Discovering Relationships Among Human Cancer, Gene Expression Profile And Drug Responses: A Relational Clustering Approach

Elisabetta Fersini, Ilaria Giordani, Enza Messina, Francesco Archetti

University of Milano-Bicocca, 20126 Milan, Italy
{fersini, giordani, messina, archetti}@disco.unimib.it

Abstract: The combined analysis of the microarray and drug-activity datasets has the potential of revealing valuable knowledge about various relations among gene expressions and drug activity patterns in tumor cells. However, the huge amount of biological data needs appropriate data mining models in order to extract interesting patterns and useful information. In this paper, the NCI60 dataset has been analyzed for the molecular pharmacology of cancer. In particular, we proposed a novel relational clustering algorithm joint with Bayesian network inference engine for linking gene expression profiles to drug activity patterns. Our analysis could be an initial step for predicting potential useful drugs according to the gene expression level of tumor tissues.

Keywords: NCI60 dataset analysis, Relational Clustering, Bayesian Networks

1. Introduction

One of the most challenging problem in biomedical research is related to the discovery of embedded relationships among human cancer, gene expression profile and drug activity. Highlighting these relationships is of crucial importance for several objectives, among others: identification of mechanisms of the cancer development, design of new molecular targets for anti-cancer drugs and definition of an individual therapy driven by a specific gene profile. A number of investigations available into the literature [1] [4] highlight two main interesting remarks: (1) drug activity patterns are less related to the organ of origin compared to the gene expression profile. This suggests us that a gene expression profile of a cell line plays a fundamental role, independently from the tissue of origin, to understand anticancer therapy responses; (2) learning causal relationships reveals interesting interaction among subset of genes, drugs and cancer types. Inspired by these remarks a two-folds analysis is proposed. In the first fold we perform different cluster analysis aimed at linking gene expression profiles to drug activity patterns. In particular, a novel relational clustering algorithm aimed at investigating whether drug response can be related to subsets of gene patterns is proposed. In the second fold we exploit the output of cluster analysis to induce a specific Bayesian Network able to predict the response of a set of drugs. Computational results show that the proposed Relational Clustering algorithm joint with Bayesian Networks inference engine yields to obtain an interesting link between gene expression profiles, tissues of origin and drug responses.

2. Methodology

In order to investigate embedded relationships between gene expression profile, cell lines and drug responses w.r.t. the pharmacology of cancer, we studied the well known NCI60 dataset, established at the National Cancer Institute U.S.A. It consists of 60 cell lines from 9 kinds of cancers, all extracted from human patients. The NCI60 dataset, presented in [8], can be viewed as a set Ω into the space R^{m+n}:

$$\Omega = \left\{ x \mid x = \left(x^G, x^D \right), x^G \in R^m, x^D \in R^n \right\} \quad (1)$$

where x is a cell line, x^G represents the gene expression level into a space R^m and x^D

denotes the drug response into a space R^n. In particular, x^G has been derived by using the cDNA microarray and x^D by assessing the grown inhibition activities (GI50) after 48 hours of drug treatment through Sulphorhodamine B. We consequently define Ω^G and Ω^D as:

$$\Omega^G = \left\{ x^G \mid x = \left(x^G, x^D \right) \in \Omega \right\} \text{ and } \Omega^D = \left\{ x^D \mid x = \left(x^G, x^D \right) \in \Omega \right\} \quad (2)$$

In order to obtain a meaningful representation of these data, in terms of discriminative features that can be used by the machine learning algorithms a preprocessing step has been performed. We defined Ω^2 by removing from the original dataset those genes and drugs for which at least one cell line had a missing value. In this case, each cell line is represented into the gene expression and drug activity spaces $R^{m=555}$ and $R^{n=836}$ respectively. Cell lines have been normalized in order to have mean equal to 0.

2.1 Relational Clustering

The main idea of the proposed relational clustering is to create groups of cell lines into the drug space by constraining this clustering process in order to consider also the relationships between couples of cell lines into the gene space. This is motivated by the fact that if two cell lines show a similar gene expression profile they will likely provide a similar response to the same drugs. Our proposed relational clustering is characterized by two main phases: in the first phase we learn relationships between cell lines over the gene space, while into the second phase we incorporate these relationships along with an underlying objective function over the drug space. The first phase, aimed at discovering relationships between cell lines over the gene space, can be viewed as a clustering problem and can be solved by the k-Means algorithm. The obtained set of clusters into the gene space leads us to define two kind of relationships: affinity-link, defined if two cell lines are placed into the same cluster, and diversity-link if two cell lines are placed into different clusters. Affinity and diversity links between are established between each couple of cell lines, according to the k-Means clustering output, and are weighted by using two matrices. A diversity link matrix R^U is a $|\Omega| \times |\Omega|$ matrix whose elements r_{ik}^U represent the weights of the *diversity-links* between elements belonging to different clusters. It will suggest, in the following phase that two elements should not be placed in the same cluster (module). An affinity link matrix R^A, having the same dimension of R^U, represents the weight of the *affinity-links* of elements belonging to the same cluster and suggests that two elements should be placed in the same cluster (module). r_{ik}^U and r_{ik}^A are defined as follows:

$$r_{ik}^U = \begin{cases} dist\left(x_i^G, x_k^G\right) & \text{if } x_i^G \in C_\alpha^G \text{ and } x_k^G \in C_\beta^G \neq C_\alpha^G \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$r_{ik}^A = \begin{cases} dist\left(x_i^G, x_k^G\right) & \text{if } x_i^G \wedge x_k^G \in C_\alpha^G \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

