Cluster validation techniques for genome expression data

N. Bolshakova* and F. Azuaje

Department of Computer Science, Trinity College Dublin

* Corresponding author. Department of Computer Science, Trinity College, Dublin 2, Ireland
Tel.: +353-1-608-3688, *e-mail address:* Nadia.Bolshakova@cs.tcd.ie

Abstract

Several clustering algorithms have been suggested to analyse genome expression data, but fewer solutions have been implemented to guide the design of clustering-based experiments and assess the quality of their outcomes. A cluster validity framework provides insights into the problem of predicting the correct the number of clusters. This paper presents several validation techniques for gene expression data analysis. Normalisation and validity aggregation strategies are proposed to improve the prediction about the number of relevant clusters. The results obtained indicate that this systematic evaluation approach may significantly support genome expression analyses for knowledge discovery applications.

Keywords: Genome expression; Clustering; Cluster validation; Genomic data mining.

1. Introduction

Several clustering techniques have been applied to the analysis of genome expression data [1,2]. Clustering can support the identification of existing underlying relationships among a set of variables such as biological conditions or perturbations. It may represent a basic tool not only for the classification of known categories, but also for the discovery of relevant classes. In the genome expression domain it has provided the basis for novel clinical diagnostic and prognostic studies [3].

On the other hand fewer solutions to systematically evaluate the quality of the clusters have been presented [4,5]. The prediction of the correct number of clusters is a fundamental problem in unsupervised classification problems. Many clustering algorithms require the definition of the number of clusters beforehand. To overcome this problem, various cluster validity indices have been proposed to assess the quality of a clustering partition [6]. This approach consists of running a clustering algorithm several times and obtaining different partitions, and the clustering partition that optimises a validity index is selected as the best partition. Thus, the main goal of a cluster validity technique is to identify the partition of clusters for which a measure of quality is optimal [5]. Cluster validity techniques include the *Silhouette method* [7], *Dunn's based index* [8,9], *Davies-Bouldin index* [10] and the *C-index* [11]. For a review on validation techniques the reader is referred to [12].

The remainder of the paper is organised as follows. In Section 2 the experimental data are described. Section 3 introduces relevant validation methods and their applications to the analysis of expression data. A comparative study is presented in Section 4. Conclusions are presented in Section 5.

2. Genomic Expression Data

Genome expression data reflect the level of activity of several genes in parallel under different biochemical conditions [13]. This method is based on the idea that genes that are contained in a particular pathway should exhibit similar patterns of expression. The data studied in this paper consisted of two expression sets originating from *leukaemia* [14] and *B-cell lymphoma* [15] data. For related work on clustering based on these data sets the reader is referred to [16, 17, 18].

2.1. Leukaemia data

The data comprised 38 samples (27 *acute lymphoblastic leukaemia* (ALL) and 11 *acute myeloid leukaemia* (AML)) described by the expression levels of 50 genes with suspected roles in this type of cancer. These data were obtained from a study published by Golub and co-workers [14]. They presented a model to distinguish two sub-classes of ALL sample, known as *B-cell ALL* and *T-cell ALL.* The original data and experimental methods are available at http://www.genome.wi.mit.edu/MPR.

2.2. B-cell lymphoma data

The data consisted of 63 samples (45 *diffuse large B-cell lymphoma* (DLBCL) and 18 normal) described by the expression levels of 23 genes. These data were obtained from a study published by Alizadeh and colleagues [15], who identified subgroups of DLBCL based on the analysis of the patterns generated by a specialized cDNA microarray technique. The study distinguished two categories of DLBCL: *GC B-like* DLBCL (22 samples) and *Activated B-like* DLBCL (23 samples). Data sets and experimental methods are available at http://llmpp.nih.gov/lymphoma.

In this paper clustering is performed using the *Kohonen Self-organising Maps* (SOM) algorithm, which have been applied to analyse expression profiles in several biomedical studies [19]. This model is relatively computationally inexpensive and it shows significant advantages in comparison to other algorithms [19].

3. Validation techniques

This section introduces three validation methods known as the *Silhouette*, the *Dunn's* and the *Davies-Bouldin indices*, which have shown to be robust strategies for the prediction of optimal clustering partitions [7, 9, 10].

