CVAP: VALIDATION FOR CLUSTER ANALYSES

Kaijun Wang[1]*, Baijie Wang[2], and Liuqing Peng[2]

**[1]School of Mathematics and Computer Science, Fujian Normal University, Fuzhou 350007, P. R. China*
Email: wangkjun@yahoo.com
[2]School of Computer Science and Technology, Xidian University, Xian 710071, P. R. China.

ABSTRACT

Evaluation of clustering results (or cluster validation) is an important and necessary step in cluster analysis, but it is often time-consuming and complicated work. We present a visual cluster validation tool, the Cluster Validity Analysis Platform (CVAP), to facilitate cluster validation. The CVAP provides necessary methods (e.g., many validity indices, several clustering algorithms and procedures) and an analysis environment for clustering, evaluation of clustering results, estimation of the number of clusters, and performance comparison among different clustering algorithms. It can help users accomplish their clustering tasks faster and easier and help achieve good clustering quality when there is little prior knowledge about the cluster structure of a data set.

Keywords: Cluster validation, Validity indices, Visual cluster analysis environment

Availability of the program: The programs in Matlab are available from:
http://www.mathworks.com/matlabcentral/fileexchange/authors/24811

1 INTRODUCTION

Cluster analysis is an important technique in many research areas such as data mining, information science, agriculture technology, and biomedicine. For example, in cluster analysis of gene expression data, several clustering algorithms such as K-means, partitioning around medoids (PAM), hierarchical clustering (HC), and self-organizing map (SOM) are widely used (Gordon et al., 2005; Shamir et al., 2005; Thalamuthu et al., 2006). Different clustering algorithms with different properties tend to give somewhat or much different solutions, and there is no single "best" clustering method for all possible data sets. Then, an important effort is to select a proper or best one for a data set from candidate clustering algorithms. Once clustering results are obtained by a clustering algorithm, the next important step is to evaluate clustering solutions to determine an optimal solution or cluster structure for the data set, usually the number of clusters (NC). This step depends on evaluation of clustering results or cluster validation that aims to find a clustering solution that best fits the given data set.

It is usually time-consuming work to accomplish a clustering task because cluster analysis has many aspects to be treated carefully such as data preprocessing, similarity metrics, number of clusters, parameters of clustering algorithms, validity indices, the evaluation of clustering solutions, and so on. A good tool for cluster analysis can enhance work efficiency, achieve better results, and avoid possible mistakes on account of neglect of important factors in clustering processes. Hence, to better accomplish a clustering task, one expects a good clustering tool, which provides the necessary methods for cluster analysis, especially the technique of cluster validation. In this paper, we present an efficient cluster validation tool to serve the above purpose.

2 METHODS OF CLUSTER VALIDATION

One of the most important issues in cluster analysis is the evaluation of clustering results to find the partitioning that best fits the underlying data (Halkidi et al., 2001). For the evaluation of clustering solutions, it is usually the validity indices that are used to measure the quality of clustering results. There are two kinds of validity indices: external indices and internal indices. An external index is a measure of agreement between two partitions where the first partition is the *a priori* known clustering structure, and the second results from the clustering procedure (Dudoit et al., 2002). Internal indices are used to measure the goodness of a clustering structure without external information (Tseng et al., 2005). For external indices, we evaluate the results of a clustering algorithm based on a known cluster structure of a data set (or cluster labels). For internal indices, we evaluate the results using quantities and features inherent in the data set. The optimal NC is usually determined based on an internal validity index.

