Comparison of the Data-based and Gene Ontology-based Approaches to Cluster Validation Methods for Gene Microarrays

Nadia Bolshakova, Anton Zamolotskikh and Pádraig Cunningham
Department of Computer Science, Trinity College Dublin, Ireland
{nadia.bolshakova, zamolota, padraig.cunningham}@cs.tcd.ie

Abstract

The paper presents a comparison of the data-based and Gene Ontology (GO)-based approaches to cluster validation methods for gene microarray analysis. We apply a homogeneous approach to obtaining metrics from different GO-based similarity measures and a normalization of validation index values, that allows us to compare them to each other as well as to data-based validation indices. The results show strong correlation between both GO-based and data-based validation indices. The results suggest that this may represent an effective tool to support biomedical knowledge discovery tasks based on gene expression data.

1. Introduction

In recent years, gene microarray technology, also known as gene chips, has significantly impacted on genomic and post-genomic studies. The microarrays allow the measurement of the expression of thousands of genes in parallel and under multiple experimental conditions. Biomedical research such as disease diagnosis and drug discovery benefits from this DNA microarray technology.

Processing of the tremendous amount of data obtained from microarray experiments requires advanced data mining methods. Several approaches have been applied to analyse gene expression data including supervised [1] and unsupervised [2] learning. Unsupervised learning, covers clustering which is aimed at detecting samples or genes with similar expression patterns. Various methods have been applied, such as self-organizing maps, k-means, hierarchal clustering and so on.

Since different clustering algorithms or different runs of the same algorithm generate different solutions for the same dataset, the question of choosing an appropriate algorithm with appropriate parameters for the dataset becomes a critical problem. Methods for the systematic evaluation of the quality of the clusters based on the data have been also proposed [3, 4, 5]. Several data-based cluster validity indices are described in the literature, such as Dunn's index [6], Rand index [7], Figure of Merit [8], Silhouette index [9] or Davies-Bouldin index [10] and many of them have already been used with gene expression data.

Data-driven methods mainly include statistical tests or validity indices applied to the data clustered. Even though the results of the approaches are valuable, the methods do not apply external biological information. In our previous research, a knowledge-driven cluster validity assessment system for microarray data clustering had been implemented [11]. Unlike traditional methods that only use (gene expression) data-derived indices, our method consisted of validity indices that incorporate similarity knowledge originating from the GO and a GO-driven annotation database [11, 12]. A number of tools has been developed for ontological analysis of gene expression data [13] recently.

This paper is devoted to the comparison of the data-based and Gene Ontology-based approaches to cluster validation methods for gene microarrays to estimate the optimal cluster partition from a collection of candidate partitions. We show that there is strong agreement between the two approaches.

2. Biological Ontologies

The Gene Ontology Consortium [14] initiated the Gene Ontology (www.geneontology.org) project in 2004. The ontology is intended for annotating gene products with a consistent, controlled and structured vocabulary. The GO is independent from any biological species and is rapidly growing. The GO represents terms in a Directed Acyclic Graph (DAG), comprising three independent hierarchies: molecular

function (MF), biological process (BP) and cellular component (CC). Terms are allowed to have multiple parents as well as multiple children. Two different kinds of relationship exist: the is-a" and the "part-of" relationships. It is possible to represent relationships between gene products and annotation terms encoded in these hierarchies. Previous research has applied GO information to detect over-represented functional annotations in clusters of genes obtained from expression analyses [15, 16, 17].

Recently, new ontologies covering other biological or medical aspects are being developed. For instance, the Protein Ontology project [18], which defines a common structured vocabulary for researchers who need to share knowledge in the proteomics domain. It includes concepts (type definitions), which are data descriptors for proteomics data and the relations among these concepts [18].

For a review on biological ontologies the reader is referred to [19].