3.1. Silhouette index

For a given cluster, X_j (j = 1,…, c), this method assigns to each sample of X_j a quality measure, $s(i)$ (i = 1,…, m), known as the *Silhouette width.* The Silhouette width is a confidence indicator on the membership of the i^{th} sample in cluster X_j. The Silhouette width for the i^{th} sample in cluster X_j is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \qquad (1)$$

where $a(i)$ is the average distance between the i^{th} sample and all of the samples included in X_j; 'max' is the maximum operator, and $b(i)$ is the minimum average distance between the i^{th} sample and all of the samples clustered in X_k ($k = 1,\ldots, c$; $k \neq j$). From this formula it follows that $-1 \leq s(i) \leq 1$.

When a $s(i)$ is close to 1, one may infer that the i^{th} sample has been "well-clustered", i.e. it was assigned to an appropriate cluster. When a $s(i)$ is close to zero, it suggests that the i^{th} sample could also be assigned to the nearest neighbouring cluster. If $s(i)$ is close to –1, one may argue that such a sample has been "misclassified" [7]. Thus, for a given cluster, X_j ($j = 1,\ldots, c$), it is possible to calculate a cluster Silhouette S_j, which characterises the heterogeneity and isolation properties of such a cluster:

$$S_j = \frac{1}{m}\sum_{i=1}^{m} s(i)\,, \qquad (2)$$

where m is number of samples in S_j.

It has been shown that for any partition $U \leftrightarrow X$: $X_1 \cup \ldots X_i \cup \ldots X_c$, a *Global Silhouette value*, GS_u, can be used as an effective validity index for U.

$$GS_u = \frac{1}{c}\sum_{j=1}^{c} S_j \qquad (3)$$

Furthermore, it has been demonstrated that equation (3) can be applied to estimate the most appropriate number of clusters for U. In this case the partition with the maximum S_u is taken as the optimal partition.

3.2. *Dunn's Index*

This index identifies sets of clusters that are compact and well separated. For any partition $U \leftrightarrow X$: $X_1 \cup \ldots X_i \cup \ldots X_c$, where X_i represents the i^{th} cluster of such partition, the Dunn's validation index, D, is defined as:

$$D(U) = \min_{1\leq i\leq c}\left\{\min_{\substack{1\leq j\leq c\\ j\neq i}}\left\{\frac{\delta(X_i, X_j)}{\max\limits_{1\leq k\leq c}\{\Delta(X_k)\}}\right\}\right\}, \qquad (4)$$

where $\delta(X_i, X_j)$ 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 of partition U. The main goal of this measure is to maximise intercluster distances whilst minimising intracluster distances. Thus large values of D correspond to good clusters. Therefore, the number of clusters that maximises D is taken as the optimal number of clusters, c.

3.3. *Davies-Bouldin Index*

As the Dunn's index, the Davies-Bouldin index aims at identifying sets of clusters that are compact and well separated. The Davies-Bouldin validation index, DB, is defined as:

$$DB(U) = \frac{1}{c}\sum_{i=1}^{c}\max_{i\neq j}\left\{\frac{\Delta(X_i)+\Delta(X_j)}{\delta(X_i, X_j)}\right\}, \qquad (5)$$

where U, $\delta(X_i, X_j)$, $\Delta(X_i)$, $\Delta(X_j)$ and c are defined as in equation (4). Small values of DB correspond to clusters that are compact, and whose centres are far away from each other. Therefore, the cluster configuration that minimizes DB is taken as the optimal number of clusters, c.

Different methods may be used to calculate intercluster and intracluster distances [5]. Thirty-six indices based on equations (4) and (5) were calculated. These indices consist of different combinations of intercluster and intracluster distance methods. 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, D_{13}, represents a Dunn's 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 described in the following section.

3.4 Distances used to implement the validation methods

3.4.1 Basic distance metrics

The distance between two samples, $d(x,y)$, was calculated using the well-known *Euclidean, Manhattan* and *Chebychev* metrics [20].

3.4.2 Intercluster distances

Six intercluster distances are used for the calculation of the Dunn's and Davies-Bouldin validity indices. The *single linkage* distance defines the closest distance between two samples belonging to two different clusters. The *complete linkage* distance represents the distance between the most remote samples belonging to two different clusters. The *average linkage* distance defines the average distance between all of the samples belonging to two different clusters. The *centroid linkage* distance reflects the distance between the centres of two clusters. The *average of centroids linkage* represents the distance between the centre of a cluster and all of samples belonging to a different cluster. *Hausdorff metrics* are based on the discovery of a maximal distance from samples of one cluster to the nearest sample of another cluster. These intercluster distances are defined as follows.

Single linkage:

$$\delta_1(S,T) = \min_{x \in S, y \in T}\left\{ d(x,y) \right\}, \qquad (6)$$

where S and T are clusters from partition U; $d(x,y)$ defines the distance between any two samples, x and y, belonging to S and T respectively; $|S|$ and $|T|$ provide the number of samples included in clusters S and T respectively.

