A knowledge-driven approach to cluster validity assessment

Nadia Bolshakova[a, *]*, Francisco Azuaje*[b] *and Pádraig Cunningham*[a]

[a]*Department of Computer Science, Trinity College Dublin, Dublin 2, Ireland*
[b]*School of Computing and Mathematics, University of Ulster, Jordanstown, BT37 0QB, U.K*

ABSTRACT

Summary: This paper presents an approach to assessing cluster validity based on similarity knowledge extracted from the Gene Ontology.

Availability: The program is freely available for non-profit use on request from the authors.

Supplementary information: http://www.cs.tcd.ie/Nadia.Bolshakova/GOtool.html

The automated integration of background knowledge is fundamental to support the generation and validation of hypotheses about the function of gene products. One such source of prior knowledge is the *Gene Ontology* (GO), which is a structured, shared vocabulary that allows the annotation of gene products across different model organisms. The GO comprises three independent hierarchies: molecular function (MF), biological process (BP) and cellular component (CC). Researchers can represent relationships between gene products and annotation terms in these hierarchies. Previous research has applied GO information to detect overrepresented functional annotations in clusters of genes obtained from expression analyses. It has also been suggested to assess gene sequence similarity and expression correlation. For additional information on the GO and its applications, the reader is referred to its website (http://www.geneontology.org) and (Wang *et al*., 2004).

Topological and statistical information extracted from the GO in relation to a set of annotated gene products may be used to measure similarity between them. Different GO-driven similarity assessment methods may be then implemented to perform clustering or to quantify the quality of the resulting clusters. Cluster validity assessment may consist of *data- and knowledge-driven methods*, which aim to estimate the *optimal cluster partition* from a collection of candidate partitions. Data-driven methods mainly include statistical tests or validity indices applied to the data clustered. Knowledge-driven methods are proposed to enhance the predictive reliability and biological relevance of the results. A data-driven, cluster validity assessment platform was previously reported by (Bolshakova and Azuaje, 2003).

Traditional GO-based cluster description methods have consisted of statistical analyses of the enrichment of GO terms in a cluster. The application of GO-based similarity to perform clustering and validate clustering outcomes has not been widely investigated. A recent contribution by Speer *et al.* (2004) presented an algorithm that incorporates GO annotations to cluster genes. They applied the Davies-Bouldin index (Bolshakova and Azuaje, 2003) to estimate the quality of the clusters.

We implemented a knowledge-driven cluster validity assessment system for microarray data clustering. It consists of validity indices that incorporate similarity knowledge originating from the GO (we used only non-IEA annotations and the May 2004 release version). A well-known gene expression dataset from the yeast cell cycle (Cho *et al*., 1998) was analysed to illustrate its application. Several cluster partitions, obtained with the *k*-means algorithm, were analysed to estimate the optimum number of clusters for this dataset. An *information content* technique proposed by Resnik (1995) was implemented to measure similarity between gene products based on the GO. Detailed descriptions on this and other GO-based similarity assessment techniques are presented in (Wang *et al*., 2004) and the Supplementary Section.

This research applies two approaches to calculating cluster validity indices. The first approach process overall similarity values, which are calculated by taking into account the combined annotations originating from the three GO hierarchies. The second approach is based on the calculation of independent similarity values, which originate from each of these hierarchies. The second approach allows one to estimate the effect of each of the hierarchies on the validation process.

We applied the C-index (Hubert and Schultz, 1976), which is an effective cluster validity estimator for different types of clustering applications. Clustering was performed with the *Machaon CVE* tool (Bolshakova and Azuaje, 2003). The data comprised 64 genes described by their expression values during the yeast cell cycle (Cho *et al*., 1998). Previous research has shown that disjoint clusters of genes are significantly expressed in each of the five cell cycle stages: early G1, late G1, S, G2, M.

Figure 1(a) shows the predictions made by the validity indices at each number of clusters, *c*, for *c* =2 to 6. The bold entries correspond to the optimal values of the indices. The validity indices based on similarity information from the MF, BP and the combined hierarchies indicated that the optimal number of clusters is *c* = 5, which is consistent with the cluster structure expected (Cho *et al*., 1998). Only the method based on the CC hierarchy suggested the partition with two clusters as the optimal partition, which confirms that cellular localization information does not adequately reflect relevant functional relationships in this dataset.

