Validation of clustering techniques for microarray gene expression data

by Nadia Bolshakova and Pádraig Cunningham

Recent advances in microarray technology have enabled the measurement of the simultaneous expression of thousands of genes under multiple experimental conditions. The methods implemented in this research may contribute to the validation of clustering results and the estimation of the number of clusters. For instance, these tools may be used for the identification of new tumour classes using gene expression profiles. The results show that this estimation approach may represent an effective tool to support biomedical knowledge discovery and healthcare applications.

One of our major tasks is to advance data analysis and integration capabilities in genomic expression pattern discovery and classification. It has consisted of the implementation of algorithms and tools to organise and categorise genome expression data. It has integrated and improved a number of machine learning techniques, which may aid in the identification of relevant features for diagnostic, prognostic and system biology studies. These tools may also be applied to other information management domains such as biomedical informatics. Moreover, automated discovery solutions may assist the design of novel techniques for intelligent information retrieval and knowledge and meta-knowledge representation, which are crucial aspects for the integration of information over the global network.

An important step in the analysis of gene expression data is the detection of samples or genes with similar expression patterns. The accurate classification of tumours is essential for a successful diagnosis and treatment of cancer. One of the problems associated with cancer tumour classification is the identification of unknown classes using gene expression profiles. Several clustering algorithms have been developed for gene expression data. Also techniques to systematically evaluate the quality of the clusters have been presented. 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. It is useful not to rely on one single clustering or validation method, but to apply a variety of approaches. Therefore, a combination of these methods may be successfully used for the estimation of the number of clusters. It has been shown that these methods may support the prediction of the optimal partition and computational diagnosis.

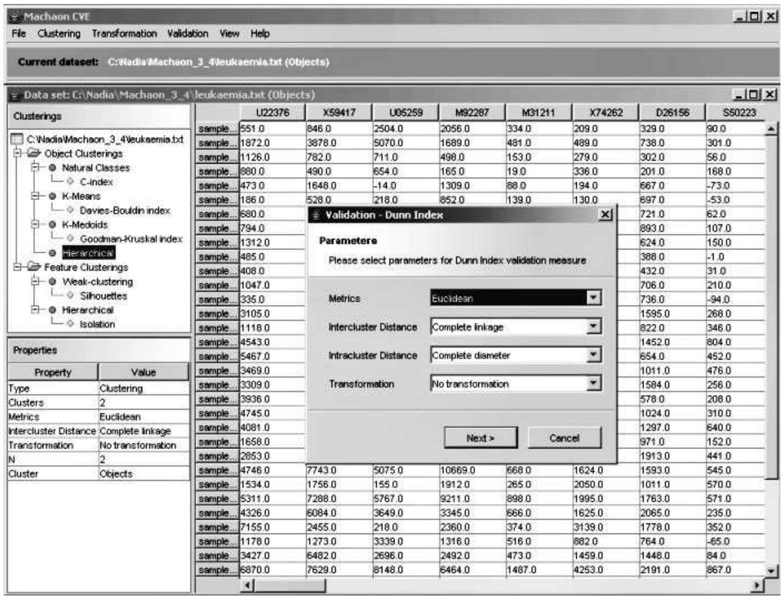
We have developed the Machaon Cluster Validation Environment (Machaon CVE) for the application of different clustering and validation algorithms to experiment on gene expression data. This tool may improve the quality of the data analysis results, and may support the prediction of the number of relevant clusters in the microarray datasets. The major stages of the system can be summarised as follows:

- *Clustering*. In this step we extract clusters that correspond to the pre-defined number of clusters for a particular dataset. It offers a number of the well-established clustering methods that are available in the literature as well as some recently developed ensemble techniques.
- *Validation of clustering techniques.* The clustering methods can find a partition in a dataset, based on certain assumptions. Thus, an algorithm may result in different clustering schemes for a dataset assuming different parameter values. Machaon evaluates the results of clustering algorithms based on quality indices and selects the clustering scheme that best fits the data. The definition of these indices is based on two fundamental criteria of clustering quality: cluster compactness and isolation.

