On the Feasibility of Heterogeneous Analysis of Large Scale Biological Data

Ivan G. Costa, Alexander Schliep

Department of Computational Molecular Biology
Max Planck Institute for Molecular Genetics, Berlin, Germany
{ivan.filho,alexander.schliep}@molgen.mpg.de

Abstract. Secondary information such as Gene Ontology (GO) annotations or location analysis of transcription factor binding is often relied upon to demonstrate validity of clusters, by considering whether individual terms or factors are significantly enriched in clusters. If such an enrichment indeed supports validity, it should be helpful in finding biologically meaningful clusters in the first place. One simple framework which allows to do so and which does not rely on strong assumptions about the data is semi-supervised learning. A primary data source, gene expression time-courses, is clustered and GO annotation or transcription factor binding information, the secondary data, is used to define soft pair-wise constraints for pairs of genes for the computation of clusters. We show that this approach improves performance when high quality labels are available, but naive use of the heterogeneous data routinely used for cluster validation will actually decrease performance in clustering.

1 Introduction

A fundamental task in the analysis of gene expression time-courses is to find groups of genes undergoing the same transcriptional program or sharing similar functions. The numerous clustering methods proposed in the literature [2] are often validated by showing a statistically significant enrichment of individual Gene Ontology (GO) terms or transcription-factor binding information in some or all clusters. If the validity of a cluster is concluded from secondary data shared by its elements, a clustering procedure which prefers such clusters in the computation should yield superior results.

A natural, simple and mostly assumption-free framework is semi-supervised learning. Methods make use of labels which are available for a subset of objects in a combination of supervised and unsupervised learning. One particular type of methods is called clustering with constraints and it makes weaker assumptions about the labels by encoding secondary information as pair-wise constraints. We use either Gene Ontology annotation (`GO`) [1] or data from location analysis of transcription regulators bindings (`TR`) [6] as secondary information. For this data, the use of a clustering with constraints method, instead of a joint analysis approach [10, 11], has two advantages: (1) `GO` and `TR` is not available for all genes from expression experiments; and (2) gene expression time-courses provide

ECML/PKDD-2006 Workshop on Data and Text Mining for Integrative Biology
M. Hilario and C. Nédellec (Eds.), pp. 55-60, Berlin, 2006.

one view of the biological process under investigation, which is very unlikely to provide the same level of details as GO or TR data. Using such data as secondary information we can limit the results to biologically more plausible solutions.

One challenge of using GO or TR data as secondary knowledge is their complex and overlapping structure. GO, for example, consists of three directed acyclic graphs (DAG), composed of terms describing either molecular functions, processes or components. Most genes are directly annotated with several terms. Furthermore, if a gene is annotated with one term, it is also associated with all parent nodes of this term. Even though the structure of TR is simpler, genes are often associated with more than one transcription regulator and vice-versa. This work makes use of soft pair-wise constraints to model the secondary information, following the approaches of [5] and [7], and extending the semi-supervised approach applied for gene expression proposed in [9]. The challenge in this method is the formulation of the constraints between pairs of genes, which ideally should extract as much information from the secondary data as possible.

2 Mixture Model Estimation with Constraints

A standard mixture model can be defined as $\mathbf{P}[x_i|\Theta] = \sum_{k=1}^{K} \alpha_k \mathbf{P}[x_i|\theta_k]$, where $X = \{x_i\}_{i=1}^{N}$ is the set of observed vectors and $\Theta = (\alpha_1, ..., \alpha_K, \theta_1, ..., \theta_K)$ are the model parameters. By including a set of hidden labels $Y = \{y_i\}_{i=1}^{N}$, where $y_i \in \{1, .., K\}$ defines the component generating the x_i, we obtain the complete data likelihood, which can be estimated with the EM method.

$$\mathbf{P}[X, Y|\Theta] = \mathbf{P}[X|Y, \Theta]\mathbf{P}[Y|\Theta]$$

The constraints are incorporated in the estimation by extending the prior probability of the hidden variable to $\mathbf{P}[Y|\Theta, W] = \mathbf{P}[Y|\Theta]\mathbf{P}[W|Y, \Theta]$, where W is the set of positive constraints $w_{ij}^{+} \in [0, 1]$ and negative constraints $w_{ij}^{-} \in [0, 1]$, for all $1 \leq i < j \leq N$. As in Lu and Leen 2005, we use the following distribution from the exponential family to model $\mathbf{P}[W|Y, \Theta]$.