The principles of some widely-used internal indices for NC-estimation and clustering quality evaluation are introduced in the following:

- Silhouette index: A composite index reflecting the compactness and separation of clusters; a larger average Silhouette index indicates a better overall quality of the clustering result, so the optimal NC is the one that gives the largest average Silhouette value (Kaufman & Rousseeuw, 1990; Chen et al., 2002).
- Davies-Bouldin index: A measure of the average similarity between each cluster and its most similar one; small values correspond to clusters that are compact and have centers that are far away from each other; therefore, its minimum value determines the optimal NC (Dimitriadou et al., 2002; Bolshakova & Azuaje, 2003).
- Calinski-Harabasz index: The measures of between-cluster isolation and within-cluster coherence; its maximum value determines the optimal NC (Dudoit et al., 2002; Shu, 2003).
- Dunn index: A measure that maximizes the inter-cluster distances while minimizing the intra-cluster distances; its large values indicate the presence of compact and well-separated clusters, so the NC that maximizes the index is taken as the optimal NC (Halkidi et al., 2001; Bolshakova & Azuaje, 2003).
- RMSSTD index: A measure of the homogeneity of the formed clusters (or the variance of clusters) at each step of the HC algorithm; a lower RMSSTD value means better clustering, so its minimum determines the optimal NC (Halkidi et al., 2001; Kovács et al., 2005).(note: it is designed for HC)

One may choose a validity index to estimate an optimal NC, where the optimal clustering solution is found from a series of clustering solutions under different NCs. However, to find the best clustering solution for a clustering task depends on not only a validity index but also the appropriate clustering procedure. An obvious case is that using different clustering algorithms or different validity indices results in different clustering solutions for a specific clustering task. Therefore, there is still much complex work in the process of cluster validation.

3 CLUSTER VALIDATION TOOL CVAP

We report on a cluster validation tool for cluster analysis to deal with the aspects of cluster analysis mentioned in the introduction section, especially the optimal NC-estimation and the evaluation of clustering results. This cluster validation tool is developed using Matlab and called the Cluster Validity Analysis Platform (CVAP). The appropriate clustering procedures and necessary methods for cluster analysis are designed and included in the CVAP, covering a clustering process and a validation process (see Figure 1).

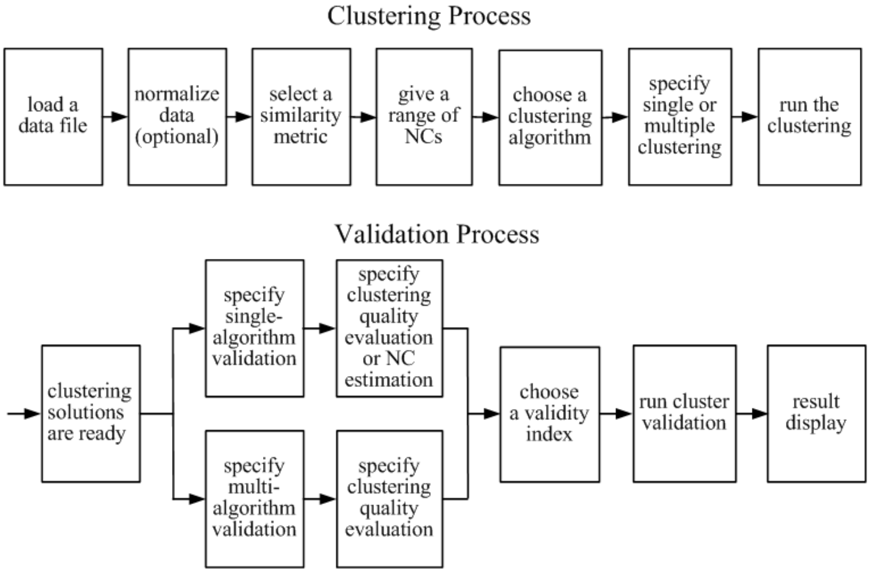

Clustering Process
load a data file
normalize data (optional)
select a similarity metric
give a range of NCs
choose a clustering algorithm
specify single or multiple clustering
run the clustering
Validation Process
clustering solutions are ready
specify single-algorithm validation
specify clustering quality evaluation or NC estimation
specify multi-algorithm validation
specify clustering quality evaluation
choose a validity index
run cluster validation
result display

Figure 1. The clustering process and validation process in the CVAP

The most important feature of the CVAP is that three important items of cluster analysis are included: the performance comparison of clustering algorithms, validity evaluation of clustering solutions, and (automatic) estimation of the number of clusters.