The *Machaon CVE* (Bolshakova and Azuaje, 2003) has been updated to support this technique. It aims to partition samples or genes into groups characterised by similar expression patterns, and to evaluate the quality of the clusters obtained. Figure 1(b) depicts screenshots from the *Machaon CVE.* Future research will include the comparison and combination of different data- and knowledge-driven cluster validity indices. This study contributes to the development of techniques for facilitating the statistical and biological validity assessment of data mining results in functional genomics.

ACKNOWLEDGEMENTS

This research is partly based upon works supported by the Science Foundation Ireland under Grant No. S.F.I.-02IN.1I111.

REFERENCES

Bolshakova,N. and Azuaje,F. (2003) Machaon CVE: cluster validation for gene expression data. *Bioinformatics*, **19**, 2494-5.

Cho,R.J., Campbell,M.J., Winzeler,E.A., Steinmetz,L., Conway,A., Wodicka,L., Wolfsberg,T.G., Gabrielian,A.E., Landsman,D., Lockhart,D.J. and Davis,R.W. (1998) A genomewide transcriptional analysis of the mitotic cell cycle, *Molecular Cell* , 2, 65-73.

Hubert,L. and Schultz,J. (1976) Quadratic assignment as a general data-analysis strategy. *British Journal of Mathematical and Statistical Psychologie*. 190-241.

Resnik,P. (1995) Using information content to evaluate semantic similarity in a taxonomy, *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 448-453.

Speer,N., Spieth,C. and Zell,A. (2004) A memetic clustering algorithm for the functional partition of genes based on the gene ontology, in *Proceedings of the 2004 IEEE Symposium on*

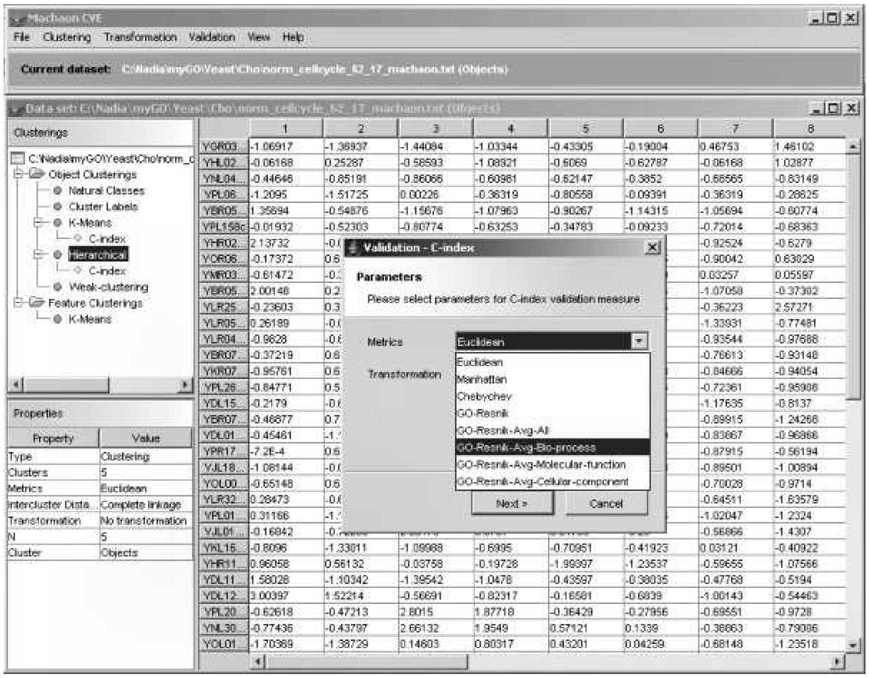
Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2004), San Diego, USA, pp. 252-259, IEEE Press.

Wang, H., Azuaje,F., Bodenreider,O. and Dopazo,J. (2004) Gene expression correlation and gene ontology-based similarity: An assessment of quantitative relationships", in *Proc. of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, IEEE Press, October 7-8, La Jolla-California, pp. 25 – 31.

