An empirical comparison of popular algorithms for learning gene networks

Vera Djordjilović [1], *Monica Chiogna* [2], *Jirka Vomlel* [3]

[1] djordjilovic@stat.unid.it, Department of Statistical Sciences, University of Padova, Italy.
[2] monica@stat.unipd.it, Department of Statistical Sciences, University of Padova, Italy.
[3] vomlel@utia.cas.cz, Department of Decision Making Theory, Institute Of Information Theory and Automation, Czech Republic.

Abstract

We perform an empirical comparison of learning algorithms used to reconstruct networks of genes, represented by directed acyclic graphs. We compare approaches designed for categorical and continuous data and study the impact of including prior information. Our results suggest that categorizing continuous gene expression measurements and including, even vague, prior information can significantly improve the predictive accuracy of learned models.

Keywords: Gene networks, Structure learning.

1. Introduction

Reverse engineering is the process of reconstructing a structure of a dynamic system, reasoning backwards from observations of its behaviour. Although an area of active interest in many scientific disciplines, this is especially true in systems biology; biological systems are notoriously complex, and understanding how different pieces come together is a challenging task. Here, we treat one particular problem of reverse engineering: reconstructing networks of genes, represented by directed acyclic graphs (DAGs), from their expression measurements.

We consider a number of popular structure learning algorithms and apply them to the experimental data from the *Drosophila Melanogaster* experiment performed by the University of Padova [4]. Our focus is on prediction and so we evaluate different methods on the basis of their predictive accuracy. Given the small sample size ($n = 28$ observations of $p = 12$ genes), to assess the predictive accuracy we adopt a "leave-one-out" approach, where in each step the chosen learning algorithm is applied to the data from which a single observation has been removed. In the second step, the removed observation is used to evaluate the predictive accuracy: prediction of the value of every variable is computed given the values of all other variables. To measure the distance between the observed value and the predicted value for each gene, we use the Brier score, introduced in [1]. The Brier score measures the squared distance between the forecast probability distribution and the observed value. It can assume values between 0 (the perfect forecast) and 1 (the worst possible forecast). For every algorithm we thus have n predictions, one for each observation that is being left out. We measure the predictive accuracy of the algorithm with a scalar measure B

$$B = \sum_{j=1}^{n} \sum_{i=1}^{p} {}_j b_i, \tag{1}$$

where ${}_jb_i$ is the Brier score corresponding to the prediction of the $i$the gene, based on the model learned after excluding the jth observation from the dataset. Obviously, algorithms having lower score are preferred.

Gene expression measurements are continuous, and since we aim to compare algorithms designed for categorical and continuous data we need a categorization procedure. We employ a data driven categorization based on the idea that genes can assume only a few functional states, such as "under-expressed", "normal", and "over-expressed". The actual measurements depend on these functional states and the amount of biological variability and technical noise. A plausible model for such data is a mixture of K (where above $K = 3$) normal distributions, each centered at one of the K functional states. Since it is not always plausible to assume that all K states are present in a single experiment, we propose to estimate the number of components varying from one (corresponding to a gene with only one observed state) to K (all functional states are present in the data) from the data for each gene independently. The approach that simultaneously estimates the number of components in the mixture and parameters pertaining to different components and then classifies each observation according to the estimated model is called Model Based Clustering and was introduced in [5]. We used its implementation in the R package `mclust`.

The learning algorithms that work with continuous data produce predictions on the continuous scale. In order to make them comparable with categorical predictions, we combine discriminant analysis with the proposed categorization procedure. We classify continuous predictions into one of the gene specific components estimated in the initial categorization.

2. Considered algorithms

In this empirical study, we consider a number of variants of the PC algorithm [6], the K2 algorithm [2] and the exact Gobnilp method [3]. Of the examined approaches, the K2 algorithm and all modifications of the K2 algorithm considered here, include the prior information. The prior information is in the form of the topological ordering of the studied genes. To specify the topological ordering, we relied on public databases of biological knowledge. In particular, we considered a WNT pathway of the KEGG database.

In summary, the considered options are the following:

PC The PC algorithm using χ^2 test of independence at the 5% significance level.

PC20 The PC algorithm using χ^2 test of independence at the 20% significance level.

K2 The original K2 algorithm.

