Machaon CVE: Cluster validation for gene expression data

Nadia Bolshakova[a,*] *and Francisco Azuaje*[b]

[a]*Department of Computer Science, Trinity College Dublin, Dublin 2, Ireland*
[b]*School of Computing and Mathematics, University of Ulster, Jordanstown Co. Antrim, BT37 0QB Northern Ireland, U.K*

[*]To whom correspondence should be addressed

ABSTRACT

Summary: This paper presents a cluster validation tool for gene expression data. *Machaon-CVE* (Clustering and Validation Environment) system aims to partition samples or genes into groups characterised by similar expression patterns, and to evaluate the quality of the clusters obtained.

Availability: The program is freely available for non-profit use on request at http://www.cs.tcd.ie/Nadia.Bolshakova/Machaon.html

Contact: Nadia.Bolshakova@cs.tcd.ie

Supplementary information: http://www.cs.tcd.ie/Nadia.Bolshakova/Machaon.html

INTRODUCTION

DNA microarray technologies allow measuring the expression of thousands of genes in parallel under multiple experimental conditions. Genomic and post-genomic studies (Schena *et al.*, 1995), such as disease diagnosis, drug discovery and toxicological research have been benefited from it (Debouck and Goodfellow, 1999, Gray *et al.*, 1998).

An important step in the analysis of gene expression data is the detection of samples or genes with similar expression patterns. Several clustering algorithms have been developed for gene expression data. Also solutions to systematically evaluate the quality of the clusters have been presented (Bolshakova and Azuaje, 2002). The prediction of the correct number of clusters in a data set is a critical problem in unsupervised classification. Various cluster validity indices have been proposed to measure the quality of clustering results (Azuaje, 2002; Dudoit. and Fridlyand, 2002). Clustering algorithms may require the a-priori definition of the number of clusters. Hence, a clustering algorithm can be executed several times, with different number of clusters in each run, and the clustering partition that optimises a validity index is selected as the best partition. Previous studies have not provided integrated tools for both clustering and automatically assessing the quality of the resulting clusters. Therefore, there is the need to design software platforms, which integrate clustering and validation methods for predicting the optimal number of clusters in gene expression data analyses.

The *Machaon CVE* is a cross-platform Java-based tool, which offers multiple clustering and validity methods for DNA microarray data analysis. It aims: a) to partition samples or genes into groups characterised by similar expression patterns, and b) to evaluate the quality of the clusters obtained.

SYSTEM OVERVIEW

The software is implemented as a multi-window Java application, which allows working with different datasets, clustering (hierarchical and k-means) and validation (C-index, Davis-Bouldin, Dunn's, Goodman-Kruskal and Silhouette indices) algorithms, and results simultaneously. For further information on the implementation, of these algorithms the reader is referred to the supplementary information page. The system supports several modifications of tabular data format widely used by third-party clustering tools (Herrero *et al.*, 2001) Moreover, an XML-based format is being designed to address some of the limitations observed in traditional formats, such as inability to store multiple clustering and validation results within a dataset.

Multiple clustering may be applied to a single dataset and the results may be easily compared. Every clustering result may be selected and validated across a number of parameterised validation methods. Both clustering and validation results are represented as two-level tree in the bottom of the corresponding data set window (Figure 1). Clustering indices are also displayed in additional columns of a data set table. Every such column is associated with a single partition. The results of a hierarchical clustering can also be displayed using dendrograms. Users may choose from a collection of clustering and validation techniques, compare the results from each method and generate interpretations.

Several methods for measuring gene-to-gene (or sample-to-sample), intercluster and intracluster distances can be used in any combination. This is important to research the influence of different distance metrics on both clustering and validation. Machaon CVE provides data normalization functionality, which may be either selected as an option of clustering/validation or used to produce a normalized dataset.

Apart from the clustering and validation results, the system shows, if known, the natural classification structure of the data (leukemia types in the example illustrated in Figure 1), which allows comparisons against clustering results and validation analyses across natural classes.

Despite the fact that *Machaon CVE* was developed for DNA microarray expression analysis applications, it may be effectively used for clustering/validating other biomedical and physical data with no limitations.

ACKNOWLEDGEMENTS

This contribution was partly supported by the Enterprise Ireland Research Innovation Fund 2001.

REFERENCES

Azuaje,F. (2002) A cluster validity framework for genome expression data, *Bioinformatics*, **18**, 319-320.