Complete linkage:

$$\delta_2(S,T) = \max_{x \in S, y \in T}\left\{ d(x,y) \right\} \qquad (7)$$

Average linkage:

$$\delta_3(S,T) = \frac{1}{|S||T|}\sum_{\substack{x \in S \\ y \in T}} d(x,y) \qquad (8)$$

Centroid linkage:

$$\delta_4(S,T) = d(vs, vt), \qquad (9)$$

where $vs = \frac{1}{|S|}\sum_{x \in S} x$, $vt = \frac{1}{|T|}\sum_{y \in T} y$

Average to centroids linkage:

$$\delta_5(S,T)=\frac{1}{|S|+|T|}\left(\sum_{x\in S}d(x,vt)+\sum_{y\in T}d(y,vs)\right) \quad (10)$$

Hausdorff metrics:

$$\delta_6(S,T)=\max\{\delta(S,T),\delta(T,S)\}, \quad (11)$$

where $\delta(S,T)=\max_{x\in S}\left\{\min_{y\in T}\{d(x,y)\}\right\}$, $\delta(T,S)=\max_{y\in T}\left\{\min_{x\in S}\{d(x,y)\}\right\}$

3.4.3. Intracluster distances

Three intracluster distances are used to calculate the Dunn's and Davies-Bouldin validity indices. The *complete diameter* distance represents the distance between the most remote samples belonging to the same cluster. The *average diameter* distance defines the average distance between all of the samples belonging to the same cluster. The *centroid diameter* distance reflects the double average distance between all of the samples and the cluster's centre. These intracluster distances are defined as follows.

Complete diameter:

$$\Delta_1(S)=\max_{x,y\in S}\{d(x,y)\}, \quad (12)$$

where *S* is a cluster from partition *U*; *d*(*x*,*y*) defines the distance between any two samples, *x* and *y*, belonging to *S*; $|S|$ represents the number of samples included in clusters *S*.

Average diameter:

$$\Delta_2(S)=\frac{1}{|S|\cdot(|S|-1)}\sum_{\substack{x,y\in S\\ x\neq y}}d(x,y) \quad (13)$$

Centroid diameter:

$$\Delta_3(S)=2\left(\frac{\sum_{x\in S}d(x,\bar{v})}{|S|}\right), \quad (14)$$

where $\bar{v}=\frac{1}{|S|}\sum_{x\in S}x$

4. Comparison of validation techniques

The Dunn's and Davies-Bouldin validity indices require the definition of at least two clusters. The same situation applies to the Silhouette method. To compute the minimum average distance between the sample in one cluster and all of the samples from not the same cluster, the Silhouette width formula (1) requires at least two clusters. Thus, calculations for null-case are not considered here.

Golub with colleagues suggested 50 "informative" genes, which are correlated with the AML/ALL cancer types. The problem of feature selection is a crucial problem, which is not part of the goals of our paper. This problem is accentuated when biological data sets are described in terms of many tens or hundreds of features.

Feature selection methods have been recently applied to improve decision support tasks in expression studies [21, 22].

Tables 1 and 2 show the Global Silhouette values, GS_u, for each partition, and the Silhouette values, S for each number of clusters, c, for $c = 2$ to $c = 6$, using the leukaemia and DLBCL data respectively. The bold entries correspond to the optimal values predicted by this validation method. In this case $c = 2$ is suggested as the best clustering configuration for both expression data sets. Table 1 suggests that the partition consisting of 4 clusters may also be considered as a useful partition, because it generates the second highest GS_u. An examination of this partition confirms that it represents relevant information relating to the detection of the ALL subclasses, B-cell and T-cell, as demonstrated by Golub and colleagues [14]. In the case of the DLBCL data, a partition consisting of three clusters is predicted as the second best choice according to both Silhouette (Table 2) and Dunn's index [5] validation techniques. This is a relevant partition as it distinguishes the groups GC B-like DLBCL, activated B-like DLBCL and normal cells [5].

Table 1. Silhouette values for expression clusters originating from leukemia samples. The entries represent the Global Silhouette values, GS_u, for each partition, and the Silhouette values, S, for each cluster defining a partition. Bold entries highlight the optimal number of clusters, c, predicted by this method.