For the NC-estimation designed here, the basic idea is that a range of number of clusters are searched to find an optimal NC under a given clustering algorithm, i.e., the clustering algorithm is run to yield a clustering solution at each NC for different NCs, and then values of a chosen validity index are calculated, and the index-value graph for the clustering solutions under different NCs is plotted. Finally the optimal NC from the graph is determined according to the principle of the validity index (refer to the principles of some indices listed in the last section).

For performance comparison of clustering algorithms, we designed the computation and display (in a GUI window) of validity-index values of clustering solutions from different clustering algorithms (e.g., K-means, PAM, HC, and SOM) in the CVAP. With this function, we may find a best clustering algorithm for a data set by comparing validity-index values of clustering solutions from candidate algorithms and according to the principle of a chosen validity index. For example, PAM is inferred to be the best clustering algorithm if PAM has the largest Silhouette value among competing clustering algorithms.

For validity evaluation of clustering solutions, we designed three procedures for single-algorithm validation (solutions from one clustering algorithm are evaluated), multi-algorithm validation (the solutions from several clustering algorithms are evaluated comparatively), and the clustering quality evaluation (the target is to evaluate the goodness of a clustering result) respectively. Moreover, the CVAP supplies 18 commonly-used validity indices:

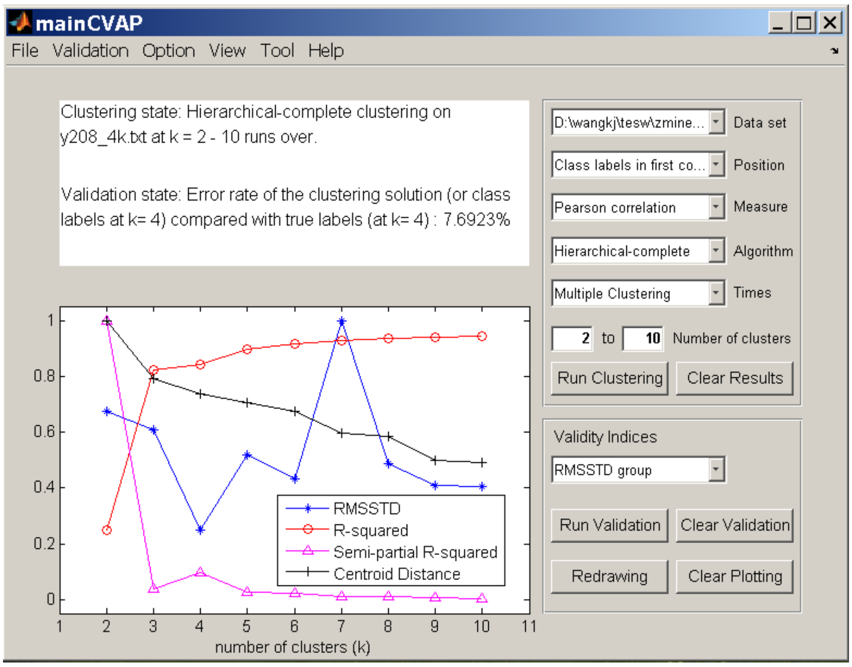
- The external indices include Rand, adjusted Rand, Jaccard, and Fowlkes-Mallows (FM) indices (Halkidi et al., 2001; Dudoit et al., 2002).
- The internal indices are: Silhouette, Davies-Bouldin, Calinski- Harabasz, Dunn, Hubert-Levin (C-index), Krzanowski-Lai and Hartigan indices (Bolshakova & Azuaje, 2003; Bolshakova & Azuaje, 2006; Dudoit et al., 2002); the Root-mean-square standard deviation (RMSSTD), R-squared, Semi-partial R-squared (SPR) and Distance between two clusters (CD) indices (Halkidi et al., 2001; Halkidi et al., 2002); the weighted inter-intra index (Strehl, 2002); and the Homogeneity and Separation indices (Sharan et al., 2003).

In brief, the CVAP is an efficient cluster validation tool with dedicatedly-designed validation methods, validation procedures, and visual analysis environment for cluster analysis.