$$\mathbf{P}[W|\Theta, Y] = \frac{1}{Z} \exp^{\sum_i \sum_{j \neq i} -\lambda^{+} w_{ij}^{+} 1\{y_j \neq y_i\} - \lambda^{-} w_{ij}^{-} 1\{y_j = y_i\}}$$

Lange *et al.* [5] showed that this distribution follows the Maxent principle, where λ^{+} and λ^{-} are Lagrange parameters defining the penalty weights of positive and negative constraints violations. In this formulation, however, one cannot assume independence between elements in Y in the estimation step. Exact inference of the posterior $\mathbf{P}[y_i = k|x_i, \Theta]$ involves the marginalization over all objects with some non-zero constraint with the ith object. Such computation is only feasible when the constraints are highly decoupled, which is not the case of the structures in this study. One way to approximate the posterior distribution is to use a mean field approximation [5]. More formally, the posterior assignments will take the form

$$\mathbf{P}[y_i = k|Y_i', X, \Theta, W] = \frac{\alpha_k \mathbf{P}[x_i, \theta_k]}{Z} \exp\left(\sum_{j \neq i} -\lambda^{+} w_{ij}^{+}(1 - r_{j,k}) - \lambda^{-} w_{ij}^{-} r_{j,k}\right),$$

where $r_{j,k} = \mathbf{P}[y_j = k|Y'_j, X, \Theta, W]$. When there is no overlap in the annotations—more exactly, $w^+_{ij} \in \{0,1\}$, $w^-_{ij} \in \{0,1\}$, $w^+_{ij}w^-_{ij} = 0$, and $\lambda^+ = \lambda^- \sim \infty$—we obtain hard constraints as the ones used in [9], or as implicitly performed in [8].

2.1 Constraints Definitions

Each DAG of gene ontology is composed of a set of terms $T = \{t_1, ..., t_p\}$ and a set of parent child relations between pairs of terms $P(t_l, t_m) \in \mathcal{P}$. The annotation of a set of genes $G = \{g_1, ..., g_N\}$ can be defined as $A(t_l, g_i) \in \mathcal{A}$. Furthermore, we also have the property that genes annotated with a term are also annotated with the whole set of parents of this term, or $(P(t_l, t_m) \in \mathcal{P}) \wedge (A(t_m, g_i) \in$ A$) \rightarrow A(t_l, g_i) \in \mathcal{A}$. The main idea for calculating the constraint is to account for the similarity of the sub-dags $D(g_i) = \{t_m | A(t_m, g_i) \in \mathcal{A}, t_m \in \mathcal{T}\}$ associated with the gene pairs. More formally, for all pair of genes g_i and g_j, we define the constraints as (non-annotated genes have constraints equal to zero):

$$w^+_{ij} = \frac{\#\{t_m | t_m \in D(g_i) \cap D(g_j)\}}{\#\{t_m | t_m \in D(g_i) \cup D(g_j)\}}, \text{and } w^-_{ij} = \frac{\#\{t_m | t_m \in D(g_i) \uplus D(g_j)\}}{\#\{t_m | t_m \in D(g_i) \cup D(g_j)\}}.$$

Similarly, the formula above can be used for a set of of transcription factors $F = \{f_1, ..., f_q\}$, where $A'(f_l, g_i) \in \mathcal{A}'$ indicates that factor f_l bounds to g_i and $D'(g_i)$ is the set of factors associated with g_i.

3 Results

We use the expression profiles of 384 genes during Yeast mitotic cell division assigned to one of the five cell cycle phases classes [4], which we refer to as `YC5`. Even though this data set is biased towards profiles showing periodic behavior, and some of the class assignments are ambiguous, it is one of the few with a complete expert labeling of genes. The relation between regulators and target genes where obtained from large scale location analysis [6], comprising data from 142 candidate regulators. Relations were obtained after thresholding the confidence that the factor binds to a particular gene as in the source literature. In relation to GO, the SGD Saccharomyces cerevisiae annotation was used and for simplicity, we only included the DAG molecular process in our analysis.

