An Experimental Evaluation of Lifted Gene Sets

Ondřej Kuželka and Filip Železný

Czech Technical University, Prague, Czech Republic
{kuzelon2,zelezny}@fel.cvut.cz,

1 Introduction

In this paper, we study the question whether small sets of correlated genes can be better characterized by direct conventional methods or by indirect methods using first-order-logic descriptions. We introduce so-called lifted gene sets and show that they are able to predict correlations of genes better than conventional methods. The motivation for this study is not directly the estimation of correlations of genes but rather the possibility to find relational descriptions of sets of highly correlated genes because such sets are important in machine learning applications such as set-level predictive classification methods or group-lasso-based regression and classification methods [3].

2 Lifted Gene Sets

In order to be able to predict correlations of gene sets based on their relational descriptions we need to work with training examples which have both *structure* and *real parameters*. One example may e.g. describe a measurement of the expression of several genes; here the structure would describe functional relations between the genes and the parameters would describe their measured expressions. Note that we allow different structures in different examples. For example, a training set thus may consist of measurements pertaining to different gene sets, each giving rise to a different structure of mutual relations between the genes.

To describe the training examples as well as the lifted gene sets, we use a conventional first-order logic language $\mathcal{L}$ whose alphabet contains two distinguished sets of constants $\{r_1, r_2, \dots r_n\}$ and $\{g_1, g_2, \dots g_n\}$ and two distinguished sets of variables $\{R_1, R_2, \dots R_m\}$ and $\{G_1, G_2, \dots G_m\}$. Any substitution in our framework must map variables (other than) R_i only to terms (other than) r_j and variables (other than) G_i only to terms (other than) g_j. The structure of an example is described by a (Herbrand) interpretation H, in which the constants r_i represent uninstantiated real parameters and g_i represent the genes. The parameter values are then determined by a real vector $\boldsymbol{\theta}$. Thus each example is a pair $(H, \boldsymbol{\theta})$. A *lifted gene set* is simply a logic formula which has some free distinguished variables. Intuitively, a lifted gene set extracts some of the genes g_i and their respective expression levels r_i from the examples. For example, the intentionally simplistic lifted gene set

$$\exists G_1, G_2 \;\; \mathrm{expr}(G_1, R_1) \wedge \mathrm{expr}(G_2, R_2) \wedge \mathrm{regulates}(G_1, G_2) \qquad (1)$$

contains just two distinguished gene variables G_1, G_2 and two distinguished variables R_1, R_2 corresponding to gene-expression levels of G_1 and G_2.

Let us now first introduce some more notation so that we could clarify what we mean by *extracting genes and their expressions from examples.* If v is a real vector (an ordered list of genes, respectively) then v_i denotes the i-th element of v. If $I \subseteq [1;n]$ then $v_I = (v_{i_1}, v_{i_2}, \dots v_{i_{|I|}})$ where $i_j \in I$. For the largest k such that $\{R_1/r_{i_1}, R_2/r_{i_2}, \dots, R_k/r_{i_k}\} \subseteq \vartheta$ we denote $I_R(\vartheta) = (i_1, i_2, \dots i_k)$ and analogically for the largest k such that $\{G_1/g_{i_1}, G_2/g_{i_2}, \dots, G_k/g_{i_k}\} \subseteq \vartheta$ we denote $I_G(\vartheta) = (i_1, i_2, \dots i_k)$. Given an example $e = (H, \boldsymbol{\theta})$ and a lifted gene set φ, the *sample set* of φ and e is the multi-set $\mathcal{S}(\varphi, e) = \{\boldsymbol{\theta}_{I_R(\vartheta)} | H \models \varphi\vartheta\}$ where ϑ are r-substitutions grounding all free variables[1] in φ, and $H \models \varphi\vartheta$ denotes that $\varphi\vartheta$ is true under H. Similarly, we define the *ground gene sets of a lifted gene set* φ *and example* e as $\mathcal{S}(\varphi, e) = \{\boldsymbol{\theta}_{I_G(\vartheta)} | H \models \varphi\vartheta\}$. For example, the the set of ground gene sets corresponding to the example lifted gene set 1 is the set of all pairs of genes which are in the relation of *regulation.*

