Application Examples for Cluster validity analysis platform (CVAP)

Kaijun WANG, Xidian University (sunice9@yahoo.com)

The quality assessment of clustering results is an important topic in cluster analysis. CVAP is a cluster validation tool to evaluate clustering quality, integrating validity indices and clustering algorithms. Here we illustrate the clustering process and validation process in CVAP through applying CVAP to cluster analysis of two gene expression datasets: one is the leukemia dataset with 72 tissue samples in three classes; another is the yeast data set with 208 genes in four clusters.

"mainCVAP" is used in Matlab command window to start CVAP, and the clustering process should be performed before cluster validation process, but no clustering process is needed for a solution file where class labels (or clustering solutions) are from a clustering algorithm not included in CVAP. We first use PAM to cluster the leukemia dataset (data file "leuk72_3k.txt") into *k* clusters (*k*=2,3,4, …,10) respectively, and then find out under which *k* the clustering solution is the optimal for this dataset.

We click "Load Data File" in File menu bar to load data file "leuk72_3k.txt" and the file name appears in pop-up menu "Data set"; and then the following items are specified: the default "Clustering Rows" in Option menu is applicable (so no action is needed), the default true "Class labels in first column/row" in pop-up menu "Position" is applicable (no action needed), similarity metric "Pearson correlation" is selected from the pop-up menu "Measure", clustering algorithm "PAM" is chosen from pop-up menu "Algorithm", the default "Multiple Clustering" in pop-up menu "Times" is applicable (no action needed), giving the range of number of clusters "2 - 10" in Edit boxes for multiple clustering; finally, we press button "Run Clustering" to perform clustering process and the running information is displayed in running-state window (see Figure 1).

Once the clustering process runs over, we turn to the validation process. To estimate the optimal number of clusters from the above clustering solutions of PAM, we click "Estimate Number of Clusters" in Validation menu, and then select Calinski-Harabasz index from pop-up menu "Validity Indices", finally press button "Run Validation". Now index values across the number of clusters are displayed in the plotting window. As the maximum value of Calinski-Harabasz index indicates the optimal number of clusters (NC), the optimal NC with the largest Calinski-Harabasz value at *k*=3, where the clustering solution is the optimal, is indicated by a square symbol for PAM clustering on leukemia dataset (see Figure 1).

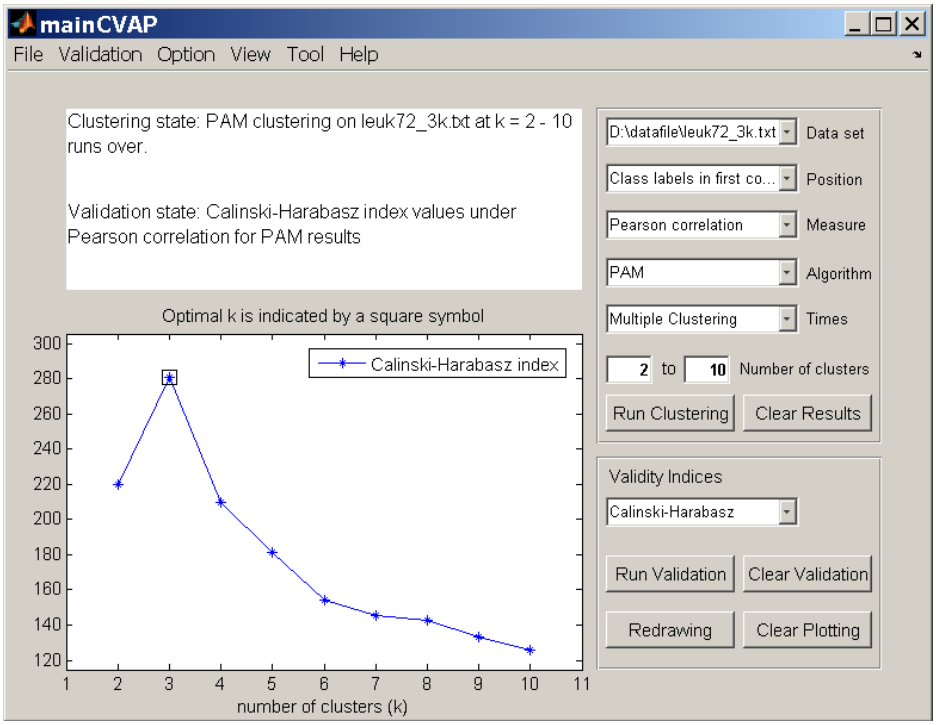

Figure 1. The optimal number of clusters, *k*=3, is given by Calinski-Harabasz index (internal index) for PAM clustering on leukemia dataset.

When "Hierarchical-centroid" clustering algorithm is selected from pop-up menu "Algorithm", we run the similar clustering process, but we need to choose special "RMSSTD group" from pop-up menu "Validity

Indices" to estimate the optimal NC for the hierarchical clustering. After pressing button "Run Validation", the index values of RMSSTD, R-squared, Semi-partial R-squared and Distance between two clusters (Centroid Distance) indices are displayed in the plotting window. As the steepest knee/elbow indicates optimal NC, i.e., the greater jump of these indices' values from larger to smaller NC, we infer that the optimal NC is at *k*=3, where the clustering solution is optimal (see Figure 2).