K2-BIC A modified K2 algorithm, where the criterion used to score competing DAGs is BIC, while the search strategy remains the one step greedy search.

G-BIC The Gobnilp algorithm with the BIC scoring criterion.

G-BICm The Gobnilp algorithm with the modified BIC criterion (the penalty term is multiplied by a factor of 10^{-3}).

G-BICl The Gobnilp algorithm with the modified BIC criterion (the penalty term is multiplied by 10^{-9}).

CK2 The CK2 algorithm proposed in [4]. A modification of the K2 algorithm for continuous data, where the K2 score is replaced with the BIC criterion for the multivariate normal distribution. The only algorithm in this study that is applied to the continuous measurements.

Full graph Corresponds to the complete directed acyclic graph, which is a directed acyclic graph whose skeleton is a complete graph. In other words, the set of conditional independence relations entailed by such a DAG is empty.

Empty graph Corresponds to the DAG containing no arrows. In other words, the variables of such a graph form a system of independent random variables. This is a very naive prediction method, but it may serve as a reference for comparison with more advanced methods.

3. Results

	PC	PC20	K2-BIC	K2	G-BIC	G-BICm	G-BICl	CK2	Full	Empty
psn	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.01	3.00	2.88
nkd	4.65	2.61	0.00	0.00	1.00	0.00	0.00	8.50	3.00	7.53
dally	5.36	6.97	5.34	5.30	6.29	5.30	5.30	13.56	6.30	9.72
por	1.00	0.00	0.00	0.00	1.00	1.00	0.00	1.98	3.00	2.88
daam	4.81	3.81	4.87	3.95	5.13	3.95	3.44	2.99	5.44	6.15
fz	4.07	1.19	2.41	1.12	1.12	1.12	1.12	0.01	3.12	6.15
rho1	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.41	3.00	2.88
dco	3.29	3.05	1.25	1.50	1.25	2.50	2.50	1.29	3.50	3.69
sgg	0.18	0.00	0.00	0.00	1.00	0.00	1.00	0.99	3.00	2.88
pont	1.10	3.16	1.50	1.50	1.50	1.50	1.50	1.98	3.50	7.37
	24.46	21.79	16.37	14.37	18.29	15.37	14.86	32.72	36.86	52.13

Table 1: Evaluation of the prediction accuracy: the B Score.

In Table 1, we report the B score for each of the considered methods. Two genes were excluded from the analysis, since in the categorized dataset they assumed only one value. In our study K2 reaches the minimal B score, followed by the Gobnilp's likelihood method G-BICl. The K2 algorithm with the BIC score, K2-BIC, together with the remaining Gobnilp methods, G-BICm and G-BIC, also perform reasonably well with a slightly inferior score with respect to the leading twosome. On the other hand, the PC algorithm gives significantly less accurate predictions. The CK2 algorithm, seems to fail in this case. Its B score is almost comparable to the one of the full graph (Full).

4. Discussion and conclusions

We compared a number of different approaches of inferring a network of genes on the basis of gene expression measurements. In terms of prediction accuracy the most promising one seems

to be the K2 algorithm that, in addition to the experimental data in the form of categorized measurements, requires information regarding the topological ordering of genes. The possible reasons for the success of K2 are twofold: its inferred graphs are more dense with respect to graphs inferred by other methods (the property related to the K2 scoring criterion), and the use of prior information that seems to point the search towards "better" models, at least when it comes to prediction considerations. On the other hand, we attribute the somewhat surprisingly low performance of CK2, in large part, to the use of the continuous measurements that makes predictions much more sensitive and less robust. This might indicate that genomics is one of the few settings in which the advantages of categorization reflected in attenuating technical noise outweigh the incurred information loss.

5. Bibliography

[1] Brier, G.M. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*. 1 (78).

[2] Cooper, G.F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine learning*. 4 (9).

[3] Cussens, J. and Bartlett, M. (2013). GOBNILP 1.4. 1 User/Developer Manual1.

[4] Djordjilović, V. (2105). *Graphical modelling of biological pathways*. PhD thesis, University of Padova.

[5] Fraley, C. and Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 458 (97).

[6] Spirtes, P. and Glymour, C.N. and Scheines, R. (2000). *Causation, prediction, and search*. MIT press.