C	GS_u	S_1	S_2	S_3	S_4	S_5	S_6
2	**0.43**	**0.17**	**0.57**				
3	0.14	0.11	0.35	0.11			
4	0.25	0.15	0.31	0.31	0.26		
5	0.19	0.07	0.45	0.23	0.23	0.21	
6	0.23	0.28	0.23	0.28	0.42	0.14	0.14

Table 2. Silhouette values for expression clusters originating from DLBCL samples. The entries represent the Global Silhouette values, GS_u, for each partition, and the Silhouette values, S, for each cluster defining a partition. Bold entries highlight the optimal number of clusters, c, predicted by this method.

C	GS_u	S_1	S_2	S_3	S_4	S_5	S_6
2	**0.26**	**0.23**	**0.28**				
3	0.17	0.17	0.26	0.13			
4	0.17	0.11	0.32	0.18	0.13		
5	0.14	0.10	0.12	0.21	0.21	0.10	
6	0.15	0.22	0.08	0.26	0.26	0.03	0.17

The results shown in Tables 1 and 2 were obtained using the well-known Euclidean distance between samples. Previous studies have shown that the estimation of the optimal partition is generally insensitive to the type of metrics selected to implement equations (1) and (2) [4].

One may conclude that the Silhouette method is suitable for estimating only the first choice or best partition. Its ability to rank cluster partitions in terms of their biological and statistical validity is debatable. Nevertheless, this method may be successfully used in combination with other validation techniques for predicting different optimal clustering partitions. This application is currently being investigated [23].

The eighteen Dunn's indices and the average index at each number of clusters, *c*, for *c* = 2 to *c* = 6 were computed in [4, 5]. An examination of these results confirms that *c* = 2 represent the most appropriate prediction for both leukaemia and DLBCL data.

The application of different intercluster/intracluster distance combinations may produce validation indices of different scale ranges. Hence those indices with higher values may have a stronger effect on the calculation of the average index values. This may result in a biased prediction of the optimal number of clusters. For example, the bottom lines of the Tables 3 and 4 represent the average values of the Davies-Bouldin validation index (for leukaemia and DLBCL data respectively), which are strongly influenced by the values based on the *complete diameter* distance (12).

Table 3. Predicting the correct number of clusters: Davies-Bouldin validity indexes for expression clusters originating from leukaemia data. The entries represent the Davies-Bouldin values using 3 types of intracluster measures and 6 types of intercluster measures. Normalised Davies-Bouldin validity indexes are given between brackets. Bold entries represent the optimal number of clusters, *c*, predicted by each index.

Normalised validity index	*c* = 2	*c* = 3	*c* = 4	*c* = 5	*c* = 6
DB_{11}	8.89 (1.75)	5.70 (-0.15)	5.14 (-0.48)	5.26 (-0.40)	**4.72 (-0.73)**
DB_{21}	**1.29 (–1.21)**	1.36 (-0.77)	1.68 (1.31)	1.53 (0.33)	1.53 (0.34)
DB_{31}	**2.30 (-1.40)**	2.48 (0.13)	2.64 (1.41)	2.47 (0.04)	2.44 (-0.17)
DB_{41}	**2.83 (-1.69)**	3.86 (0.89)	3.72 (0.53)	3.62 (0.28)	3.50 (-0.01)
DB_{51}	**2.56 (-1.58)**	2.88 (0.29)	3.03 (1.17)	2.86 (0.18)	2.82 (-0.06)
DB_{61}	**1.69 (-1.07)**	1.73 (-0.92)	2.30 (1.31)	2.01 (0.18)	2.09 (0.50)
DB_{12}	4.97 (1.71)	3.69 (-0.05)	**3.14 (-0.82)**	3.55 (-0.24)	3.30 (-0.59)
DB_{22}	**0.72 (-1.36)**	0.87 (-0.54)	0.97 (-0.01)	1.13 (0.84)	1.18 (1.07)
DB_{32}	**1.28 (-1.56)**	1.60 (0.14)	1.52 (-0.29)	1.71 (0.73)	1.75 (0.97)
DB_{42}	**1.58 (-1.65)**	2.48 (0.66)	2.13 (-0.25)	2.46 (0.60)	2.48 (0.65)
DB_{52}	**1.43 (-1.60)**	1.85 (0.23)	1.74 (-0.25)	1.97 (0.72)	2.00 (0.89)
DB_{62}	**0.95 (-1.38)**	1.11 (-0.65)	1.33 (0.30)	1.42 (0.69)	1.50 (1.05)
DB_{13}	6.87 (1.70)	4.92 (0.04)	**4.01 (-0.74)**	4.53 (-0.30)	4.06 (0.70)
DB_{23}	**1.00 (-1.51)**	1.16 (-0.46)	1.28 (0.30)	1.35 (0.72)	1.38 (0.95)
DB_{33}	**1.77 (-1.68)**	2.12 (0.72)	1.99 (-0.17)	2.10 (0.54)	2.10 (0.59)
DB_{43}	**2.18 (-1.63)**	3.30 (1.01)	2.81 (-0.15)	3.07 (0.46)	3.01 (0.32)
DB_{53}	**1.98 (-1.68)**	2.46 (0.73)	2.29 (-0.15)	2.43 (0.55)	2.43 (0.55)
DB_{63}	**1.31 (-1.42)**	1.48 (-0.64)	1.75 (0.61)	1.72 (0.45)	1.84 (1.00)
Average	**2.53 (-0.96)**	2.50 (0.03)	2.41 (0.20)	2.51 (0.35)	2.45 (0.37)