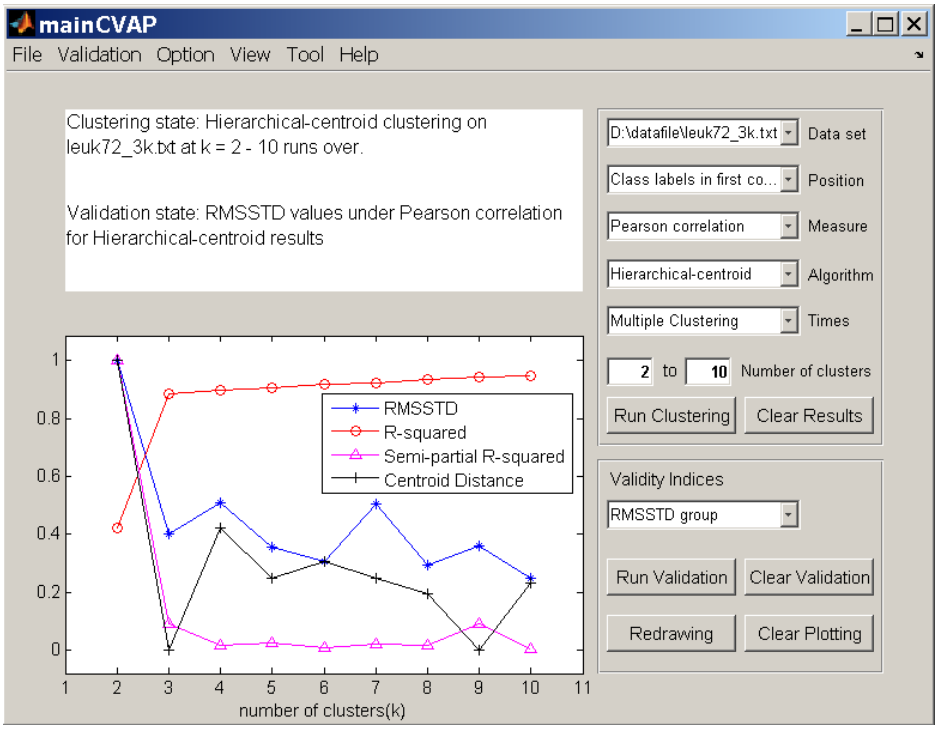

Figure 2. Steepest knee and elbows of RMSSTD, R-squared, Semi-partial R-squared and Centroid Distance indices (internal indices) indicate the optimal number of clusters at *k*=3 for hierarchical clustering on leukemia dataset.

For the yeast data set, we aim to find out the most suitable algorithm through performance comparison between candidate clustering algorithms: PAM, K-means, SOM and an external clustering algorithm, whose clustering solutions are stored in solution file "y208-solution.txt". Same as that in, Silhouette index, a composite index reflecting the compactness and separation of the clusters, is adopted as the criterion of clustering quality evaluation for performance comparison. Thus, the "best" algorithm will be used for cluster analysis of the data.

After loading the data file "y208_4k.txt" and clicking "Multi-Algorithm Validation" in Validation menu, we run a similar clustering process stated above; when clustering solutions from PAM are ready, we select Silhouette index from pop-up menu "Validity Indices"; and then press button "Run Validation" and Silhouette values across the number of clusters are displayed in the plotting window for PAM algorithm. Suppose that an external algorithm has clustered the data set and its clustering solutions are stored in solution file y208-solution.txt. Similarly, we run the clustering and validation processes for the K-means, external algorithm (denoted by solution file y208-solution.txt in the following figures) and SOM step by step (but no clustering process for the solution file). Now the Silhouette values corresponding to the four algorithms at *k*=2-10 are ready, and the index values at *k*=2-6 are redrawn for better observation by pressing button "Redrawing" (see Figure 3). From Figure 3 we may find that K-means has larger Silhouette values at *k*=2-4 and its Silhouette value at *k*=4 is the largest one among all the values. As a larger Silhouette value indicates a better quality of a clustering result, we prefer K-means to be the most suitable algorithm for the yeast data set.

Then, we cluster the data with K-means, and find the best clustering solution at the optimal NC, *k*=4 indicated by a square symbol, by Silhouette index, whose largest value indicates the optimal NC.

When a priori partition is known, external indices are used to assess the degree of agreement between the priori partition (e.g. "true" class labels) and a clustering result (e.g. cluster labels obtained from a cluster procedure). We can use them to achieve performance comparison between candidate clustering algorithms, and here we choose external index "Jaccard" from pop-up menu "Validity Indices" for performance

comparison. Similarly, we run the clustering and validation processes for the K-means, external algorithm and SOM step by step (but no clustering process for a solution file). When clustering solutions from an algorithm are ready, we press button "Run Validation" and Jaccard values across the number of clusters are displayed in the plotting window (redrawing at *k*=2-6 for good observation) (see Figure 4). In Figure 4 one may see that K-means has the largest Jaccard value at *k*=4. As higher the score better the solution, K-means has the best performance or is most suitable for the yeast data set.

