Analyzing time series gene expression data with predictive clustering rules

Bernard Ženko[1], Jan Struyf[2], and Sašo Džeroski[1]

[1] Department of Knowledge Technologies, Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana, Slovenia
{bernard.zenko, saso.dzeroski}@ijs.si
[2] Department of Computer Science, Katholieke Universiteit Leuven
Celestijnenlaan 200A, B-3001 Leuven, Belgium
jan.struyf@cs.kuleuven.be

Abstract. Under specific environmental conditions, co-regulated genes and/or genes with similar functions tend to have similar temporal expression profiles. Identifying groups of genes with similar temporal profiles can therefore bring new insight into understanding of gene regulation and function. The most common way of discovering such groups of genes is with short time series clustering techniques. Once we have the clusters, we can also try to describe them in terms of some common characteristics of the comprising genes, e.g., (Ernst et al., 2005). An alternative way are the so-called constrained clustering techniques; here only clusters with valid descriptions are considered, and as a result, we obtain clusters and their descriptions in one single step.

We present a novel constrained clustering method for short time series, which uses the approach of predictive clustering. Predictive clustering (Blockeel et al., 1998) combines clustering and predictive modeling; it partitions the instances in a set of clusters like the regular clustering does, however, it also constructs predictive model(s) that describes each of the clusters. So far, predictive models can take the form of decision trees (Blockeel et al., 1998) or rules (Ženko, 2007). Predictive clustering trees, together with a qualitative time series distance measure (Todorovski et al., 2002), have already been used for clustering of short time series (Džeroski et al., 2007). Here we present predictive clustering rules for short time series, which use the same qualitative distance measure, but describe clusters with decision rules instead of trees.

The advantage of rules over trees is that each rule describing a cluster can be interpreted independently of other rules (clusters), while a tree describes all the clusters simultaneously. In addition, within rules we can easily introduce an additional constraint that rule conditions only comprise tests on the presence of gene descriptors and not on their absence. Trees by their nature have to include both types of tests (a set of instances is split into a cluster where the gene descriptor is present, and another set where the descriptor is absent), even if tests on absence are not biologically meaningful.

We demonstrate the benefits of our method on a publicly available collection of data sets (Gasch et al., 2000), which records the changes over time in the expression levels of yeast genes in response to a change in several environmental conditions. As the gene descriptors we use the Gene Ontology terms (Ashburner et al., 2000). The results show that rules give rise to clusters of genes with similar statistical properties (e.g., intra cluster variance and size) as trees, however,

the descriptions of the clusters are easier to interpret since they only include the presences of gene descriptors.

Keywords: time series, predictive clustering, rule learning

References

1. M. Ashburner et al. Gene Ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, 25(1):25–29, 2000.
2. H. Blockeel, L. De Raedt, and J. Ramon. Top-down induction of clustering trees. In *15th Int'l Conf. on Machine Learning*, pages 55–63, 1998.
3. J. Ernst, Nau G.J., and Bar-Joseph Z. Clustering short time series gene expression data. *Bioinformatics*, 21(Suppl. 1):159–168, 2005.
4. L. Todorovski, B. Cestnik, M. Kline, N. Lavrač, and S. Džeroski. Qualitative clustering of short time-series: A case study of firms reputation data. In *ECML/PKDD-'2 Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning*, pages 141–149, 2002.
5. S. Džeroski, V. Gjorgjioski, I. Slavkov, and J. Struyf. Analysis of time series data with predictive clustering trees. In *5th Int'l Workshop on Knowledge Discovery in Inductive Databases: Revised Selected and Invited Papers*, LNCS, Volume 4747, pages 63-80. Springer, 2007.
6. B. Ženko. *Learning predictive clustering rules.* PhD thesis, University of Ljubljana, Slovenia, 2007.