3. Methods

The dataset described the response of human fibroblasts to serum on cDNA microarrays in order to study growth control and cell cycle progression. These data were obtained from a study published by Iyer and colleagues [20]. The authors found 517 genes whose expression levels varied significantly [20]. The original data and experimental methods are available at http://genome-www.stanford.edu/serum. We used these 517 genes for which the authors provide National Center for Biotechnology Information (NCBI) accession numbers. After mapping with GeneLynx (http://www.genelynx.org), 63 genes showed one or more links to the GO database. These genes were used for the clustering and cluster validation methods.

Several cluster partitions (with numbers of clusters from two to ten clusters), obtained with the k-means algorithm, were analysed to estimate the optimum number of clusters for this dataset. Clustering and further validation were performed with the Machaon CVE tool [5].

Cluster validation was performed using two validity indices: the Dunn's index [6] and the Silhouette index [9], whose data-driven versions have been shown to be effective cluster validity estimators for different types of clustering applications [4].

For the data-based cluster validity approach, the distance between genes was calculated using the well-known Euclidean metric [21].

To calculate GO-based similarity measures we use two approaches: Wu and Palmer [22] and Resnik's [23] methods.

In [12] a specific approach to transformation of the similarity values into metrics (dissimilarity values) for each of the methods is proposed. Instead of that we suggest to consider a similarity measure to be a kernel function, i.e. an inner product in some latent feature space.

The following relation between an inner product and metric holds true:

$$d^2(x,y) = k(x,x) + k(y,y) - 2 \cdot k(x,y), \qquad (1)$$

where x and y are two vectors, k is a kernel function, assuming k is in fact a dot product in the complete metric space, in which d is a metric.

The difference to the methods proposed in [12] in terms of the calculation of the GO-based metrices is in the calculation of the similarity measures themselves. Both Wu and Palmer and Resnik's similarity measures between two gene products are now calculated as follows:

$$sim(g_k, g_m) = \max_{i,j} \left(sim(t_{ki}, t_{mj}) \right), \qquad (2)$$

where g_k and g_m are gene products, t_{ki} and t_{mj} are terms directly associated with those gene products and *sim* denotes a similarity measure between terms.

In [12] an average value is calculated instead of maximum. It means that we applied both Wu and Palmer and Resnik similarity measures to terms and then had to introduce an artificial technique (averaging) to calculate the similarity measures between gene products. In this paper we avoided that considering gene products and their associations with terms as a part of GO. Indeed calculating the maximum of similarities in each pair of terms associated with the gene products is equivalent to calculating of Wu and Palmer and Resnik's similarity measure between the gene products directly.

Finally the metric is calculated from each of the similarity measures as follows:

$$d(g_k, g_m) = \sqrt{sim(g_k, g_k) + sim(g_m, g_m) - 2 \cdot sim(g_k, g_m)}. \qquad (3)$$

4. Results

The clustering algorithms were applied to produce different partitions consisting of 2 to 10 clusters each. Then, the validity indices were computed for each of these partitioning results. One data-based and two GO-based similarity assessment techniques introduced

above were used for all cases to calculate the distances between the genes. The validity values obtained were normalized to values between zero and one.

Table 1 shows normalized indices for three validation methods applied to give nine partitionings of the dataset.

Table 1. Normalised validity indices for nine *k-mean* partitionings (number of clusters equals from two to ten) for the fibroblasts data.

	Dunn's validity indices			**Silhouette's validity indices**		
	Euclidean	**Wu-Palmer**	**Resnik**	**Euclidean**	**Wu-Palmer**	**Resnik**
2	1	1	1	1	1	1
3	0.87	0.57	0.79	0.91	0.6	0.76
4	0.65	0.3	0.68	0.89	0.44	0.56
5	0.46	0.26	0.54	0.77	0.01	0.07
6	0.16	0.26	0.52	0.26	0.03	0.09
7	0.14	0.05	0.36	0.2	0.01	0.04
8	0.08	0.05	0.25	0.17	0	0.01
9		0.12	0.34	0	0.01	0.03
10	0	0	0	0.03	0.02	0

The results show strong correlation between two GO-based validity indices as well as correlation between each of them and the data-based validity index. For instance, Figure 1 depicts the value of the Wu and Palmer metric-based Dunn's validation index as a function of the value of the data-based Dunn's index.

