BY MOHITH MANJUNATH, PhD, AND YI ZHANG,
Carl R. Woese Institute for Genomic Biology,
University of Illinois at Urbana-Champaign

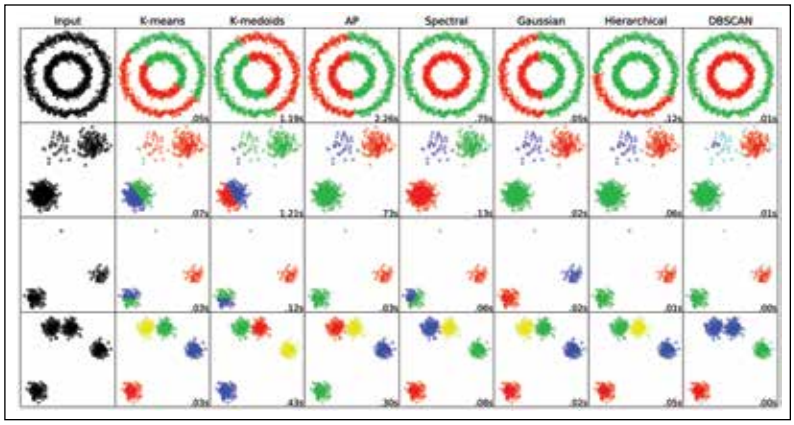
EXPLORING PATTERNS IN BIG DATA USING ClusterEnG, A CLUSTERING ENGINE FOR GENOMICS

In this age of genomic data deluge, researchers need integrated resources that can efficiently identify hidden structures in data. Researchers often use clustering, a popular machine-learning technique, to explore similarities within data. But they must choose from several clustering algorithms that may yield different results depending on input data and algorithm-specific metrics. Experimental biologists or even bioinformaticians may not be aware of the pros and cons of diverse clustering algorithms that vary in their complexity and ease of use; their ability to handle noisy data, outliers, or datapoints that aren't well separated; their computational expense; and how well they work on non-linear datasets.

ClusterEnG is a web resource that aims to address that problem. First, it provides a tutorial that defines and describes the pros and cons of various popular clustering algorithms: k-means, k-medoids, affinity propagation, spectral clustering, Gaussian mixture model, hierarchical clustering and DBSCAN. Second, it offers the opportunity for users to upload a file with numeric data in a tabular format and perform clustering analysis on that data. If a user is not sure which algorithm to choose, ClusterEnG offers the option of selecting several algorithms to explore the results. This gives researchers an idea of the structure of their data (see Figure).

ClusterEnG provides visualization of clustering results by performing principal component analysis, a commonly used dimensional reduction technique. The first three principal components of the input dataset are plotted two at a time for 2D visualization, while all three are also plotted together in 3D. The ability to pan, zoom and select datapoints in the 3D visualization is particularly helpful in revealing the hidden patterns in data.

The user can also explore the similarities and differences between various algorithms by selecting one of the sample datasets available on the ClusterEnG site. One option is the NCI60 gene expression dataset, which provides data for cell lines from 9 different types of cancer tissue of origin. Using this high-dimensional dataset for which the labels of the samples are known, users can see how some algorithms perform better at clustering different cancer types. For this dataset, for example, k-medoids and spectral clustering prove to be better at clustering like with like.



This table compares clustering results and run-times for six different algorithms (columns) applied across four input datasets (rows) with different structures (non-linear noisy circles, boxes with different densities, data with an outlier, and close boxes). Colors represent cluster labels. Spectral clustering and DBSCAN beat other methods when the dataset has circular structure and boxes with different densities (top two rows). For the second type of dataset, affinity propagation works better than others in most cases. The third dataset includes an outlier not far from the clusters, and Gaussian mixture model clustering does best at finding the outlier. The last row shows a dataset with four clusters, two of them are close to each other. All the algorithms do well, but the ways they partition the two close clusters are different. Image courtesy of the authors.

For high-dimensional datasets where the structure is unknown, the process of identifying the best clustering algorithm for the data may require some biological intuition about the data's structure as well as iterative trial and error—i.e., visualizing the principal component plots and experimenting with several algorithms until a meaningful pattern emerges.

Clustering can be a powerful way to explore data, but it is important to understand which methods to use and how to use them correctly. By allowing users to explore their data using multiple clustering algorithms, the ClusterEnG web resource provides much-needed assistance for biomedical researchers dealing with Big Data. □

DETAILS

Mohith Manjunath is a postdoctoral research associate and Yi Zhang is a graduate student in Jun Song's lab at the Carl R. Woese Institute for Genomic Biology at the University of Illinois at Urbana-Champaign. They are part of the development team for ClusterEnG, which was developed as part of the KnowEnG BD2K Center and can be accessed at http://education.knoweng.org/clustereng.