Our aim in this paper will be to discover lifted gene sets which correspond to highly correlated ground gene sets. Therefore we need to be able to compute *correlations* of genes within the lifted gene sets. For this we employ methods from our recently introduced framework called *Gaussian logic* [5]. Here we only briefly mention the way a covariance matrix is computed for a lifted gene set from which correlations may be easily obtained. The theoretical justification of the procedure can be found in [5].

Given a non-empty sample set $\mathcal{S}(\varphi, e)$, we define the $\boldsymbol{\Sigma}$-*matrix* as

$$\boldsymbol{\Sigma}(\varphi, e) = \frac{1}{|\mathcal{S}(\varphi, e)|} \sum_{\boldsymbol{\theta} \in \mathcal{S}(\varphi, e)} (\boldsymbol{\theta} - \boldsymbol{\mu}(\varphi, e)) (\boldsymbol{\theta} - \boldsymbol{\mu}(\varphi, e))^T \tag{2}$$

Using the above, we define the estimate $\widehat{\boldsymbol{\Sigma}}_\varphi$ over the entire training set $\{e_1, e_2, \dots e_m\}$

$$\widehat{\boldsymbol{\Sigma}}_\varphi = \frac{1}{m} \sum_{i=1}^{m} \left(\boldsymbol{\Sigma}(\varphi, e_i) + \boldsymbol{\mu}(\varphi, e_i) \boldsymbol{\mu}(\varphi, e_i)^T \right) - \widehat{\boldsymbol{\mu}}_\varphi \widehat{\boldsymbol{\mu}}_\varphi^T \tag{3}$$

where

$$\boldsymbol{\mu}(\varphi, e) = \frac{1}{|\mathcal{S}(\varphi, e)|} \sum_{\boldsymbol{\theta} \in \mathcal{S}(\varphi, e)} \boldsymbol{\theta} \text{ and } \widehat{\boldsymbol{\mu}}_\varphi = \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\mu}(\varphi, e_i). \tag{4}$$

We can extract the *lifted-gene-set correlations* from this matrix $\widehat{\boldsymbol{\Sigma}}_\varphi$.

[1] Note that an interpretation H does not assign domain elements to variables in $\mathcal{L}$. The truth value of a *closed* formula (i.e., one where all variables are quantified) under H does not depend on variable assignment. For a general formula though, it does depend on the assignment to its free (unquantified) variables.

3 Experiments

In this section we will experimentally assess the next three questions which should give us clues about the usefulness of lifted gene sets.

- Q1: *Are lifted gene sets better predictors of correlation than ground gene sets?*
- Q2: *Are complex lifted gene sets constructed by relational machine learning techniques better predictors of correlation than gene sets corresponding to biological pathways?*
- Q3: *Are lifted-gene-set correlations stable across gene-expression datasets?*

We will answer each of these questions experimentally in the following three subsections. In all the experiments we used algorithms from [5] for construction of lifted gene sets. In each lifted gene set φ we added $\neq$ constraints for all pairs of variables except the R_i variables but for brevity we will not display them. So when we will write $\varphi = g(G_1, R_1), g(G_2, R_2)$ we will mean $\varphi = g(G_1, R_1), g(G_2, R_2), G_1 \neq G_2$ etc. In all the experiments we used a set of seven gene-expression datasets obtained from the *Gene-Expression Omnibus* database [1] and relational descriptions of biological pathways from the *KEGG* database [4]. We worked with a subset of genes contained in a set of 50 KEGG pathways. Several examples of constructed logical formulas describing lifted gene sets and their average correlations are shown in Table 1.

