IMPROVING EXPRESSION DATA MINING THROUGH CLUSTER VALIDATION

N. Bolshakova, F. Azuaje
Department of Computer Science, Trinity College Dublin, Ireland

Abstract-**This paper presents several cluster evaluation techniques for gene expression data analysis. Normalisation and validity aggregation strategies are proposed to improve the prediction of the number of relevant clusters. The effect of different intracluster and intercluster distances on this prediction process is studied. This approach is applied to a publicly released medulloblastomas tumour data set. The results suggest that it may represent an effective tool to support biomedical knowledge discovery tasks based on gene expression data.**

Keywords **- Gene expression, data mining, clustering, cluster evaluation, validity indices.**

I. INTRODUCTION

Recent advances in DNA microarray technology, also known as gene chips, allow measuring the expression of thousands of genes in parallel and under multiple experimental conditions [1]. This technology is having a significant impact on genomic and post-genomic studies. Disease diagnosis, drug discovery and toxicological research benefit from the of microarray technology. A main step in the analysis of gene expression data is the detection of samples or genes with similar expression patterns. A number of data mining techniques have been applied to the analysis of gene expression data. Clustering is a fundamental approach to gene expression knowledge discovery [2, 3]. Solutions for the systematic evaluation of the quality of the clusters have been recently proposed [4, 5]. Moreover, the prediction of the correct number of clusters is a critical problem in unsupervised classification problems. Many clustering algorithms require the number of clusters given as an input parameter. Different cluster validity indices have been suggested to address this problem [6]. A cluster validity index indicates the quality of a resulting clustering process. Thus, the clustering partition that optimises the validity index under consideration is chosen as the best partition [4]. This paper presents cluster validity techniques for gene expression data analysis. Normalisation and validity aggregation strategies are proposed to improve the prediction of the number of relevant clusters.

II. METHODOLOGY

This section introduces the gene expression data, clustering and validation methods under consideration. Three validation methods were applied: the *Silhouettes* [7], the *Dunn's* [8] and the *Davies-Bouldin* [9] indices, which have shown to be robust strategies for the prediction of optimal clustering partitions

The data comprised 34 *medulloblastoma* tumour samples (9 *desmoplastic medulloblastomas* and 25 *classic medulloblastomas*) described by the expression levels of 140 genes with suspected roles in these subtypes of cancer. These data were obtained from a study published by Pomeroy and co-workers [10]. They developed a classification system based on DNA microarray gene expression data to distinguish *desmoplastic* and *classic medulloblastomas*. It allowed the prediction of clinical outcomes in children with *medulloblastomas* on the basis of the expression profiles of their tumours at diagnosis [10]. The original data and experimental methods are available at http://www.genome.wi.mit.edu/MPR/CNS.

The validation methods are illustrated using the *K-Means* algorithm, which has been applied to analyse expression profiles in several biomedical and systems biology studies [11]. This agglomerative clustering technique finds clusters in a set of unlabeled data based on the selection of the desired number of, *K*, classes. The performance of the *K-Means* clustering algorithm may be improved by estimating the number of clusters represented in the data. For further information on the implementation and analysis of this algorithm the reader is referred to [12, 13]. The cluster validity techniques described here have also been illustrated using other clustering methods [14, 15].

1.1 Cluster validity techniques

A. Silhouette method

For a given cluster, X_j (j = 1,…, c), the silhouette technique assigns to the i*th* sample of X_j a quality measure, $s(i)$ (i = 1,…, m), known as the *silhouette width*. This value is a confidence indicator on the membership of the ith sample in cluster X_j and it is defined as:

$$s(i) = (b(i) - a(i)) / \max\{a(i), b(i)\}, \quad (1)$$

where *a(i)* is the average distance between the ith sample and all of the samples included in X_j; and *b(i)* is the minimum average distance between the ith sample and all of the samples clustered in X_k (k = 1,…, c; $k \neq j$). From this formula it follows that $s(i)$ has a value between –1 and 1.

When $s(i)$ is close to 1, one may infer that the ith sample has been assigned to an appropriate cluster. When $s(i)$ is close to zero, it suggests that the sample could also be assigned to the nearest neighbouring cluster, i.e. such a sample lies equally far away from both clusters. 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, it is possible to calculate a *cluster silhouette* S_j, which characterises the heterogeneity and isolation properties of such a cluster. It is calculated as the sum of all samples' *silhouette widths* in X_j. Moreover, for any partition, a *global silhouette value* or *silhouette index,* GS_u, can be used as an effective validity index for a partition U.

$$GS_u = \frac{1}{c}\sum_{j=1}^{c} S_j \tag{2}$$

Furthermore, it has been demonstrated that equation (2) can be applied to estimate the "correct" number of clusters for partition U [7]. In this case the partition with the maximum silhouette index value is taken as the optimal partition.