Table 4. Predicting the correct number of clusters: Davies-Bouldin validity indexes for expression clusters originating from DLBCL data. The entries represent the Davies-Bouldin values using 3 types of intracluster measures and 6 types of intercluster measures. Normalised Davies-Bouldin validity indexes are given between brackets. Bold entries represent the optimal number of clusters, *c*, predicted by each index.

Normalised validity index	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$
DB_{11}	6.57 (1.59)	5.80 (0.31)	5.25 (-0.58)	**5.02** (**-0.96**)	5.39 (-0.36)
DB_{21}	**1.29** (**-1.73**)	1.69 (0.19)	1.78 (0.62)	1.69 (0.20)	1.80 (0.73)
DB_{31}	**2.63** (**-1.72**)	2.78 (0.60)	2.74 (-0.03)	2.78 (0.64)	2.77 (0.50)
DB_{41}	**3.78** (**-1.49**)	4.58 (-0.30)	4.75 (-0.05)	5.45 (0.99)	5.36 (0.85)
DB_{51}	**3.06** (**-1.59**)	3.35 (0.04)	3.31 (-0.15)	3.50 (0.91)	3.48 (0.79)
DB_{61}	**2.26** (**-1.60**)	2.39 (-0.34)	2.51 (0.77)	2.50 (0.73)	2.47 (0.44)
DB_{12}	3.79 (1.48)	3.44 (0.08)	**3.16** (**-1.03**)	3.21 (-0.82)	3.49 (0.29)
DB_{22}	**0.75** (**-1.63**)	1.00 (-0.15)	1.08 (0.31)	1.10 (0.45)	1.20 (1.01)
DB_{32}	**1.51** (**-1.44**)	1.65 (-0.31)	1.67 (-0.15)	1.80 (0.94	1.80 (0.97)
DB_{42}	**2.18** (**-1.39**)	2.71 (-0.43)	2.89 (-0.12)	3.52 (1.00)	3.48 (0.93)
DB_{52}	**1.76** (**1.40**)	1.98 (-0.36)	2.02 (-0.18)	2.26 (0.97)	2.26 (0.96)
DB_{62}	**1.30** (**-1.39**)	1.42 (-0.60)	1.53 (0.19)	1.61(0.74)	1.66 (1.06)
DB_{13}	**5.26** (**1.59**)	4.68 (0.07)	4.33 (-0.84)	**4.31** (**-0.88**)	4.67 (0.06)
DB_{23}	**1.04** (**-1.65**)	1.36 (-0.13)	1.47 (0.37)	1.48 (0.42)	1.60 (0.98)
DB_{33}	**2.10** (**-1.42**)	2.24 (-0.36)	2.27 (-0.13)	2.42 (0.95)	2.42 (0.96)
DB_{43}	**3.02** (**-1.39**)	3.69 (-0.45)	3.94 (-0.10)	4.73 (1.01)	4.67 (0.93)
DB_{53}	**2.45** (**-1.39**)	2.70 (-0.38)	2.75 (-0.18)	3.04 (0.98)	3.04 (0.96)
DB_{63}	**1.81** (**-1.37**)	1.93 (-0.66)	2.08 (0.25)	2.18 (0.79)	2.21 (1.00)
Average	**2.59** (**-1.00**)	2.74 (-0.18)	2.75 (-0.06)	2.92 (0.50)	2.99 (0.73)

To resolve this problem the following normalisation technique has been applied. Given a cluster configuration consisting of *c* clusters, for any partition $U_c \leftrightarrow X$: $X_1 \cup ... \cup X_c$, normalised Dunn's indices - D_{ij}^*, are calculated as:

$$D_{ij}^*(U_c) = \frac{D_{ij}(U_c) - \overline{D}_{ij}}{\sigma D_{ij}} , \tag{15}$$

$$\overline{D}_{ij} = \frac{1}{n} \sum_k D_{ij}(U_k), \tag{16}$$

where *i* reflects the selection of intercluster distance calculation method ($i = 1, \ldots, 6$), *j* is the selection of intracluster distance calculation method ($j = 1, ..., 3$), $D_{ij}(U_c)$ is the value of a Dunn's validity index, *n* is the number of partitions, σD_{ij} – standard deviation of $D_{ij}(U_c)$ across all values of *c*. The normalised Davis-Bouldin indices may be calculated by formula (15) using the Davis-Bouldin indices instead of Dunn's ones.

