Similarity in Computational Sciences

Tina Eliassi-Rad and Terence Critchlow
Center for Applied Scientific Computing, Lawrence Livermore National Laboratory
P.O. Box 808, L-560, Livermore, CA 94550, *{eliassi, critchlow}@llnl.gov*

The advent of fast computer systems has enabled scientists to visualize and analyze complex phenomena (such as explosions of stars and expressions of genes) [2][3][7][8][9][10][12][13]. These complex phenomena (whether simulated or observed) generate large-scale data sets. For instance, simulations of supernovae easily produce terabytes of data [1]. Given such massive amounts of data, it is not surprising that clustering algorithms are quite popular in computational sciences. In particular, *non-projected* clustering algorithms (such as *k*-means, *k*-mediods, hierarchical, and smooth clustering) are widely used [4][5][6]. A crucial input to any of these clustering algorithms is the similarity function that assigns data objects to specific groups. Depending on the purpose of clustering and the characteristics of data objects, the task of selecting an appropriate similarity function can be nontrivial. In computational sciences, data objects are represented by *n*-dimensional vectors in space and time [10][11]. In particular, each element of an *n*-dimensional data object can be either a scalar quantity (such as density) or a vector quantity (such as velocity). Due to the presence of both scalar and vector quantities within the data objects, it is important that a similarity function encodes both magnitude and direction. For instance, the popular Euclidean distance can only measure dissimilarities in scalar quantities and magnitude of vector quantities. That is, two vectors $\vec{\alpha}$ and $\vec{\beta}$ are considered identical when $\vec{\alpha} \equiv \vec{\beta}$ and $\angle\vec{\alpha} \neq \angle\vec{\beta}$! In contrast, the Pearson's correlation coefficient can only measure similarities in directions of vector quantities. That is, two vectors $\vec{\alpha}$ and $\vec{\beta}$ are considered identical when $\vec{\alpha} \neq \vec{\beta}$ and $\angle\vec{\alpha} \equiv \angle\vec{\beta}$! In both of these examples, a similarity function that captures both magnitude and direction will not consider $\vec{\alpha}$ and $\vec{\beta}$ as identical!

In the gene expression literature, the *uncentered correlation coefficient* has been singled out as a highly desirable metric for non-projected clustering of gene expression data [5][6][7]. In particular, Eisen *et al* [7] describe the uncentered correlation coefficient as a function that compares both magnitude and direction. In this presentation, we show that the uncentered correlation coefficient (as is commonly described) does not satisfy the aforementioned condition. Specifically, an arbitrary value for the offset parameter of an uncentered correlation coefficient can reduce the coefficient to merely measure similarities in direction (like the Cosine similarity function). In addition, we describe the properties and behaviors of different offsets and derive a quasi-optimal offset value for capturing both magnitude and direction of data objects.

For our empirical study, we input different similarity functions to non-projected clustering algorithms and build clusters on gene expression data and (simulated) supernovae data. We evaluate the clusters by using validity metrics that measure intra-cluster cohesion and inter-cluster separation. Our results illustrate the need for using similarity functions that capture both magnitude and direction in computational science data sets.

Topic: data mining
Preference: oral/poster

Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by the University of California Lawrence Livermore National Laboratory under contract No. W-7405-ENG-48.1. UCRL-ABS-209716.

References

[1] Abdulla, G., Critchlow, T., and Arrighi, W. (2004) Simulation Data as Data Streams, In *SIGMOD Record*, **33**(1).

[2] Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D., and Levine, A. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, In *PNAS U.S.A.*, **96**(12): 6745-6750.

[3] Baldwin, C., Abdulla, G., and Critchlow, T. (2003) Multi-resolution Modeling of large scale scientific simulation data, In *Proceedings of the 12th International Conference on Information and Knowledge Management*, ACM Press, 40-48.

[4] Bolshakova, N., Azuaje, F., and Cunningham, P. (2005) An integrated tool for microarray data clustering and cluster validity assessment, *Bioinformatics*, **21**(4): 451-455.

[5] Dadgostar, H., Zarnegar, B., Hoffmann, A., Qin, X.-F., Truong, U., Rao, G., Baltimore, D., and Cheng, G. (2002) Cooperation of multiple signaling pathways in CD40-regulated gene expression in B lymphocytes. In *PNAS U.S.A.*, **99**(3):1497-1502.

[6] De Hoon, M., Imoto, S., and Miyano, S., (2002) A comparison of clustering techniques for gene expression data. In *Proc. of the 10th Int'l Conf. on Intelligent Systems for Molecular Biology*.

[7] Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998) Cluster analysis and display of genome wide expression patterns. In *PNAS U.S.A.*, **95**(25):14863-14868.

[8] Eliassi-Rad, T., Critchlow, T., and Abdulla, G. (2002) Statistical modeling of large-scale simulation data, In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, 488-494.

[9] Freitag, L.A., and Loy, R.M. (1999) Adaptive, multiresolution visualization of large data sets using a distributed memory octree. In *Proceedings of the 1999 Supercomputing Conference*, ACM Press, Article 60.

[10] Heyer, L., Kruglyak, S., and Yooseph, S. (1999) Exploring Expression Data: Identification and Analysis of Coexpressed Genes, *Genome Research*, 9:1106-1115.

[11] Musick, R., and Critchlow, T. (1999) Practical lessons in supporting large-scale computational science, In *SIGMOD Record, 28*, 4.

[12] Slonim, D., Tamayo, P., Mesirov, J., Golub, T., and Lander, E. (2000) Class Prediction and Discovery Using Gene Expression Data, *RECOMB* 2000, 263-272.

[13] Wang, J., Wang, X., Lin, K.-I., Shasha, D., Shapiro, B.A., Zhang, K. (1999) Evaluating a class of distance-mapping algorithms for data mining and clustering, In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM Press, 307-311.