B. Dunn's and Davies-Bouldin methods

These indices aim to identify sets of clusters that are compact and well separated. For any partition $U \leftrightarrow X$: $X_1 \cup ... X_i \cup \ldots X_c$, where X_i represents the ith cluster of such partition U, the Dunn's validation index, D, is defined as:

$$D(U) = \min_{1\le i\le c}\left\{\min_{\substack{1\le j\le c\\ j\neq i}}\left\{\frac{\delta(X_i, X_j)}{\max_{1\le k\le c}\{\Delta(X_k)\}}\right\}\right\}, \tag{3}$$

where $\delta(X_i, X_j)$ defines the *intercluster distance* between clusters X_i and X_j ; $\Delta(X_k)$ represents the *intracluster distance* ("diameter") 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 Dunn's validity index correspond to good cluster partitions. Therefore, the number of clusters that maximises D is taken as the optimal number of clusters, c.

The Davies-Bouldin validation index, DB, is defined as:

$$DB(U) = \frac{1}{c}\sum_{i=1}^{c}\max_{i\neq j}\left\{\frac{\Delta(X_i)+\Delta(X_j)}{\delta(X_i, X_j)}\right\}, \tag{4}$$

where $\Delta(X_i)$, $\Delta(X_j)$ and $\delta(X_i, X_j)$ are defined as above. In this case, small index values correspond to good clusters, that is to say, the clusters are compact and their centers 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 [4, 5]. Six intercluster, δ_i, $1 \le i \le 6$ (*single, complete, average, centroid, average of centroids linkage* and *Hausdorff metrics*); and three intracluster, Δ_j, $1 \le j \le 3$ (*complete, average* and *centroid diameters*) distances are used for the implementation of the validity indices. Thirty-six indices based on equations (3) and (4) were calculated. These indices consist of different combinations of intercluster and intracluster distance methods. Thus, for example, D_{13}, represents a Dunn's index based on an intercluster distance, δ_1, and an intracluster distance Δ_3; and DB_{31}, represents a Davies-Bouldin validity index based on an intercluster distance, δ_3, and an intracluster distance Δ_1.

It has been shown that using different intercluster/intracluster distance combinations may produce validation indices of different scale ranges [4]. 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. To overcome this problem the following normalisation technique has been applied. Given a cluster configuration consisting of c clusters, for any partition $U_c \leftrightarrow X$: $X_1 \cup ... \cup X_c$, the normalised Dunn's indices - D_{ij}^*, are calculated as:

$$D_{ij}^*(U_c) = \left(D_{ij}(U_c) - \overline{D}_{ij}\right)\Big/\sigma D_{ij}, \tag{5}$$

$$\overline{D}_{ij} = \frac{1}{n}\sum_k D_{ij}(U_k), \tag{6}$$

where i reflects the selection of the intercluster distance calculation method (i = 1,…, 6), j is the selection of the 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} is the standard deviation of $D_{ij}(U_c)$ across all values of c. The normalised Davis-Bouldin indices may be calculated by equation (5) using the Davis-Bouldin index instead of the Dunn's index.

III. RESULTS

The cluster validity methods have been implemented, using the well-known Euclidean distance between samples. Table I depicts the global silhouette values, *GSu*, for each partition, and the silhouette values, *S*, for each number of clusters, *c*, for c = 2 to c = 6. In this case $c = 2$ is suggested as the best clustering configuration for the examined data set.

TABLE I
GLOBAL SILHOUETTE VALUES FOR EACH PARTITION, *GSu*, AND THE SILHOUETTE VALUES, *S*, FOR EACH CLUSTER DEFINING A PARTITION

c	*GSu*	S_1	S_2	S_3	S_4	S_5	S_6
2	0.31	0.42	0.16				
3	0.25	0.25	0.13	0.36			
4	0.26	0.18	0.23	0.38	0.23		
5	0.29	0.31	0.21	0.37	0.22	0.27	
6	0.19	0.22	0.60	0.01	0.56	0.14	0.33

The normalised values of the eighteen Dunn's and Davies-Bouldin validity indices and their average indices at each number of clusters, c, for c = 2 to c = 6 are shown in Tables II and III respectively. An examination of these results indicates that c = 2 represents the most appropriate partition for the data under analysis.