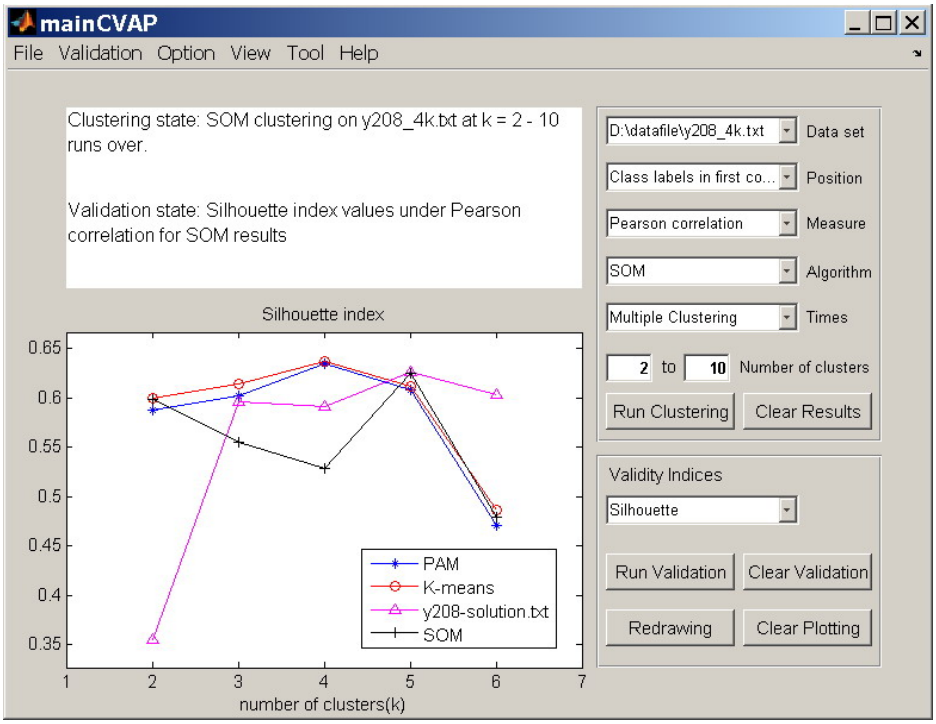

Figure 3. Silhouette (internal index) values of clustering solutions on yeast dataset corresponding to PAM, K-means, external algorithm (denoted by y208-solution.txt) and SOM are plotted for performance comparison between these algorithms. Larger the Silhouette values, better the clustering quality.

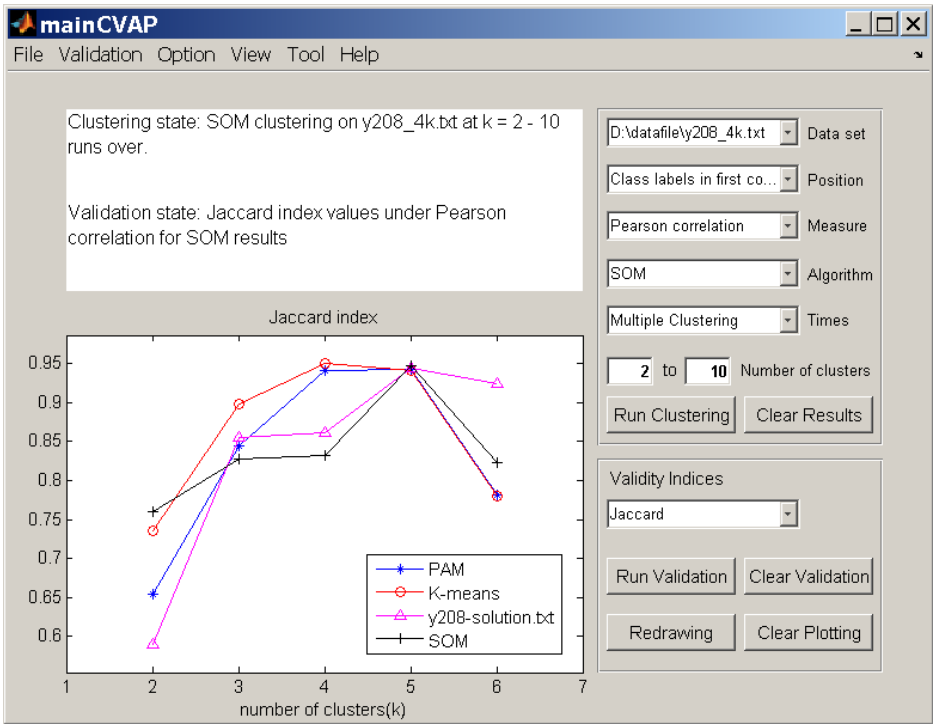

Figure 4. External index Jaccard is used for performance comparison between PAM, K-means, external algorithm (denoted by y208-solution.txt) and SOM on yeast dataset. Higher the score, better the solution.