Predictive clustering relates gene annotations to phenotype properties extracted from images

Dragi Kocev[1], Bernard Ženko[1], Petra Paul[2], Coenraad Kuijl[2], Jacques Neefjes[2], and Sašo Džeroski[1]

[1] Department of Knowledge Technologies, Jozef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
{Dragi.Kocev,Bernard.Zenko,Saso.Dzeroski}@ijs.si
[2] Division of Cell Biology and Centre for Biomedical Genetics, the Netherlands Cancer Institute
Plesmanlaan 121, 1066 CX Amsterdam, Netherlands
{p.paul,c.kuijl,j.neefjes}@nki.nl

We address the task of grouping genes resulting in highly similar phenotypes upon siRNA mediated downregulation. The phenotypes are described by features extracted from images of the corresponding cellular assays. Both freely available general-purpose software, such as CellProfiler [4], and custom-made proprietary software can be used for this purpose. The features capture properties (such as intensity or texture) of the cells or their parts (nuclei, citoplasm, Golgi apparatus ...) in the images.

Clustering [6] produces partitions of the objects of interest (genes) into groups that are similar in a given feature space. In the context of the application of interest, this is a set of features extracted from the images of cellular assays. Besides finding clusters, e.g., groups of genes, we also aim to find descriptions/explanations for the clusters. The groups are explained in terms of a set of descriptors from a separate space, i.e., annotations of genes in terms of, e.g., the Gene Ontology [2] or the KEGG Pathway Database [5].

The typical approach to the problem at hand is to first cluster the phenotypes and elucidate the characteristics of the obtained clusters later on. Instead, we perform so-called constrained clustering, which yields both the clusters and their symbolic descriptions all in one step. The constrained clustering can be performed by using predictive clustering trees (PCTs) [3, 8, 9, 7], predictive clustering rules [10, 11] or ensembles of predicitve clustering rules [1]: These exemplify the paradigm of predictive clustering, which combines clustering and prediction.

In the presentation, we will describe the methods of building predictive clustering trees and ensembles of predictive clustering rules. We will also describe its application to the analysis of image data resulting from siRNA screens. These approaches have been used to analyze image data from a siRNA screen designed to study MHC Class II antigen presentation.

An example predictive clustering tree obtained in this domain is given in Figure 1. The tree has been produced by clustering phenotypes as described by 13 image features (such as intensity, texture, etc.). The cosine distance/similarity metric has been used for the clustering.

The internal nodes of the tree contain GO terms with which the genes are annotated. The leaves of the tree correspond to the clusters/groups of genes. For example, one such group (C1) includes the genes involved in the biological processes of 'defense response' (GO0006952) and 'regulation of metabolic processes' (GO001922).

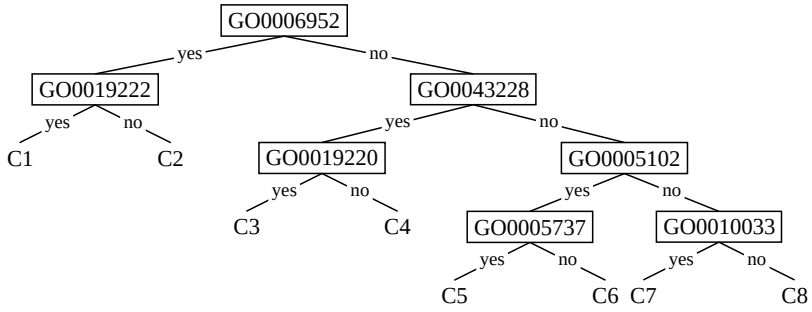

Fig. 1. A predictive clustering tree obtained from a siRNA screen for studying the MHC Class II antigen presentation. The internal nodes of the tree contain GO terms with which the genes are annotated. Leaves of the tree correspond to clusters of genes.

An example predictive clustering rule obtained in this domain is given in Table 2. The tree has been produced by clustering phenotypes as described by 6 image features. Feature selection was performed on the GO terms and only the selected subset of features was used to explain the clusters. The Euclidean distance measure was used. The cluster contains genes which are involved in 'regulation' (GO0065007) and in particular 'cellular nucleobase, nucleoside, nucleotide and nucleic acid metabolic process' (GO0006139).

Table 1. A predictive clustering rule obtained from a siRNA screen for studying the MHC Class II antigen presentation. The conditions in the antecedent describe the genes in the group in terms of their GO annotations.

```
IF GO0006139 = 1 AND
   GO0065007 = 1
THEN ClusterD1
```

In sum, we have applied predictive clustering to data from a siRNA screen designed to study MHC Class II antigen presentation. As a result of the predictive clustering process, we obtain clearly defined/described groups of genes, which yield similar phenotypes upon siRNA mediated downregulation. Groups of this kind can be used to identify pathways regulating the processes of interest (such as MHC Class II antigen presentation).

References

1. Timo Aho, Bernard Ženko, and Sašo Džeroski. Rule ensembles for multi-target regression. In Wei Wang, Hillol Kargupta, et al., editors, *Proceedings of the Ninth IEEE International Conference on Data Mining (ICDM 2009)*, pages 21–30, Los Alamitos, CA, 2009. IEEE Computer Society.
2. Michael Ashburner et al. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25:25–29, 2000.
3. Hendrik Blockeel, Luc De Raedt, and Jan Ramon. Top-down induction of clustering trees. In J. Shavlik, editor, *Proceedings of the 15th International Conference on Machine Learning*, pages 55–63. Morgan Kaufmann, 1998.
4. Anne Carpenter, Thouis Jones, Michael Lamprecht, Colin Clarke, In Kang, Ola Friman, David Guertin, Joo Chang, Robert Lindquist, Jason Moffat, Polina Golland, and David Sabatini. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, 7(10):R100+, 2006.
5. Minoru Kanehisa, Susumu Goto, Miho Furumichi, Mao Tanabe, and Mika Hirakawa. Kegg for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38(DB):D355–360, 2010.
6. Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, 1990.
7. Ivica Slavkov, Valentin Gjorgjioski, Jan Struyf, and Sašo Džeroski. Finding explained groups of time-course gene expression profiles with predictive clustering trees. *Molecular BioSystems*, 6(4):729–740, 2010.
8. Jan Struyf and Sašo Džeroski. Constraint based induction of multi-objective regression trees. In *Proc. of the 4th International Workshop on Knowledge Discovery in Inductive Databases KDID - LNCS 3933*, pages 222–233. Springer, 2006.
9. Celine Vens, Jan Struyf, Leander Schietgat, Sašo Džeroski, and Hendrik Blockeel. Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214, 2008.
10. Bernard Ženko and Sašo Džeroski. Learning classification rules for multiple target attributes. In Takashi Washio, Einoshin Suzuki, et al., editors, *Proceedings of the 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2008)*, LNCS, pages 454–465, Berlin, Germany, 2008. Springer.
11. Bernard Ženko, Jan Struyf and Sašo Džeroski. Analyzing time series gene expression data with predictive clustering rules. In Sašo Džeroski, Pierre Geurts and Juho Rousu, editors. *Machine learning in systems biology: proceedings of the Third International Workshop,* September 5-6, 2009, Ljubljana, Slovenia, pages 177-178. (Julkaisusarja - Helsingin yliopisto. Tietojenksittelytieteen laitos, report B-2009-1), 2009. University of Helsinki, Department of Computer Science.