TABLE II
NORMALISED DUNN'S VALUES USING 3 TYPES OF INTRACLUSTER MEASURES AND 6 TYPES OF INTERCLUSTER MEASURES

Validity index	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$
D_{11}	1.17	0.37	-1.50	0.32	-0.36
D_{21}	1.71	-0.07	-0.22	-0.64	-0.78
D_{31}	1.70	0.03	-0.30	-0.67	-0.76
D_{41}	1.62	0.17	-0.23	-0.59	-0.97
D_{51}	1.70	0.05	-0.34	-0.76	-0.65
D_{61}	1.77	-0.57	-0.21	-0.59	-0.40
D_{12}	1.37	0.46	-1.18	0.05	-0.71
D_{22}	1.69	-0.02	-0.19	-0.64	-0.84
D_{32}	1.66	0.11	-0.24	-0.66	-0.86
D_{42}	1.60	0.20	-0.20	-0.60	-1.00
D_{52}	1.66	0.12	-0.27	-0.73	-0.78
D_{62}	1.76	-0.42	-0.17	-0.60	-0.57
D_{13}	1.25	0.20	-1.50	0.31	-0.27
D_{23}	1.72	-0.15	-0.17	-0.65	-0.75
D_{33}	1.72	-0.08	-0.24	-0.69	-0.71
D_{43}	1.65	0.08	-0.18	-0.61	-0.94
D_{53}	1.72	-0.07	-0.28	-0.78	-0.60
D_{63}	1.75	-0.64	-0.16	-0.60	-0.35
Average	1.62	-0.01	-0.42	-0.51	-0.68

TABLE III.
NORMALISED DAVIES-BOULDIN VALUES USING 3 TYPES OF INTRACLUSTER MEASURES AND 6 TYPES OF INTERCLUSTER MEASURES

Validity index	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$
DB_{11}	0.94	0.33	0.86	-1.13	-1.00
DB_{21}	-1.64	-0.26	0.49	0.60	0.80
DB_{31}	-0.03	1.03	0.85	-1.42	-0.43
DB_{41}	-1.54	0.52	0.99	-0.43	0.46
DB_{51}	-0.19	1.06	1.00	-0.78	-1.11
DB_{61}	-1.49	1.23	0.26	0.31	-0.32
DB_{12}	0.60	-0.25	1.29	-1.36	-0.28
DB_{22}	-1.33	-0.46	-0.07	0.55	1.31
DB_{32}	-1.07	-0.55	-0.07	0.10	1.59
DB_{42}	-1.34	-0.45	0.15	0.25	1.39
DB_{52}	-1.33	-0.51	0.14	0.33	1.36
DB_{62}	-1.70	0.38	-0.05	0.55	0.81
DB_{13}	0.93	0.11	1.00	-1.13	-0.92
DB_{23}	-1.51	-0.31	0.03	0.77	1.01
DB_{33}	-1.52	-0.11	0.93	-0.17	0.88
DB_{43}	-1.65	-0.11	0.69	0.22	0.86
DB_{53}	-1.38	0.32	1.25	0.36	-0.55
DB_{63}	-1.66	0.99	0.09	0.50	0.09
Average	-0.94	0.16	0.55	-0.10	0.33

Another approach to estimate the optimal partition consists of the implementation of an aggregation method based on a weighed voting strategy. An example is shown in Table IV based on the Dunn's indices. This table was obtained from Table II by replacing the index values by weighed votes, whose values range from 1 to 5. Thus, for example, D_{11} represents the highest index value and suggests the partition $c = 2$ as the optimal partition, hence its weighed vote is equal to 5. On the other hand D_{11} represents the smallest index value for partition $c = 4$, hence its weighed vote is equal to 1. The average weighed vote for each cluster partition confirms that $c = 2$ represents the most appropriate prediction.

TABLE IV.
PREDICTING THE CORRECT NUMBER OF CLUSTERS BY WEIGHED VOTING TECHNIQUE. THE ENTRIES REPRESENT VOTE VALUES BASED ON DUNN'S VALIDATION INDEX USING 3 TYPES OF INTRACLUSTER AND 6 TYPES OF INTERCLUSTER MEASURES