4 EXAMPLE OF CLUSTER VALIDATION BY CVAP

With the above-mentioned functions and features, the CVAP can help us obtain good clustering quality especially when there is little prior knowledge about the cluster structure of a data set. Here we illustrate the clustering task for a yeast dataset with 208 genes in four well-separated clusters (Ben-Hur et al., 2002) (assuming that there is little prior knowledge about its cluster structure). Assuming that one chooses, because of prior experience, clustering algorithm HC (under complete linkage) to cluster the data set and then adopts validity index RMSSTD (the lower RMSSTD value means better clustering) to estimate the optimal NC and find the optimal clustering solution. The final result is: the optimal clustering solution under the NC of 4 has an error rate of 7.69% (see Figure 2), compared with the known cluster labels. However, the HC chosen from experience might not be the best clustering algorithm for the data set.

We can obtain better clustering quality by using the CVAP when a more suitable algorithm for the data set is found. For the yeast dataset, K-means is selected as the best clustering algorithm from the performance comparison among candidates K-means, PAM, HC, and SOM based on the Silhouette index (a larger value indicates better quality of the clustering result) because K-means has the largest Silhouette value 0.6363 at the NC of 4 (see Figure 3). Then the optimal NC for the clustering solutions of K-means is estimated to be 4, where the Silhouette value is the largest for K-means (see Figure 4). The optimal clustering solution by K-means (the K-means clustering repeats 10 times and returns the solution with the minimum mean-square-error) at the NC of 4 has an error rate of 2.4038%, compared with the known cluster labels. This best result of K-means equals that of PAM (error rate of 2.4038%), better than that of HC (error rate of 7.69%) and that of SOM (error rate of 12.50%).

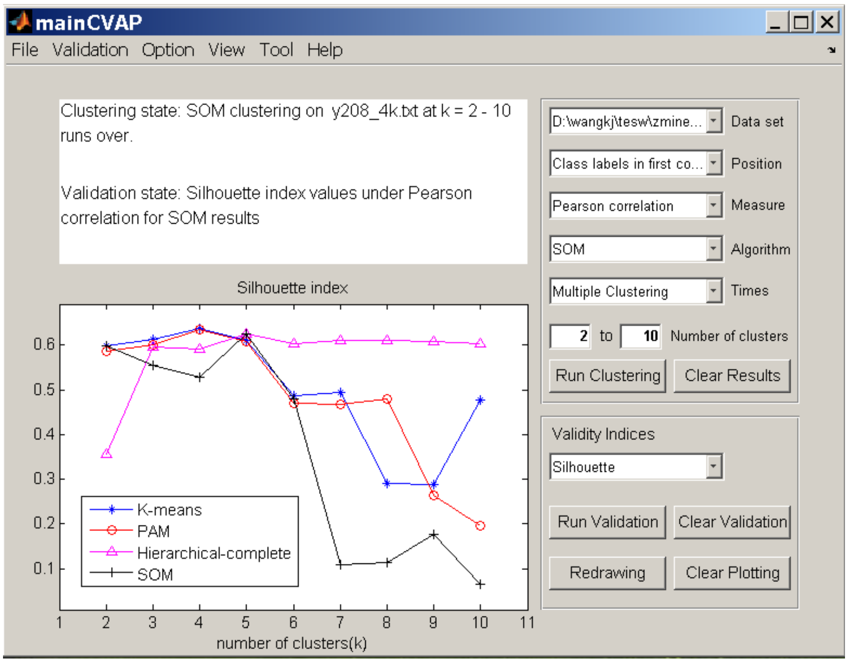


Clustering state: Hierarchical-complete clustering on y208_4k.txt at k = 2 - 10 runs over.

Validation state: Error rate of the clustering solution (or class labels at k= 4) compared with true labels (at k= 4) : 7.6923%

Data set: D:\wangkj\tesw\zmine...
Position: Class labels in first co...
Measure: Pearson correlation
Algorithm: Hierarchical-complete
Times: Multiple Clustering
Number of clusters: 2 to 10
Validity Indices: RMSSTD group

Figure 2. Hierarchical clustering on yeast dataset. The optimal number of clusters estimated by RMSSTD is 4 (where the RMSSTD value is the smallest), and the optimal clustering solution under *k*=4 has the error rate of 7.6923%.