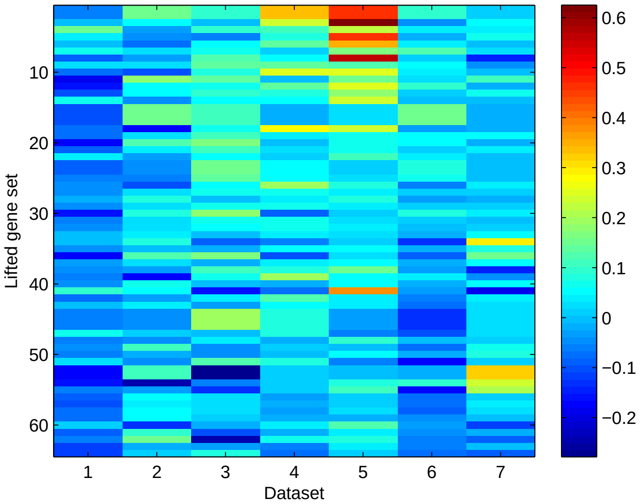

Fig. 1. Difference between the error of ground gene sets and lifted gene sets across 7 gene-expression datasets (the more positive the number the better for lifted gene sets).

Table 1. Logical formulas defining lifted gene sets and their average correlations across the seven gene-expression datasets.

Formula	**Avg. corr.**
$\exists X : g(G_1, R_1) \wedge \mathit{expresses}(G_1, X) \wedge \mathit{expresses}(X, G_2) \wedge g(G_2, R_2)$	0.29
$\exists X : g(G_1, R_1) \wedge \mathit{activates}(G_1, X) \wedge \mathit{indirect}(X, G_2) \wedge g(G_2, R_2)$	0.24
$\exists X, Y : \mathit{binds/associates}(X, G_1) \wedge \mathit{inhibits}(G_1, G_2) \wedge g(G_1, R_1) \wedge$ $\wedge g(G_2, R_2) \wedge \mathit{activates}(G_1, Y)$	0.15

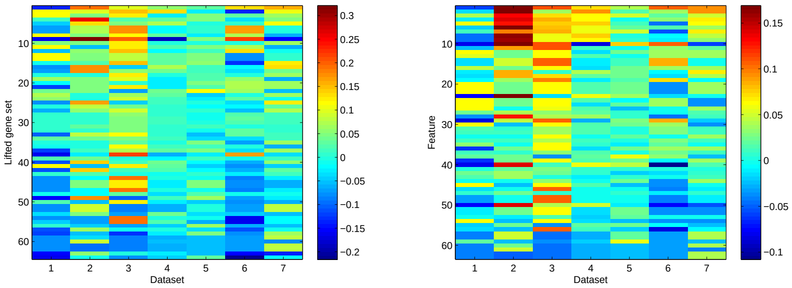

Fig. 2. Differences between the error of pathway-based gene sets and lifted gene sets across 7 gene-expression datasets (left panel) and average difference between the error of pathway-based gene sets and combined gene sets - the more positive the number the better for the combined gene sets.

3.1 Comparison of Lifted Gene Sets and Ground Gene Sets

Let us start with the first question: *Are lifted gene sets better predictors of correlation among genes than ground gene sets?* In order to answer this we performed the following experiment. We constructed a set of lifted gene sets with just 2 gene-variables (to make the interpretation of the results as simple as possible) and estimated the *lifted correlations*, i.e. the correlations computed according to Formula 3, for each gene-expression dataset separately. We also estimated the correlations using a conventional shrinkage-based method [6]. After that we used these values for prediction of correlations in datasets not used for training. We compared the errors incurred when using lifted gene sets and when using the conventional estimates (ground gene sets). The results are shown in Fig. 1. On average, the error incurred when estimating the correlations using lifted gene sets was smaller by 0.03 than the error of the estimates obtained using the ground gene sets.

3.2 Comparison of Lifted Gene Sets and Pathway-based Gene Sets

The results from the previous subsection indicate that lifted gene sets perform better than ground gene sets. A simple type of a lifted gene set is a lifted gene set which assumes all pairs of genes which are contained in some pathway P.