Validity index	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$
D_{11}	5	4	1	3	2
D_{21}	5	4	3	2	1
D_{31}	5	4	3	2	1
D_{41}	5	4	3	2	1
D_{51}	5	4	3	1	2
D_{61}	5	2	4	1	3
D_{12}	5	4	1	3	2
D_{22}	5	4	3	2	1
D_{32}	5	4	3	2	1
D_{42}	5	4	3	2	1
D_{52}	5	4	3	2	1
D_{62}	5	3	4	1	2
D_{13}	5	3	1	4	2
D_{23}	5	4	3	2	1
D_{33}	5	4	3	2	1
D_{43}	5	4	3	2	1
D_{53}	5	4	3	1	2
D_{63}	5	1	4	2	3
Average	5.00	3.61	2.83	2.00	1.56

This voting strategy may also be applied to fuse the results originating from different validation methods. An example is depicted in the Table V for three validation techniques. This table was obtained from Tables I-III by calculating the average weighed vote for each technique. Thus, after computing all validity indices, the average weighed vote for each cluster partition has been calculated, and $c = 2$ is suggested as the optimal partition.

TABLE V.
PREDICTING THE CORRECT NUMBER OF CLUSTERS FOR MEDULLOBLASTOMAS DATA BY AGGREGATION OF MULTIPLE VALIDATION METHODS

Validation technique	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$
Silhouette	5.00	2.00	3.00	4.00	1.00
Dunn's	5.00	3.61	2.83	2.00	1.56
Davies-Bouldin	4.22	2.89	2.33	3.06	2.50
Average	4.74	2.83	2.72	3.02	1.69

The applied validation techniques confirm that the partition consisting of two clusters represents the most appropriate representation for the data set under consideration. This result also supports the choice of 140 genes (from the set of 7129 genes) as responsible for the *desmoplastic* and *classic medulloblastomas* distinction reported by Pomeroy and colleagues.

IV. DISCUSSION AND CONCLUSION

Several clustering techniques have been proposed to support the analysis of gene expression data. Cluster validity indices represent useful tools to guide unsupervised data analysis. They are particularly relevant for the estimation of robust clustering partitions in different applications, which may require the definition of the number of clusters beforehand. In this research three validation indices were applied to a *desmoplastic* and *classic medulloblastomas* data set, using different intracluster and intercluster distances. The combination of these methods may be used for cluster

evaluation tasks. It has been shown how these methods may support the prediction of the optimal cluster partition. The results also suggest that the normalisation of index values and a voting strategy may improve the prediction procedure. The normalisation scheme may represent a more robust mechanism to predict the correct number of clusters. Moreover, it highlights subtle differences between index values originating from different clustering configurations. The advantage of a weighed voting approach lies in an aggregation of multiple validation methods in order to improve the estimation of the most adequate clustering partition. This validation framework has been successfully tested on other data sets and clustering techniques such as the *Kohonen Self-Organising Map* algorithm [15].

These results suggest that a systematic validation approach may significantly support genome expression analyses for knowledge discovery applications. Current and future work include the comparison, combination and estimation of results obtained from different clustering algorithms, and the analysis of more complex data sets.

ACKNOWLEDGMENT

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

REFERENCES

[1] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. USA*, vol. 95, pp. 14863-8, 1998.

[2] 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, pp. 16-22, 2001.

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

[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, pp. 319-320, 2002.

[6] M. Halkidi, Y. Batistakis, M. Vazirgiannis, "On clustering validation techniques", *JIIS*, vol. 17, pp.107-145, 2001.

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

[8] J. Dunn, "Well separated clusters and optimal fuzzy partitions", *J.Cybernetics,* vol. 4, pp. 95-104, 1974.

[9] D.L. Davies, D.W. Bouldin, "A cluster separation measure", *IEEE Transactions on Pattern Recognition and Machine Intelligence*, vol. 1, No. 2, pp. 224-227, 1979.

[10] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y.H. Kim, L.C. Goumnerova, P. McL. Black, C. Lau, J.C. Allen, D. Zagzag, J.M.Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D.N. Louis, J.P. Mesirov, E.S. Lander and T.R. Golub, "Gene expression-based classification and outcome prediction of central nervous system embryonal tumors," *Nature*, vol 415, pp. 436-42, 24 January 2002.

[11] J. Quackenbush, "Computational analysis of microarray data", *Nature Reviews Genetics*, vol. 2, pp. 418-427, 2001.

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

[13] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, San Mateo, California: Morgan Kaufmann Publisher, 2000.

[14] J.C. Bezdek, N.R. Pal, "Some new indexes of cluster validity", *IEEE Transactions on Systems, Man and Cybernetics*, vol. 28, part B, pp. 301-315, 1998.

[15] N. Bolshakova, F. Azuaje, "Cluster validation techniques for genome expression data classification", *Signal Processing*, 2002, in press.