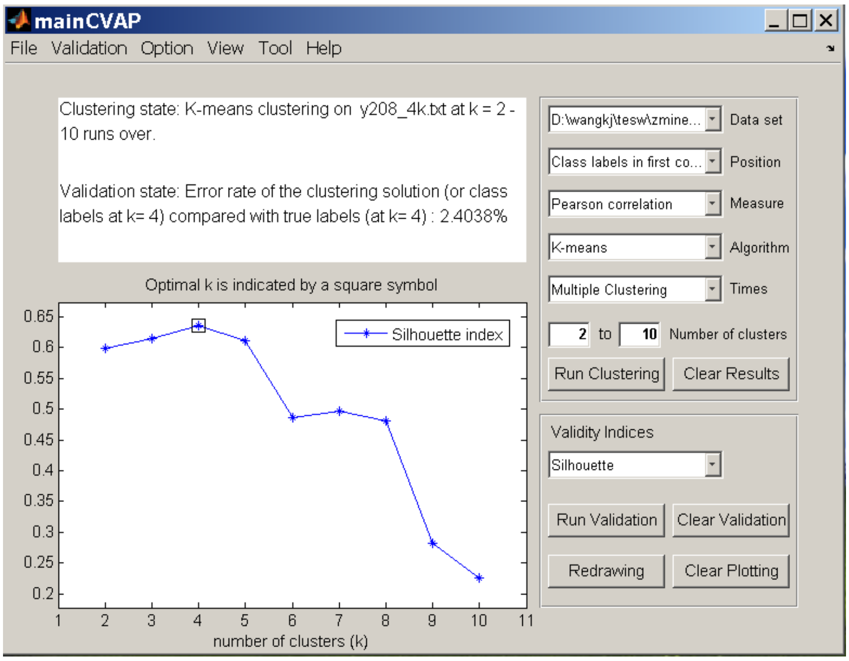


Clustering state: SOM clustering on y208_4k.txt at k = 2 - 10 runs over.

Validation state: Silhouette index values under Pearson correlation for SOM results

Data set: D:\wangkj\tesw\zmine...
Position: Class labels in first co...
Measure: Pearson correlation
Algorithm: SOM
Times: Multiple Clustering
Number of clusters: 2 to 10
Validity Indices: Silhouette

Figure 3. Performance comparison among K-means, PAM, HC and SOM based on Silhouette index. K-means is the best clustering algorithm for yeast dataset, and has the largest Silhouette value 0.6363 at *k*=4.



Clustering state: K-means clustering on y208_4k.txt at k = 2 - 10 runs over.

Validation state: Error rate of the clustering solution (or class labels at k= 4) compared with true labels (at k= 4) : 2.4038%

D:\wangkj\tesw\zmine... Data set

Class labels in first co... Position

Pearson correlation Measure

K-means Algorithm

Multiple Clustering Times

2 to 10 Number of clusters

Validity Indices

Silhouette

Figure 4. K-means clustering on yeast dataset. The optimal number of clusters estimated by Silhouette is 4 (where the Silhouette value is the largest), and the optimal clustering solution at *k*=4 has an error rate of 2.4038%.

5 DISCUSSION

The CVAP provides necessary functions and a visual analysis environment for clustering and cluster validation. Three important aspects of cluster analysis can be accomplished with the CVAP: performance comparison of clustering algorithms, validity evaluation of clustering solutions, and (automatic) NC-estimation.

With the function of NC-estimation, we can find an optimal NC easily from different clustering solutions, reducing time-consuming comparison and analysis of different clustering solutions by individual comparisons

By the function of the performance comparison of clustering algorithms, we can find the best clustering algorithm for a data set especially when there is little prior knowledge about the data set. This helps achieve better clustering quality if the clustering algorithm selected by our experience for a data set is not the best one.

For validity evaluation of clustering solutions, the CVAP provides three evaluation procedures for different conditions and supplies many external and internal indices. This function can help find the best clustering solution within the searching scope and executes the tasks of NC-estimation and performance comparison of clustering algorithms.