Tables 3 and 4 depict the non-normalised and normalised Davies-Bouldin index values for the leukaemia and DLBCL data respectively. Normalised validity indices are given between brackets. This normalisation scheme may offer a more robust mechanism to predict the correct number of clusters. It highlights the distinction between the index values from different clustering configurations.

The results shown in Tables 3 and 4 were obtained when *d(x,y)* was calculated using the Euclidean distance. Tables 5-8 summarise the effects of three measures, *d(x,y)* described in 3.4.1 on the calculation of the non-normalised and normalised Davies-Bouldin and Dunn's cluster validity indices. It suggests that the estimation of the optimal partition by normalised and non-normalised indices is not sensitive to the type of metrics, *d(x,y)*, implemented.

Table 5. Davies-Bouldin validity indexes for expression clusters originating from leukaemia data. The entries represent the average Davies-Bouldin values based on the distances shown in Tables 3, and using three measures for *d*(*x*,*y*). Normalised Davies-Bouldin validity indexes are given between brackets. Bold entries represent the optimal number of clusters, *c*, predicted by each method.

Validity index based on distances	*c* = 2	*c* = 3	*c* = 4	*c* = 5	*c* = 6
Euclidian	2.53 (**-0.96**)	2.50 (0.04)	2.41 (0.20)	2.51 (0.35)	**2.45** (0.37)
Manhattan	**3.19** (**-1.08**)	4.09 (0.30)	3.66 (0.05)	3.98 (0.47)	3.70 (0.26)
Chebychev	3.30 (**-0.67**)	2.85 (0.11)	2.82 (-0.15)	2.89 (0.61)	**2.80** (0.10)

Table 6. Davies-Bouldin validity indexes for expression clusters originating from DLBCL data. The entries represent the average Davies-Bouldin values based on the distances shown in Tables 3, and using three measures for *d*(*x*,*y*). Normalised Davies-Bouldin validity indexes are given between brackets. Bold entries represent the optimal number of clusters, *c*, predicted by each method.

Validity index based on distances	*c* = 2	*c* = 3	*c* = 4	*c* = 5	*c* = 6
Euclidian	**2.59** (**-1.00**)	2.74 (-0.18)	2.75 (-0.06)	2.92 (0.50)	2.99 (0.73)
Manhattan	**3.91** (**-0.98**)	4.31 (-0.11)	4.38 (-0.05)	4.80 (0.44)	4.79 (0.69)
Chebychev	**3.11** (**-1.11**)	3.37 (-0.27)	3.20 (-0.16)	3.48 (0.74)	3.60 (0.80)

Table 7. Dunn's validity indexes for expression clusters originating from leukaemia data. The entries represent the average Dunn's values based on the distances shown in Table 3, and using three measures for *d*(*x*,*y*). Normalised Dunn's validity indexes are given between brackets. Bold entries represent the optimal number of clusters, *c*, predicted by each method.

Validity index based on distances	*c* = 2	*c* = 3	*c* = 4	*c* = 5	*c* = 6
Euclidian	**0.93** (**1.47**)	0.48 (-0.46)	0.45 (-0.08)	0.39 (-0.55)	0.40 (-0.38)
Manhattan	**1.70** (**1.63**)	0.86 (-0.42)	0.79 (-0.09)	0.65 (-0.73)	0.66 (-0.40)
Chebychev	**0.90** (**1.29**)	0.48 (0.10)	0.49 (-0.20)	0.39 (-0.61)	0.40 (-0.58)

Table 8. Dunn's validity indexes for expression clusters originating from DLBCL data. The entries represent the average Dunn's values based on the distances shown in Tables 3, and using three measures for *d*(*x*,*y*). Normalised Dunn's validity indexes are given between brackets. Bold entries represent the optimal number of clusters, *c*, predicted by each method.