The second phase is focused on grouping cell lines using drug response information by taking into account cell- to-cell relationships coming from the previous stage. Let x_i^D be a given cell line represented by its drug response features and C_j^D, with j=1:J, be a set of clusters. The problem can be formulated as follows:

$$\min \sum_{j=1}^{J} \left[\sum_{i,k=1}^{|\Omega|} \left[dist\left(x_i^D, x_k^D\right) z_{ij} z_{kj} - dist\left(x_i^D, x_k^D\right) z_{i,j} (1 - z_{k,j}) + z_{ij} z_{kj} r_{ik}^U dist\left(x_i^D, x_k^D\right) + z_{ij} (1 - z_{kj}) r_{ik}^A dist\left(x_i^D, x_k^D\right) \right] \right] \quad (5)$$

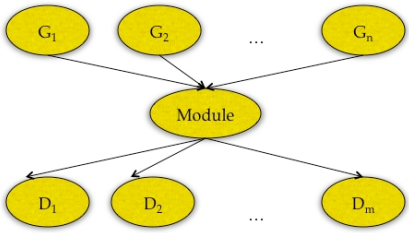
$$s.t. \quad \sum_{j=1}^{J} z_{ij} = 1 \quad \forall i; \qquad z_{ij} \in \{0,1\}$$

where z_{ik} is the decision variable representing to which cluster C_j^D the element x_i^D should be assigned. The optimization problem reported in equation (5) can be solved through an adaptation of K-means heuristic.

2.2 Bayesian Networks

For suggesting the most responsive drugs for a given cell line, we must take into account causal dependencies between the expression level of genes and the activity level of drugs.

An approach able to consider these causal relationships is represented by Bayesian Networks (BNs) [1]. The figure shows BN instantiation for gene-drug dependency analysis.



This structure of BN has been defined for inducing a probabilistic model able to predict the drug response of a new cell, only by providing its gene expression profile.

3. Experimental Investigation

In order to evaluate the quality of the clustering algorithm we used Pearson Correlation Coefficient (by considering gene expression profiles (P^G) and drug responses (P^D)), Entropy Measure (E^*) and F-Measure (F^*). For estimating the predictive power of BN we considered the number of drugs correctly predicted. In Table I we report a performance comparison over Ω^2, among the results of our relational clustering approach (RC), the traditional k-Means (KM) [3] and Soft Topographic Vector Quantization (STVQ) [2] algorithm. In particular, for STVQ algorithms, we report results obtained for the three different values of the tuning parameter α. Since all the evaluated algorithms depend on the initial choice of the representative element of each cluster (centroid), we show the average performance obtained over 1000 runs.

		P^G	P^D	F^*	E^*
KM		0.5147	0.8748	0.5231	1.0527
STVQ	α=0.0	**0.5573**	0.8646	0.5455	1.0233
	α=0.5	0.5394	0.8700	0.5307	1.0613
	α=1.0	0.4430	**0.8762**	0.5455	1.388
RC		0.5436	0.8665	**0.5619**	**0.9684**

Tab. I: Clustering Results

		Correctly Predicted Drugs
KM		11957
STVQ	α=0.0	11782
	α=0.5	11964
	α=1.0	12009
RC		**12292**

Tab. II:Computational results of Bayesian Networks

It is interesting to note that the best results w.r.t. P^D and P^G are obtained by STVQ with $\alpha = 0$ and $\alpha = 1$, albeit our approach is very close to these correlation values implying that the obtained groups of cell lines are homogeneous both from gene expression profile and drug activity response point of view. In this way, cell lines will likely respond similarly to the set of considered compounds thanks to their high drug and gene correlations. The quality of prediction of the induced network has been evaluated by counting the total amount of drug responses that, along the entire 60 cell lines, are correctly inferred. As highlighted in table II, the BN that obtains the quite encouraging result is the one trained with the clusters (modules) defined by the proposed relational clustering algorithm.

4. Conclusion

The experimental results show that the proposed clustering approach produces clusters (modules) that are homogeneous both in terms of gene expression and drug activity profiles, and its conjunction with BN represents an interesting research direction.

5. References

[1] Chang JH, Hwang KB, Zhang BT. Analysis of gene expression profiles and drug activity patterns by clustering and bayesian network learning. In Methods of Microarray Data Analysis II, chapter 11, 2002.

[2] Graepel T. et al.,"Self-organizing maps: generalizations and new optimization techniques" Journal of Neurocomputing, Vol. 21, pp.173-190, 1998.

[3] MacQueen J., Some methods for classification and analysis of multivariate observations. In Proc. of the 5th Berkeley Symp. on Mathematical Statistics and Probability, pp. 281–297, 1967.

[4] Scherf U et al., A gene expression database for the molecular pharmacology of cancer. J. of Nature Genetics 2000, 66: 236-244.