Multivariate normal distributions with diagonal covariance matrix are used as models for the expression profiles. Each parameter estimation is performed 15 times and the best model is chosen, to lessen effects of random initialization. For all experiments we varied values of λ^+ and λ^-. We use the class labels to compute sensitivity (`Sens`), specificity (`Spec`) and corrected Rand (`CR`).

As a proof of concept, we use the class labels from `YC5` to generate pair-wise constraints for 5% of all pairs of genes—positive if the genes belong to the same class, negative else—and observe the performance of the method with distinct penalizing settings (Fig. 1 top). In all cases, `CR`, `Spec` and `Sens` tend to one for λ near ten, with the exception of the experiments with positive constraints. In

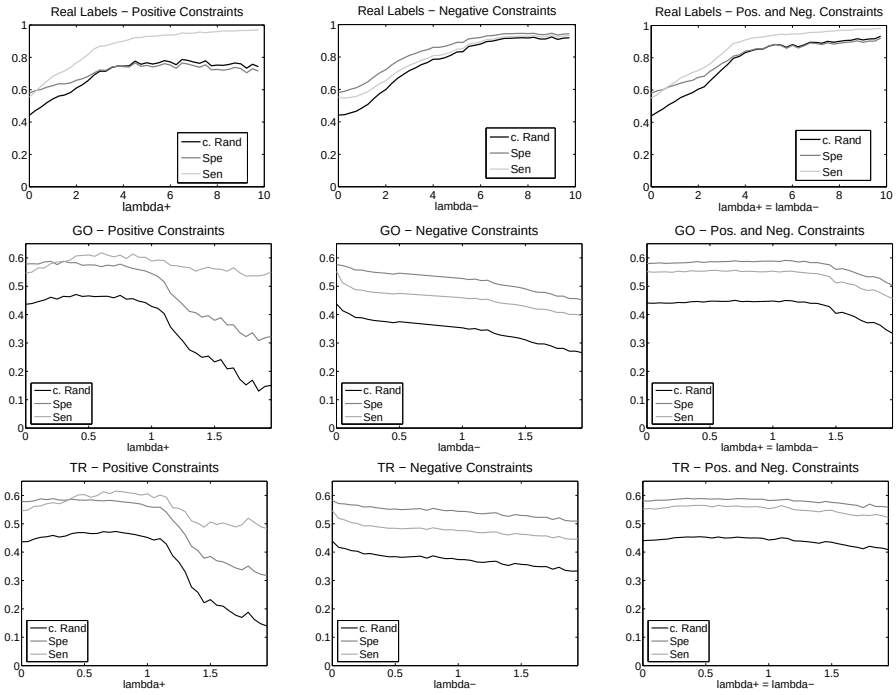

Fig. 1. We depict the `CR`, `Sens` and `Spec` after clustering `YC5` with positive (left), negative (middle) and positive and negative (right) constraints. We used either real class labels (top), `GO` (middle) or `TR` (bottom) as secondary information.

this case, one of the five components always remained empty, and two classes were joined. Furthermore, the use of positive constraints only had a stronger effect on the sensitivity, while the negative constraints affect the specificity. This is expected since these constraints penalize false negatives and false positives, respectively. It also explains the joined classes in the experiments with positive constraints, since the secondary data gives no penalty for those solutions (and the models for gene expression makes the decision).

We observe similar results if we use `GO` and `TR` as secondary data. There is a slight but significant increase of `CR` and `Sens` for the methods with positive constraints (t-test indicates an increase at $\lambda^+ = 0.5$ with p-value of $2.38e-10$), followed by a decrease in `CR`, `Sens` and `Spec` . No improvements were obtained with the use of positive and negative constraints, and the negative constraints alone only deteriorated the results. To better understand the results above, we repeated the experiments with real labels, but this time including random labels (also with 5% of pairs constrained). As seen in Fig. 2, the addition of random labels have a great impact on the recovery of the clusters. The inclusion of 20% of random labels worsen the results considerably, and for 60% of random labels the corrected Rand displays a behavior similar to `TR` and `GO`. This indicates that (1) the method is not robust in with respect to noise in the data, and (2) presence of noise or non-relevant information in `TR` and `GO`.