Validity index based on distances	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$
Euclidian	**0.99** (**1.35**)	0.79 (0.32)	0.66 (-0.39)	0.68 (-0.24)	0.60 (-1.03)
Manhattan	**1.57** (**1.25**)	1.21 (0.24)	1.02 (-0.45)	1.04 (-0.23)	0.92 (-0.81)
Chebychev	**0.97** (**1.46**)	0.79 (0.17)	0.70 (-0.21)	0.69 (-0.55)	0.63 (-0.87)

Another approach to predicting the optimal partition is an aggregation method based on a weighed voting strategy. One of such examples is shown in the Table 9 for the Davies-Bouldin indices and the leukaemia data. This table was obtained from Table 3 by replacing the index values by weighed votes, whose values range from 1 to 5. Thus, for example, DB_{11} represents the smallest index value and suggests the partition $c = 6$ as the optimal partition, hence its weighed vote is equal to 5. On the other hand DB_{11} represents the highest index value for partition $c = 2$, hence its weighed vote is equal to 1. After computing all of the Davies-Bouldin indices, the average weighed vote for each cluster partition has been calculated and it confirms that $c = 2$ represents the most appropriate prediction. In this case the partition $c = 2$ obtained an average weighed valued equal to 4.33. Nevertheless the best value for c may be disputed, one might consider the partition consisting of 4 clusters as a correct choice as these clusters capture relevant information for the discovery *of B-cell and T-cell* ALL subclasses [5]. This observation is confirmed by the average weighed value obtained for $c = 4$ in Table 9.

Table 9. Predicting the correct number of clusters for leukaemia data by weighed voting technique. The entries represent vote values based on Davies-Bouldin validation index using 3 types of intracluster and 6 types of intercluster measures.

Validity index	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$
DB_{11}	1	2	4	3	5
DB_{21}	5	4	1	3	2
DB_{31}	5	2	1	3	4
DB_{41}	5	1	2	3	4
DB_{51}	5	2	1	3	4
DB_{61}	5	4	1	3	2
DB_{12}	1	2	5	3	4
DB_{22}	5	4	3	2	1
DB_{32}	5	3	4	2	1
DB_{42}	5	1	4	3	2
DB_{52}	5	3	4	2	1
DB_{62}	5	4	3	2	1
DB_{13}	1	2	5	3	4
DB_{23}	5	4	3	2	1
DB_{33}	5	1	4	3	2
DB_{43}	5	1	4	2	3

DB_{53}	5	1	4	3	2
DB_{63}	5	4	2	3	1
Average	4.33	2.50	3.06	2.67	2.44

5. Conclusions

Several clustering techniques have been proposed for the analysis of genome expression data. Cluster validity indices represent useful tools to support such a task. They are particularly relevant in applications in which there is not a priori indication of the actual number of clusters. In this paper three validation indices were applied to two expression data sets, using different intracluster and intercluster distances. Combination of these methods may be successfully used for the assessment of cluster validity. It was shown that these methods might support the prediction of the optimal cluster partitioning for those data sets. Normalisation and weighed voting techniques are proposed to improve the prediction of the number of clusters based on multiple indices. Other validation techniques, such as Goodman-Kruskal index [24] and Hubert's Γ statistic [25], as well as the comparison and combination of results obtained from different clustering algorithms, will be part of future work.

Comparisons indicate that the normalisation of indices may improve the prediction process. Normalisation allows smoothing the effect of the highest values on the calculation of the average index values. Moreover, it effectively highlights the differences between the average index values from different clustering configurations. The advantage of a weighed voting approach lies in a robust aggregation of multiple validation methods in order to improve the estimation of the most adequate clustering partition.

The clustering of both data sets in the current research is performed using the SOM algorithm. This validation framework has also been tested on the K-Means clustering algorithm, and other methods are currently being investigated [23]. Due to time and space constraints additional analyses using artificial data sets have not been included. Two data sets with known class structures have been analysed, which have been previously assessed using systematic approaches [14, 15]. Future work includes a simulation study using artificial expression data and the analysis of more complex experimental data sets.

The methods implemented in this research may contribute to the evaluation of clustering results and the prediction of optimal cluster partitions. The results obtained suggest that such a validity approach may represent an effective tool to support biomedical knowledge discovery in genome expression data.