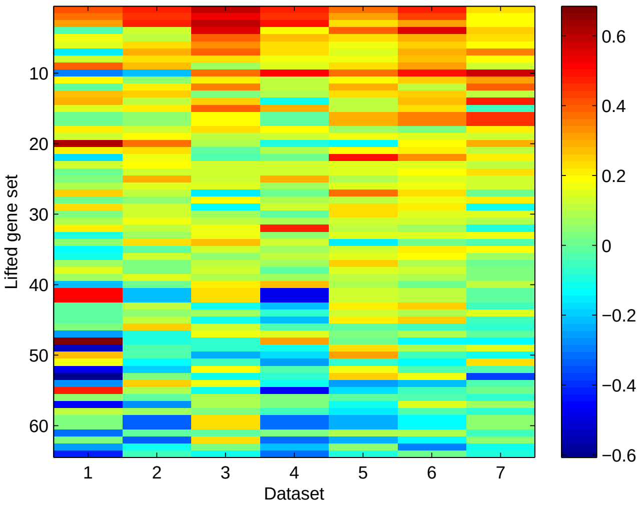

Fig. 3. Average correlations of lifted gene sets across 7 gene-expression datasets.

This type of gene sets is, in fact, quite popular in machine learning applications (pathway-based gene sets are often used in set-level predictive classification methods, e.g. [2]). It is therefore an interesting question whether our more complex lifted gene sets based on relations from KEGG are able to estimate correlations between genes more reliably than the traditional pathway-based gene sets. In order to answer this question we performed the same procedure as described in the previous subsection but instead of ground gene sets we used the pathway-based gene sets. The results are shown in Fig. 2. On average the error incurred when estimating the correlations using lifted gene sets was smaller by 0.01 than the error of the estimates obtained using the pathway-based gene sets. However, this is rather small difference and, as can be seen from Fig. 2, for many combinations *lifted gene set - dataset*, lifted gene sets performed worse than pathway-based gene sets. Therefore we performed an additional experiment in which we tested the estimates obtained by averaging the predictions of lifted gene sets and pathway-based gene sets (*combined gene sets*). In this case the average improvement over the plain pathway-based gene sets was more significant: 0.02.

3.3 Stability of Lifted-Gene-Set Correlations across Datasets

Finally, we also performed a set of experiments in order to determine stability of lifted-gene-set correlations across various gene-expression datasets. In order to do so we again estimated the correlations of the individual lifted gene sets separately for each dataset and plotted them in Fig. 3. We can notice that on average the correlation of the lifted gene sets does not change very much across datasets with some notable exceptions which may or may not be interesting from the biological point of view.

4 Discussion and Conclusions

The experiments that we have performed in this paper give some clues as for the potential of lifted gene sets. However, they also show that simple gene sets like pathway-based gene sets can perform similarly well as the more complex lifted gene sets. They also show that combinations of the simple and the more complex gene sets can sometimes improve ability to predict correlation between genes.

Acknowledgement: This work was supported by the Czech Grant Agency through project 201/09/1665 *Bridging the Gap between Systems Biology and Machine Learning* and project 103/11/2170 *Transferring ILP techniques to SRL.*

References

1. R. Edgar, M. Domrachev, and A. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, 1, 2002.
2. M. Holec, F. Železný, J. Kléma, and J. Tolar. Integrating multiple-platform expression data through gene set features. In *Proceedings of the 5th International Symposium on Bioinformatics Research and Applications*, ISBRA '09, pages 5–17. Springer-Verlag, 2009.
3. L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 433–440. ACM, 2009.
4. M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The kegg resource for deciphering the genome. *Nucleic Acids Research*, 1, 2004.
5. O. Kuželka, A. Szabóová, M. Holec, and F. Železný. Gaussian logic for predictive classification. In *To appear in Proceedings of the European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases*, ECML PKDD, 2011.
6. J. Schäffer and K.Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 2005.