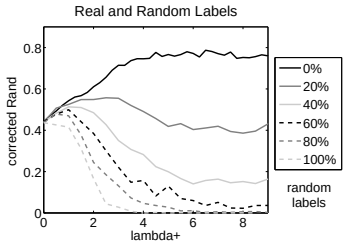

Fig. 2. We depict the `CR` obtained by clustering `YC5` with positive constraints from 5% of real labels with the inclusion of 0%, 20%, 40%, 60% and 100% random labels.

This however is not too surprising, so we attempt to estimate the maximal positive effect one can obtain from this secondary data. We perform the computation [3] for `GO` term and `TR` site enrichment used in cluster validation to obtain informative terms from the *true classes*. We repeat the experiments above with those most informative terms only. However, we observe only a slight improvement for the negative constraints and a marginal improvement with the use both positive and negative constraints in the `TR` data set (a `CR` from 0.454 to 0.472). On the other hand, no improvement was obtained after filtering terms in `GO` (data not shown).

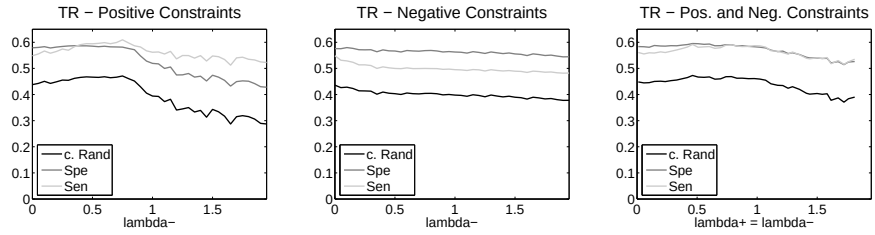

Fig. 3. We depict `CR`, `Sens` and `Spec` after clustering `YC5` with positive (left), negative (middle) and positive and negative (right) constraints after filtering of relevant `TR`.

4 Discussion

Semi-supervised learning is clearly an effective framework for joint analysis of heterogeneous data if high-quality secondary data is available as our experiments using class labels show. Surprisingly, using the very data routinely considered to support cluster validity—significantly enriched GO terms and location data—as secondary data can deteriorate cluster quality drastically. While there are parameter choices to explore, further theoretical questions to address and more data sets to repeat experiments on, the main point remains valid and clear: secondary data has little power for clustering, unless it is of very high quality, free of errors and ambiguities. Less than a percent of high-quality labels [9] have

a larger positive effect than 5% of labels of which 20% are incorrect. On one hand, this puts the economy of large-scale experiments into question. On the other hand, it stresses the importance of theoretical progress on how to reduce noise, assess reliability of individual data and how to incorporate per object quality indicators into methods.

Acknowledgments. The first author would like to acknowledge funding from the CNPq(Brazil)/DAAD.

References

1. M. Ashburner. Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, 2000.
2. Z. Bar-Joseph. Analyzing time series gene expression data. *Bioinformatics*, 20(16):2493–503, Nov 2004.
3. T. Beissbarth and T. P. Speed. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, 20(9):1464–1465, 2004.
4. R. Cho. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2:65–73, 1998.
5. T. Lange, M. H. Law, A. K. Jain, and J. M. Buhmann. Learning with constrained and unlabelled data. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 731–738, 2005.
6. T. Lee. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, 298(5594):799–804, 2002.
7. Z. Lu and T. Leen. Semi-supervised learning with penalized probabilistic clustering. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 849–856. MIT Press, Cambridge, MA, 2005.
8. W. Pan. Incorporating gene functions as priors in model-based clustering of microarray gene expression data. *Bioinformatics*, 22(7):795–801, 2006.
9. A. Schliep, I. G. Costa, C. Steinhoff, and A. A. Schnhuth. Analyzing gene expression time-courses. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(3):179–193, 2005.
10. E. Segal, R. Yelensky, and D. Koller. Genome-wide discovery of transcriptional modules from dna sequence and gene expression. *Bioinformatics*, 19(90001):273i–282, 2003.
11. C.-H. Yeang and T. Jaakkola. Time series analysis of gene expression and location data. In *Third IEEE Symposium on BioInformatics and BioEngineering (BIBE'03)*, number 305, 2003.