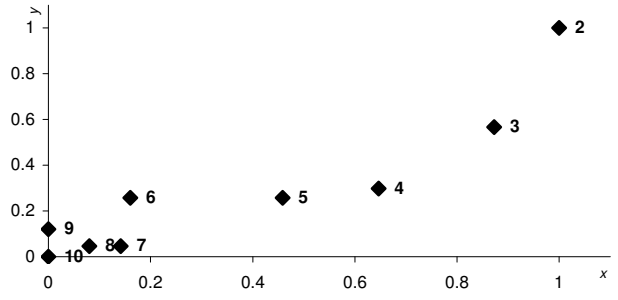

Figure 1. Correlation between the data-based Dunn's indices (*x*) and the GO-based Wu and Palmer (*y*) metrics. Labels indicate numbers of clusters in corresponding partinitionings.

The Wu-Palmer-based indices (both Silhouette and Dunn's) tend to be more sensitive than the data-based indices, producing lower values for partitionings of average quality, while for the Resnik-based indices this is not always the case. For instance, Figure 2 depicts all three Dunn's indices for all nine partitionings. For partitioning of two, three and four clusters the value of Wu-Palmer's-based Dunn's index is 0.57, 0.30 and 0.26 respectively, while Resnik-based Dunn's index produces 0.79, 0.68 and 0.54 for those partitionings. The data-based index is close to the Resnik-based one producing 0.87, 0.65 and 0.46 respectively.

In the lower band of index values the predictions of GO-based indices are less conclusive (Figure 2).

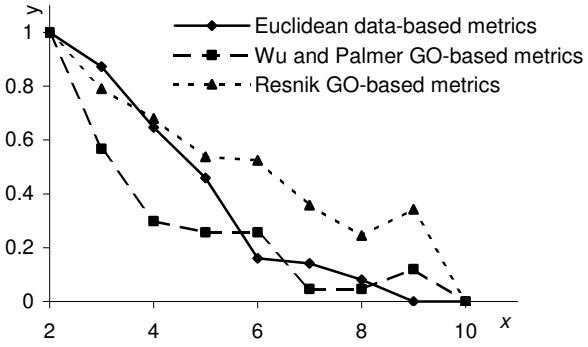

Figure 2. Normalized values of Dunn's index (*y*)-based on two GO-based and data-based matrices for two to ten (*x*) number of clusters.

5. Conclusion and Future Work

This research reports the comparison of different data- and knowledge-driven cluster validity indices.

Previous research has successfully applied validity indices using knowledge-driven methods [11, 12] to estimate the quality of the clusters. That work implemented a knowledge-driven cluster validity assessment system (based on similarity knowledge extracted from the Gene Ontology) for microarray data clustering [11, 12].

In this work we applied a more homogeneous approach to obtaining metrics from different GO-based similarity measures (i.e. Wu and Palmer and Resnik) and a normalization of validation index values, that allows us to compare them to each other as well as to data-based validation indices.

The results show strong correlation between both GO-based and data-based validation indices. It is also evident that the GO index that uses Resnik's similarity measure is far more sensitive to good partitioning than both the Wu and Palmer-based and data-based index.

Future work includes integration of data-based and GO-based validation methods into a common framework and research towards validation methods based on the comparison of GO sub-trees associated with clusters rather then on GO-based metrices between gene products.

The obtained results contribute to the development of techniques for the identification of optimal cluster partitions, which is a useful tool to support biological and biomedical knowledge discovery in microarray data analysis.

6. Acknowledgements

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

7. References

[1] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", *Science*, 1999, vol. 286, pp. 531-7.

