided by: Central Research Institute of Electric Power Industry, Center for e-Business (MIT), Eastman Kodak Company, DaimlerChrysler AG, Compaq, Honda R&D Co., Ltd., Komatsu Ltd., Merrill-Lynch, NEC Fund, Nippon Telegraph & Telephone, Siemens Corporate Research, Inc., Toyota Motor Corporation and The Whitaker Foundation.

References:

[1] F.R.K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[2] N. Cristianini, J. Shawe-Taylor, and J. Kandola. Spectral kernel methods for clustering. In *submited to Neural Information Processing Systems 2001*, 2001.

[3] A. Edelman. The probability that a random real gaussian matrix has k real eigenvalues, related distributions, and the circular law. *Multivariate Analysis*, 60:203–232, 1997.

[4] J. Freidman and L. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *Annals of Statistics*, 7:697–717, 1979.

[5] R. Kannan, S. Vempala, and A. Vetta. In clusterings: Good, bad, and sepctral. In *Proceedings of the 41st Foundations of Computer Science*, 2000.

[6] M. Meila and J. Shi. A random walks view of spectral segmentation. In *submited to Neural Information Processing Systems 2001*, 2001.

[7] T. Poggio and F. Girosi. A sparse representation for function approximation. *Neural Computation*, 10:1445–1454, 1998.