The clustering analysis example in the last section demonstrates that by using the performance-comparison function of the CVAP, we find the better clustering algorithm K-means instead of the initially-selected HC for the data set. In addition, by using the validity-evaluation and NC-estimation functions of the CVAP, we obtain the better clustering quality (or clustering error rate of 2.40%) from K-means, compared with that from HC (the error rate of 7.69%).

Hence, the CVAP can help us increase work efficiency with better results for cluster analysis tasks. Further, possible mistakes resulting from unfamiliarity with cluster validation or neglect in clustering processes can be avoided.

6 CONCLUSION

Cluster validation is an important and necessary step in cluster analysis. A visual cluster validation tool CVAP is proposed to facilitate cluster validation and cluster analysis. The CVAP provides the necessary methods and tools as well as an analysis environment for clustering and cluster validation and can help a user accomplish his clustering task faster and better.

5 ACKNOWLEDGEMENTS

This work is supported in part by Natural Science Found of China (Grant No. 60573139 and 60574039).

6 REFERENCES

Ben-Hur, A., Elisseeff, A., & Guyon, I. A. (2002) Stability based method for discovering structure in clustered data. *Pac Symp Biocomputing*, vol. 7, 6-17.

Bolshakova, N. & Azuaje, F. (2003) Cluster validation techniques for genome expression data. *Signal Processing*, 83(4): 825-833.

Bolshakova, N. & Azuaje, F. (2006) Estimating the number of clusters in DNA microarray data. *Methods of Information in Medicine*, 45(2):153-157.

Chen, G., Jaradat, S. A., Banerjee, N., Tanaka, T. S., Ko, M. S. H., & Zhang, M. Q. (2002) Evaluation and Comparison of Clustering Algorithms in Anglyzing ES Cell Gene Expression Data. *Statistica Sinica*, 12: 241-262.

Dimitriadou, E., Dolnicar, S., & Weingessel, A. (2002) An examination of indexes for determining the Number of Cluster in binary data sets. *Psychometrika*, 67(1): 137-160.

Dudoit, S. & Fridlyand, J. (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3(7): 0036.1-21.

Gordon, G. J., Rockwell, G. N., Godfrey, P. A., Jensen, R. V., Glickman, J. N., Yeap, B. Y., Richards, W. G., Sugarbaker, D. J., & Bueno, R. (2005) Validation of Genomics-Based Prognostic Tests in Malignant Pleural Mesothelioma. *Clinical Cancer Research*, Vol. 11, 4406-4414.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001) On Clustering Validation Techniques. *Intelligent Information Systems Journal*, 17(2-3): 107-145.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002) Cluster Validity Methods: Part II. *SIGMOD Record*, September 2002.

Kaufman, L. & Rousseeuw, P. J. (1990) Finding Groups in Data: An Introduction to Cluster Analysis. New York, John Wiley & Sons.

Kovács, F., Legány, C., & Babos, A. (2005) Cluster validity measurement techniques. *Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence*, Budapest, Nov. 2005, 18-19.

Strehl, A. (2002) *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining*. Ph.D thesis, The University of Texas at Austin.

Shamir, R., Maron-Katz, A., Tanay, A., Linhart, C., Steinfeld, I., Sharan, R., Shiloh, Y., & Elkon, R. (2005) EXPANDER - an integrative program suite for microarray data analysis. *BMC Bioinformatics*, 6:232.

Sharan, R., Maron-Katz, A., & Shamir, R. (2003) CLICK and EXPANDER: A System for Clustering and Visualizing Gene Expression Data. *Bioinformatics*, 19: 1787-1799.

Shu, G., Zeng, B., Chen, Y. P., & Smith, O. H. (2003) Performance assessment of kernel density clustering for gene expression profile data. *Comparative and Functional Genomics*, 4(3): 287-299.

Thalamuthu, A, Mukhopadhyay, I, Zheng, X, & Tseng, G. C. (2006) Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics*, 22(19):2405-12.

(Article history: Received 6 November 2007, Accepted 27 February 2009, Available online 24 April 2009)