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

[3] F. Azuaje. "A cluster validity framework for genome expression data", *Bioinformatics*, 2002, vol. 18, pp. 319-20.

[4] N. Bolshakova, and F. Azuaje, "Cluster validation techniques for genome expression data", *Signal Processing*, 2003a, vol. 83, pp. 825-33.

[5] N. Bolshakova, and F. Azuaje, "Machaon CVE: cluster validation for gene expression data", *Bioinformatic*s, 2003b, vol. 19, pp. 2494-5.

[6] J.C. Dunn, "Well separated clusters and optimal fuzzy partitions". *Journal of Cybernetics*, 1974, vol. 4, pp. 95–104.

[7] W. M. Rand, "Objective criteria for the evaluation of clustering methods", *Journal of the American Statistical Association*. 1971, vol. 66, pp. 846–50.

[8] K.Y. Yeung, D.R. Haynor, and W.L. Ruzzo, "Validating clustering for gene expression data", *Bioinformatics*, 2001, vol.17, pp. 309–18.

[9] P.J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational Applications in Math*. 1987, vol. 20, pp. 53–65.

[10] J.L. Davies, and D.W. Bouldin, "A cluster separation measure", IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, vol. 1, pp. 224–7.

[11] N. Bolshakova, F. Azuaje, and P. Cunningham, "A knowledge-driven approach to cluster validity assessment", *Bioinformatics*, 2005, vol. 21, pp. 2546-47.

[12] N. Bolshakova, F. Azuaj, and P. Cunningham. Incorporating biological domain knowledge into cluster validity assessment., submitted to the 4th European Workshop on Evolutionary Computation and Machine Learning in Bioinformatics (EvoBIO) 2006.

[13] P. Khatri, and S. Drăghici, "Ontological analysis of gene expression data: current tools, limitations, and open problems", *Bioinformatics*, 2005, vol. 21, pp. 3587-95.

[14] The Gene Ontology Consortium. The gene ontology (GO) database and informatics resource. Nucleic Acids Research, 2004, vol. 32, pp. D258–D261.

[15] F. Azuaje, and O. Bodenreider. Incorporating ontology-driven similarity knowledge into functional genomics: an exploratory study. Proceedings. Fourth IEEE Symposium on Bioinformatics and Bioengineering, 2004. BIBE 2004. pp. 317-24.

[16] N. Speer, C. Spieth, and A. Zel, A memetic clustering algorithm for the functional partition of genes based on the gene ontology. Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, 2004, pp. 252-259, IEEE Press.

[17] N. Speer, H. Fröhlich, C. Spieth, and A. Zell. Functional Grouping of Genes Using Spectral Clustering and Gene Ontology. Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN 2005), pp. 298-303, IEEE Press, 2005

[18] A.S. Sidhu, T.S. Dillon, E. Chang, and B.S. Sidhu, Protein ontology: vocabulary for protein data. Third International Conference on Information Technology and Applications. 2005, vol. 1, pp. 465-9.

[19] R. Stevens, C.A. Goble, and S. Bechhofer. "Ontology-based knowledge representation for bioinformatics", *Briefings in Bioinformatics*, 2000, vol. 1, pp.398-416.

[20] V.R. Iyer, M.B. Eisen, D.T. Ross, G. Schuler, T. Moore, J.C.F. Lee, J.M. Trent, L.M. Staudt, J Hudson Jr, M.S. Boguski, D. Lashkari, D. Shalon, D. Botstein, and P.O. Brown. "The transcriptional program in response of human .broblasts to serum". *Science*, 1999, vol. 283, pp. 83–7.

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

[22] Z. Wu, and M. Palmer. Verb semantics and lexical selection. Proceedings of the 32nd. Annual Meeting of the Association for Computational Linguistics. 1994. pp. 133 –138.

[23] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. Proceedings of the 14th International Joint Conference on Artificial Intelligence 1995, pp.448-53.