To support biomedical knowledge representation in gene expression data analysis, a new markup language for microarray data clustering and cluster validity assessment has been developed. cluML, a free, open, XML-based format has been designed to address some of the limitations observed in traditional formats, such as inability to store multiple clustering (including biclustering) and validation results within a dataset.

To enhance the predictive reliability and biological relevance of the validation results, a knowledge-driven cluster validity assessment approach for microarray data clustering has been implemented. It consists of validity indices that incorporate similarity knowledge originating from the Gene Ontology (GO), which is a structured, shared vocabulary that allows the annotation of gene products across different model organisms.

The methods performed in this research may bring contribute to the evaluation of clustering outcome and the prediction of optimal cluster partitions. The described estimation approach represents an effective tool to support biomedical knowledge discovery in gene expression data analysis. Despite the fact that Machaon CVE was developed for DNA microarray expression analysis applications, it may be effectively used for clustering and validating of other biomedical and physical data with no limitations.



Current dataset: C:\Nadia\Machaon_3_4\leukaemia.txt (Objects)

- C:\Nadia\Machaon_3_4\leukaemia.txt
  - Object Clusterings
    - Natural Classes
      - C-index
    - K-Means
      - Davies-Bouldin index
    - K-Medoids
      - Goodman-Kruskal index
    - Hierarchical
  - Feature Clusterings
    - Weak-clustering
      - Silhouettes
    - Hierarchical
      - Isolation

Property	Value
Type	Clustering
Clusters	2
Metrics	Euclidean
Intercluster Distance	Complete linkage
Transformation	No transformation
N	2
Cluster	Objects

Parameters

Please select parameters for Dunn Index validation measure

Metrics: Euclidean

Intercluster Distance: Complete linkage

Intracluster Distance: Complete diameter

Transformation: No transformation

	U22376	X59417	U05259	M92287	M31211	X74262	D26156	S50223
sample...	551.0	846.0	2504.0	2056.0	334.0	209.0	329.0	90.0
sample...	1872.0	3878.0	5070.0	1689.0	481.0	489.0	738.0	301.0
sample...	1126.0	782.0	711.0	498.0	153.0	279.0	302.0	56.0
sample...	880.0	490.0	654.0	165.0	19.0	336.0	201.0	168.0
sample...	473.0	1648.0	-14.0	1309.0	88.0	194.0	667.0	-73.0
sample...	186.0	528.0	218.0	852.0	139.0	130.0	697.0	-53.0
sample...	680.0						721.0	62.0
sample...	794.0						893.0	107.0
sample...	1312.0						624.0	150.0
sample...	485.0						388.0	-1.0
sample...	408.0						432.0	31.0
sample...	1047.0						706.0	210.0
sample...	335.0						736.0	-94.0
sample...	3105.0						1595.0	268.0
sample...	1118.0						822.0	346.0
sample...	4543.0						1452.0	804.0
sample...	5467.0						654.0	452.0
sample...	3469.0						1011.0	476.0
sample...	3309.0						1584.0	256.0
sample...	3936.0						578.0	208.0
sample...	4745.0						1024.0	310.0
sample...	4081.0						1297.0	640.0
sample...	1658.0						971.0	152.0
sample...	2853.0						1913.0	441.0
sample...	4746.0	7743.0	5075.0	10669.0	668.0	1624.0	1593.0	545.0
sample...	1534.0	1756.0	155.0	1912.0	265.0	2050.0	1011.0	570.0
sample...	5311.0	7288.0	5767.0	9211.0	898.0	1995.0	1763.0	571.0
sample...	4326.0	6084.0	3649.0	3345.0	666.0	1625.0	2065.0	235.0
sample...	7155.0	2455.0	218.0	2360.0	374.0	3139.0	1778.0	352.0
sample...	1178.0	1273.0	3339.0	1316.0	516.0	882.0	764.0	-65.0
sample...	3427.0	6482.0	2696.0	2492.0	473.0	1459.0	1448.0	84.0
sample...	6870.0	7629.0	8148.0	6464.0	1487.0	4253.0	2191.0	867.0

Figure 1. A screenshot of the Machaon CVE.

Links:

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