Machaon CVE: cluster validation for gene expression data

Nadia Bolshakova¹,* and Francisco Azuaje²

¹Department of Computer Science, Trinity College Dublin, Dublin 2, Ireland and
²School of Computing and Mathematics, University of Ulster, Jordanstown, Northern Ireland, UK

Received on January 24, 2003; revised on April 23, 2003; accepted on June 30, 2003

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 characterized 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; Grey 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 evaluate systematically the quality of the clusters have been presented (Bolshakova and Azuaje, 2003). 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 optimizes 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 a 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 Clustering and Validation Environment (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 characterized 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 data sets, 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 data set.

Multiple clustering may be applied to a single data set and the results may be easily compared. Every clustering result may be selected and validated across a number of parameterized validation methods. Both clustering and validation results are represented as two-level tree in the bottom of the corresponding data set window (Fig. 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 intracuster 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 data set.

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 Fig. 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 used effectively 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