



A cluster validity framework for genome expression data

F. Azuaje

Department of Computer Science, Trinity College, Dublin 2, Ireland

Received on July 31, 2001; revised on October 1, 2001; accepted on October 9, 2001

ABSTRACT

Summary: This paper presents a method for the assessment of expression cluster validity.

Availability: Executable programs are available on request from the author.

Contact: Francisco.Azuaje@cs.tcd.ie

Supplementary information: <http://www.cs.tcd.ie/Francisco.Azuaje/Cval.html>

Clustering is a useful approach to analyzing genome expression data. It aims to partition samples or genes into groups characterized by similar expression patterns. A number of clustering algorithms have been proposed (such as hierarchical clustering and neural networks), but fewer solutions to systematically evaluate the quality of the clusters obtained have been presented.

Once a clustering process is performed researchers may deal with some of the following questions: Is this a relevant partition? Should we analyze these clusters? Is there a better partition? The framework presented here aims to help researchers address these questions.

It has been shown that determining the ‘right’ number of clusters in experimental data is a complex and time-consuming process. An effective strategy may be to first decide a good estimate of the correct number of clusters. Our system predicts the optimal number of expression clusters, which may represent the best results to consider for interpretation purposes.

This system implements the Dunn’s validity index, which has been suggested as an effective estimator for different types of clustering applications (Bezdek and Pal, 1998). This index is based on the idea of identifying sets of clusters that are compact and well separated. For any partition $U \leftrightarrow X: X_1 \cup \dots \cup X_i \cup \dots \cup X_c$, where X_i represents the i th cluster of such partition, the Dunn’s validation index, V , is defined as:

$$V(U) = \min_{1 \leq i \leq c} \left\{ \min_{\substack{1 \leq j \leq c \\ j \neq i}} \left\{ \frac{\delta(X_i, X_j)}{\max_{1 \leq k \leq c} \{\Delta(X_k)\}} \right\} \right\}.$$

$\delta(X_i, X_j)$ defines the distance between clusters X_i and X_j (intercluster distance); $\Delta(X_k)$ represents the

intracluster distance of cluster X_k , and c is the number of clusters defined by the partition U . The main goal of this measure is to maximize intercluster distances whilst minimizing intracluster distances. Thus large values of V correspond to good clusters. Therefore, the number of clusters that maximizes V is taken as the optimal number of clusters, c .

Eighteen validity indexes based on the Dunn’s measure were compared. These indexes consist of different combinations of intercluster and intracluster distance techniques found in the literature. Six intercluster distances, δ_i , $1 \leq i \leq 6$; and 3 intracluster distances, Δ_j , $1 \leq j \leq 3$ were implemented. Thus for example, V_{13} , represents a validity index based on an intercluster distance, δ_1 , and an intracluster distance Δ_3 . The mathematical definitions of these intercluster and intracluster distances are included in the supplementary information section.

By way of example, this validation process is tested on expression data from a study on the molecular classification of *leukemias* (Golub *et al.*, 1998). Clustering is performed using the *GeneCluster* tool, which implements a Self-Organizing Map (SOM) algorithm. A second set of experiments performed on *B-cell lymphoma* data is described in the supplementary section. The data analyzed consisted of 38 bone marrow samples: 27 *Acute Lymphoblastic Leukemia* (ALL) and 11 *Acute Myeloid Leukemia* (AML), whose original descriptions and experimental protocols can be found on the *MIT Whitehead Institute* web site.

Figure 1a shows the values of the 18 validity indexes and the average index at each number of clusters, c , for $c = 2$ to 6. The shaded entries correspond to the optimal values of the indexes. Sixteen of the indexes indicated the correct value $c = 2$ while the remaining favour $c = 3$ and 4. Moreover V_{13} equally supported 2, 4 and 6 clusters. Figure 1b describes the clusters obtained using the optimal value $c = 2$. However, the best value for c might be debatable. One might also consider $c = 4$ as a correct choice since these clusters capture relevant information for the discovery of the ALL subclasses, B-cell and T-cell, as shown in Figure 1c. The average index predicts $c = 2$ as the first choice and $c = 4$ as the second best