Acknowledgements

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

References

[1] M. Granzow, D. Berrar, W Dubitzky, A. Schuster, F. Azuaje, R. Eils, "Tumor identification by gene expression profiles: a comparison of five different clustering methods", ACM-SIGBIO Newsletters, Vol. 21, 2001, pp. 16-22

[2] K.Y. Yeung, D.R. Haynor, W.L. Ruzzo, "Validating clustering for gene expression data", Bioinformatics, Vol. 17, 2001, pp. 309-318

[3] M. Bittner, P. Meltzer, Y. Chen, Y. Jiang, E. Seftor, M. Hendrix, M. Radmacher, R. Simon, Z. Yakhini, A. Ben-Dor, N. Sampas, E. Dougherty, E. Wang, F. Marincola,

C. Gooden, J. Lueders, A. Glatfelter, P. Pollock, J. Carpten, E. Gillanders, D. Leja, Dietrich, C. Beaudry, M. Berens, D. Alberts, V. Sondak, N. Hayward, J. Trent, “Molecular classification of cutaneous malignant melanoma by gene expression profiling”, Nature, Vol. 406, 2000, pp. 536-540

[4] F. Azuaje, N. Bolshakova, “Clustering genome expression data: design and evaluation principles”, in: D. Berrar, W. Dubitzky, M. Granzow, ed., Understanding and Using Microarray Analysis Techniques: A Practical Guide, London: Springer Verlag, 2002. *In press.*

[5] F. Azuaje, “A cluster validity framework for genome expression data”, Bioinformatics, Vol. 18, 2002, pp. 319-320

[6] M. Halkidi, Y. Batistakis, M. Vazirgiannis, “On clustering validation techniques”, JIIS, Vol. 17, 2001, pp.107-145

[7] P.J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis”, J. Comp App. Math, Vol. 20, 1987, pp. 53-65

[8] J. Dunn, “Well separated clusters and optimal fuzzy partitions”, J.Cybernetics, Vol. 4, 1974, pp. 95-104

[9] J.C. Bezdek, N.R. Pal, “Some new indexes of cluster validity”, IEEE Transactions on Systems, Man and Cybernetics, Vol. 28, Part B, 1998, pp. 301-315

[10] D.L. Davies, D.W. Bouldin, “A cluster separation measure”, IEEE Transactions on Pattern Recognition and Machine Intelligence, Vol. 1, No. 2, 1979, pp. 224-227

[11] L. Hubert, J. Schultz, “Quadratic assignment as a general data-analysis strategy”, British Journal of Mathematical and Statistical Psychologie, Vol. 29, 1976, pp. 190-241

[12] S. Gunter, H. Burke, “Validation indices for graph clustering”, in J.-M. Jolion, W. Kropatsch, M. Vento (eds.): Proc. 3rd IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition, 2001, pp. 229 – 238.

[13] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, “Quantitative monitoring of gene expression”, Vol. 270, 1995, pp. 467-470

[14] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gassenbeck, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring”, Science, Vol. 286, 1999, Vol. 531-537

[15] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson, L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Bird, D. Botstein, P.O. Brown, M. Staudt, “Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling”, Nature, Vol. 403, 2000, pp. 503-511.

[16] C. Ambroise, G.J. McLachlan, “Selection bias in gene extraction on basis of microarray gene expression data”, Proceedings of the National Academy of Sciences USA, Vol. 99, 2002, pp. 6562-6566.

[17] A Berns, “Cancer: Gene expression in diagnosis”, Nature, Vol. 403, 2000, pp. 491-492

[18] F. Azuaje, “A computational neural approach to support the discovery of gene function and classes of cancer”, IEEE Transactions on Biomedical Engineering, Vol. 48, 2001, pp. 332-339

[19] J. Quackenbush, “Computational analysis of microarray data”, Nature Reviews Genetics, Vol. 2, 2001, pp. 418-427

[20] B. Everitt, Cluster Analysis, London: Edward Arnold, 1993

[21] K. Dunne, P. Cunningham, F. Azuaje, “Solutions to Instability Problems with Sequential Wrapper-based Approaches to Feature Selection”, submitted to The Journal of Machine Learning Research, 2002

[22] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, E.S. Lander, M. Loda, P.W. Kantoff, T.R. Golub, W.R. Sellers, “Gene expression correlates of clinical prostate cancer behavior”, Cancer Cell, Vol. 1, 2002, pp. 203-9

[23] N. Bolshakova, F. Azuaje, “Improving expression data mining through cluster validation”, submitted to the 4th Annual IEEE EMBS Special Topic Conference on Information Technology Applications in Biomedicine, 2003

[24] L. Goodman, W. Kruskal, “Measures of associations for cross-validations”, J. Am. Stat. Assoc., Vol. 49, 1954, pp. 732-764

[25] L.J. Hubert, P. Arabie, “Comparing partitions”, J. Classification, Vol. 2, 1985, pp. 193-218