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

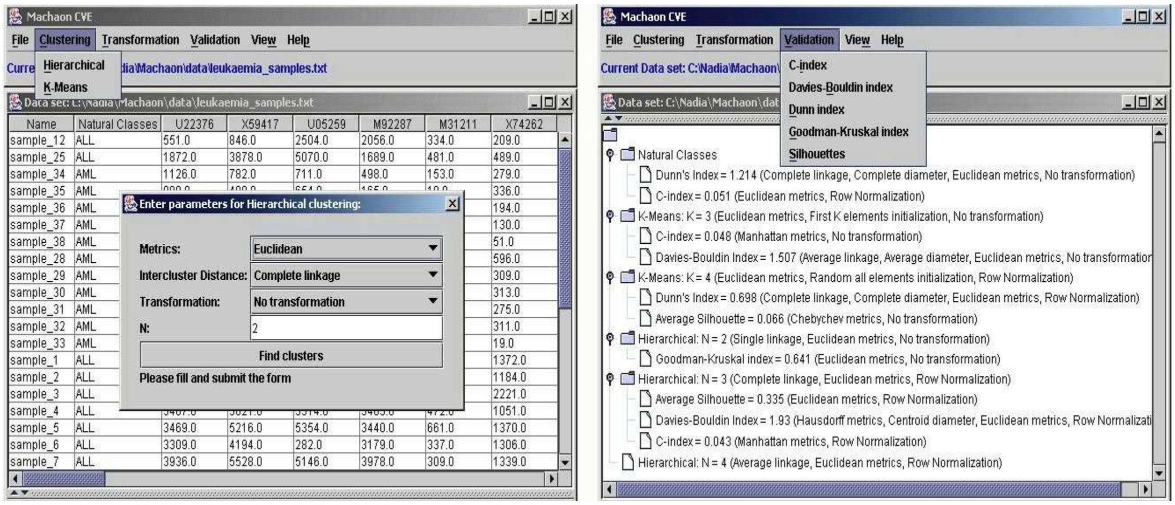
Debouck,C. and Goodfellow,P.N. (1999) DNA microarrays in drug discovery and development, *Nature Genet.*, **21**, 48-50.

Dudoit,S. and Fridlyand,J. (2002) A prediction-based resampling method for estimation the number of cluster in a dataset, *Genome Biology*, **3**, 1-21.

Grey,N.S., Wodicka,L., Thunnissen,A.M., Norman,T.C., Kwon,S., Espinoza,F.H., Morgan,D.O., Barnes,G., LeClerc,S., Meijer,L., Kim,S.H., Lockhart,D.J. and Schultz,P.G. (1998) Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science*, **281**, 533-538.

Herrero,J., Valencia,A. and Dopazo,J. (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns, *Bioinformatics*, **17**, 126-136.

Schena,M., Shalon,D., Davis,R.W. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, **270**, 467-470.



Curre[ ]dia\Machaon\data\leukaemia_samples.txt

Data set: C:\Nadia\Machaon\data\leukaemia_samples.txt

Name	Natural Classes	U22376	X59417	U05259	M92287	M31211	X74262
sample_12	ALL	551.0	846.0	2504.0	2056.0	334.0	209.0
sample_25	ALL	1872.0	3878.0	5070.0	1689.0	481.0	489.0
sample_34	AML	1126.0	782.0	711.0	498.0	153.0	279.0
sample_35	AML	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	336.0
sample_36	AML						194.0
sample_37	AML						130.0
sample_38	AML						51.0
sample_28	AML						596.0
sample_29	AML						309.0
sample_30	AML						313.0
sample_31	AML						275.0
sample_32	AML						311.0
sample_33	AML						19.0
sample_1	ALL						1372.0
sample_2	ALL						1184.0
sample_3	ALL						2221.0
sample_4	ALL	[illegible]	[illegible]	[illegible]	[illegible]	472.0	1051.0
sample_5	ALL	3469.0	5216.0	5354.0	3440.0	661.0	1370.0
sample_6	ALL	3309.0	4194.0	282.0	3179.0	337.0	1306.0
sample_7	ALL	3936.0	5528.0	5146.0	3978.0	309.0	1339.0

Enter parameters for Hierarchical clustering:

Metrics: Euclidean

Intercluster Distance: Complete linkage

Transformation: No transformation

N: 2

Please fill and submit the form

Current Data set: C:\Nadia\Machaon\

Data set: C:\Nadia\Machaon\dat

- Natural Classes
  - Dunn's Index = 1.214 (Complete linkage, Complete diameter, Euclidean metrics, No transformation)
  - C-index = 0.051 (Euclidean metrics, Row Normalization)
- K-Means: K = 3 (Euclidean metrics, First K elements initialization, No transformation)
  - C-index = 0.048 (Manhattan metrics, No transformation)
  - Davies-Bouldin Index = 1.507 (Average linkage, Average diameter, Euclidean metrics, No transformatio
- K-Means: K = 4 (Euclidean metrics, Random all elements initialization, Row Normalization)
  - Dunn's Index = 0.698 (Complete linkage, Complete diameter, Euclidean metrics, Row Normalization)
  - Average Silhouette = 0.066 (Chebychev metrics, No transformation)
- Hierarchical: N = 2 (Single linkage, Euclidean metrics, No transformation)
  - Goodman-Kruskal index = 0.641 (Euclidean metrics, No transformation)
- Hierarchical: N = 3 (Complete linkage, Euclidean metrics, Row Normalization)
  - Average Silhouette = 0.335 (Euclidean metrics, Row Normalization)
  - Davies-Bouldin Index = 1.93 (Hausdorff metrics, Centroid diameter, Euclidean metrics, Row Normalizati
  - C-index = 0.043 (Manhattan metrics, Row Normalization)
- Hierarchical: N = 4 (Average linkage, Euclidean metrics, Row Normalization)

Figure 1. Screenshots from the *Machaon CVE* explaining different aspects of its functionality.
(a) Data set window with parameter window for hierarchical clustering.
(b) Result tree illustrating clustering and validation results.