(a)

Validity indices based on:	**c=2**	**c=3**	**c=4**	**c=5**	**c=6**
Combined hierarchies	0.506	0.436	0.385	**0.347**	0.348
Biological process	0.496	0.331	0.223	**0.132**	0.141
Molecular function	0.505	0.326	0.231	**0.174**	0.191
Cellular component	**0.507**	0.648	0.674	0.71	0.732

(b)



Current dataset: C:\Nadia\myGO\Yeast\Cho\norm_cellcycle_62_17_machaon.txt (Objects)

Data set: C:\Nadia\myGO\Yeast\Cho\norm_cellcycle_62_17_machaon.txt (Objects)

Clusterings

- C:\Nadia\myGO\Yeast\Cho\norm_c
  - Object Clusterings
    - Natural Classes
    - Cluster Labels
    - K-Means
      - C-index
    - Hierarchical
      - C-index
    - Weak-clustering
  - Feature Clusterings
    - K-Means

Properties

Property	Value
Type	Clustering
Clusters	5
Metrics	Euclidean
Intercluster Dista...	Complete linkage
Transformation	No transformation
N	5
Cluster	Objects

	1	2	3	4	5	6	7	8
YGR03...	-1.06917	-1.36937	-1.44084	-1.03344	-0.43305	-0.19004	0.46753	1.46102
YHL02...	-0.06168	0.25287	-0.58593	-1.08921	-0.6069	-0.62787	-0.06168	1.02877
YNL04...	-0.44646	-0.85191	-0.86066	-0.60981	-0.62147	-0.3852	-0.66565	-0.83149
YPL06...	-1.2095	-1.51725	0.00226	-0.36319	-0.80558	-0.09391	-0.36319	-0.28625
YBR05...	1.35694	-0.54876	-1.15676	-1.07963	-0.90267	-1.14315	-1.05694	-0.60774
YPL158c	-0.01932	-0.52303	-0.80774	-0.63253	-0.34783	-0.09233	-0.72014	-0.68363
YHR02...	2.13732	-0.					-0.92524	-0.6279
YOR06...	-0.17372	0.6					-0.90042	0.63029
YMR03...	-0.61472	-0.					0.03257	0.05597
YBR05...	2.00148	0.2					-1.07058	-0.37302
YLR25...	-0.23603	0.3					-0.36223	2.57271
YLR05...	0.26189	-0.					-1.33931	-0.77481
YLR04...	-0.9828	-0.6					-0.93544	-0.97688
YBR07...	-0.37219	0.6					-0.76613	-0.93148
YKR07...	-0.95761	0.6					-0.84666	-0.94054
YPL26...	-0.84771	0.5					-0.72361	-0.95908
YDL15...	-0.2179	-0.6					-1.17635	-0.8137
YBR07...	-0.46877	0.7					-0.89915	-1.24266
YDL01...	-0.45461	-1.					-0.83867	-0.96868
YPR17...	-7.2E-4	0.6					-0.87915	-0.56194
YJL18...	-1.08144	-0.					-0.89501	-1.00894
YOL00...	-0.65148	0.6					-0.70028	-0.9714
YLR32...	0.28473	-0.6					-0.64511	-1.63579
YPL01...	0.31166	-1.					-1.02047	-1.2324
YJL01...	-0.16842	-0.7					-0.56866	-1.4307
YKL16...	-0.8096	-1.33011	-1.09988	-0.6995	-0.70951	-0.41923	0.03121	-0.40922
YHR11...	0.96058	0.56132	-0.03758	-0.19728	-1.99397	-1.23537	-0.59655	-1.07566
YDL11...	1.58028	-1.10342	-1.39542	-1.0478	-0.43597	-0.38035	-0.47768	-0.5194
YDL12...	3.00397	1.52214	-0.56691	-0.82317	-0.16581	-0.6839	-1.00143	-0.54463
YPL20...	-0.62618	-0.47213	2.8015	1.87718	-0.36429	-0.27956	-0.69551	-0.9728
YNL30...	-0.77436	-0.43797	2.66132	1.9549	0.57121	0.1339	-0.38863	-0.79086
YOL01...	-1.70389	-1.38729	0.14603	0.80317	0.43201	0.04259	-0.68148	-1.23518

Validation - C-index

Parameters

Please select parameters for C-index validation measure

Metrics: Euclidean

Transformation

Euclidean
Manhattan
Chebychev
GO-Resnik
GO-Resnik-Avg-All
GO-Resnik-Avg-Bio-process
GO-Resnik-Avg-Molecular-function
GO-Resnik-Avg-Cellular-component
Next >
Cancel

Figure 1. (a) C-index values for expression clusters originating from yeast data. Bold entries represent the optimal number of clusters, *c*, predicted by each method. Validity indices used Resnik's similarity metric. The first approach process overall similarity values, which are calculated by taking into account the combined annotations originating from the three GO hierarchies. The other indices are based on the calculation of independent similarity values, which originate from each of these hierarchies. (b) Screenshots from the *Machaon CVE.*