Relational Clustering for Gene Expression Profiles and Drug Activity Pattern Analysis

Elisabetta FERSINI[1], Cristina MANFREDOTTI[1], Enza MESSINA[1], Francesco ARCHETTI[1,2]

DISCo, Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126, Milano, Italy
{fersini, manfredotti, messina}@disco.unimib.it
Consorzio Milano Ricerche, Via Cicognara 7, 20129 Milano, Italy
{archetti}@milanoricerche.it

Abstract. The combined analysis of the micro array and drug-activity datasets has the potential of revealing valuable knowledge about various relations among gene expressions and drug activity patterns in malignant cells. However, the huge amount of biological data needs appropriate data mining models in order to extract interesting patterns and useful information. The ultimate goal of the paper is to define a model which, given the gene expression profile related to a specific tumor tissue, could help in selecting a set of most responsive drugs. This is accomplished through an unsupervised classification algorithm that associates to a cell line the set of drugs that most probably are related to its gene expression profile. The classification engine is based on a Relational Clustering algorithm which groups cell lines using drug response information and taking into account cell-to-cell relationships defined by the similarity of their gene expression profiles.

1. Introduction

Thanks to the recent progresses in biological experiment technologies, for example cDNA micro arrays, large amount of data have been collect. These evidences offer important opportunities to increase the knowledge related to complex biological phenomena. An important research field 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.

2. Motivation

The absence of a common pattern between gene expression profiles and drug-response, as showed in [1][2], might be partly due to the activity of genes related to drug sensitivity and resistance. This idea has been supported by the fact that several cell lines with a relatively high expression level of multi-drug resistance gene ABCB1 have been clustered in the same group. Consequently, this indicates that chemoresponse mechanisms are distributed across different tissues in the panel and that it should be possible to link drug activity to gene expression profiles. Several studies have been conducted with traditional distance-based clustering algorithms, in order to identify common patterns in gene expression and drug activities of cell lines. These algorithms, such as K-Means [3] and UPGMA [4], aim at partitioning data points into a pre-specified number of clusters through the minimization of a cost function related to a similarity/dissimilarity measure between the points without taking into account background information about pairs of instances for constraining their cluster placement. The conclusions drawn by previous investigations and the weakness of traditional distance-based algorithms lead us to the definition of a constraint-based clustering algorithm, following the idea of introducing specific domain constraints derived from background knowledge as presented in [5].

3. Methodology

The constraint-based algorithm proposed in this paper is aimed at assigning groups of cell lines using drug response information and taking into account cell-to-cell relationships defined by the similarity of their gene expression profiles.

Relational Constraints by the Induced Bisecting K-Means

A relation among two instances can be either an *affinity* or a *diversity* relationship and can be characterized by different degrees of intensity. In order to obtain these relations we cluster the cell lines on the basis of their gene expression profiles.

Let x be a cell line defined as a vector in R^{m+n}, where m and n represent the dimensions of the gene expression profiles and the drug activity space features respectively. Therefore we can define Ω as:

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

and Ω^G and Ω^D as the set of the cell lines $x \in \Omega$ represented through their gene expression profiles and their drug activity response respectively:

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

Given the elements $x_i^G \in \Omega^G$ and a set of clusters C_j^G with $j = 1:J$, the clustering problem consists in assigning each element x_i^G to a cluster C_j^G such that the intra-cluster distance is minimized and the inter-cluster distance is maximized. This assigning problem, which is known to be NP-Hard, is solved through an heuristic algorithm called Induced Bisecting K-Means [6] based on a distance measure defined as:

$$dist\left(x_i, x_k \right) = 1 - \cos ine_{ik} \quad (3)$$

The obtained set of clusters leads us to define two matrices of relations R^U and R^A that will be used in the subsequent clustering phase.

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. More formally, r_{ik}^U is defined as the distance between x_i^G and x_k^G, computed as in (3) i.e.

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

The 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. The element r_{ik}^A is given by

$$r_{ik}^A = \begin{cases} dist\left(x_i^G, x_k^G \right) & if\ x_i^G \wedge x_k^G \in C_\alpha^G \\ 0 & otherwise \end{cases} \quad (5)$$

Relational Constraints by the Induced Bisecting K-Means

The clustering task is aimed at minimizing the sum of the distances between elements, expressed by their drug activity response, penalized by a function that takes into account affinity and diversity constraints. Let $x_i^D \in \Omega^D$ be a given data point represented by its drug response features and C_j^D with $j = 1:J$ a set of clusters. Our Relational K-Means clustering aims at solving an optimization problem formulated as follows:

$$\min \sum_{j=1}^{J} \left[\sum_{i,k=1}^{|\Omega|} \left[dist\left(x_i^D, x_k^D \right) z_{ij} z_{kj} - dist\left(x_i^D, x_k^D \right) z_{i,j} \left(1 - z_{k,j} \right) + z_{ij} z_{kj} r_{ik}^U dist\left(x_i^D, x_k^D \right) + z_{ij} \left(1 - z_{kj} \right) r_{ik}^A dist\left(x_i^D, x_k^D \right) \right] \right] \quad (6)$$

$$s.t.$$

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

where z_{ik} is the decision variable representing to which cluster C_j^D the element x_i^D should be assigned.

In this way if an affinity-link (or a diversity-link) is not preserved, the objective function is penalized according to the weight r_{ik}^A (or r_{ik}^U) and the distance between cell lines *i* and *k* represented in terms of their drug responses features, $dist\left(x_i^D, x_k^D \right)$.

4. Results

The NCI60 dataset is composed by 60 cell lines from 9 different cancers, all extracted from human patients. These 60 cell lines have been characterized using two matrices: ***T*** matrix describes gene expression profiles and

A matrix describes in vitro drug sensitivity profiles. In particular the ***T*** matrix represents each cell line using 1375 genes with strong pattern of variation among cell lines, extracted by cDNA micro-arrays, and with less than 5 missing values. The ***A*** matrix represents each cell line by using the drug-response values of 1400 chemical compounds, tested one at time and independently; these drug activities are described through the Sulphorhodamine B assay, monitoring the changes in total cellular protein after 48 hours of drug treatment.

In order to evaluate the quality of the proposed relational clustering algorithm, we used the widely adopted average Pearson Correlation Coefficient defined as:

$$\overline{P} = \sum_{j=1}^{J} \frac{|C_j|}{|\Omega|} \sum_{i=1}^{|\Omega|} \sum_{k=1}^{|\Omega|} \frac{corr_{ik}}{|C_j|^2} z_{ij} z_{kj} \quad (7)$$

where J is the number of clusters and C_j is a cluster obtained by our relational clustering process. We estimate $\overline{P}$ in two ways: in one case we computed $\overline{P}^G$ considering the $corr_{ik}$ between instances i and k represented by their gene expression profiles and in another case $\overline{P}^D$ using their drug response profiles. To demonstrate the efficacy of the proposed algorithm we applied the relational K-Means to NCI60 dataset, comparing our results, as reported in table 1, to that obtained by Chang et al. in [2] using the Soft Topographic Vector Quantization approach (STVQ).

	Relational K-Means	STVQ with M=9	STVQ with M=16
$\overline{P}^G$	**0.34**	0.26	0.32
$\overline{P}^D$	**0.65**	0.19	0.25

Table 1: Performance comparison

Even if we set the number of clusters J in the Relational K-Means equal to 9 or equal to 16, our algorithm converges, due to the force of the constraints, to a solution that provides 6 groups of elements. The average Person Correlation Coefficients reported in table 1, for the STVQ represents the best performance obtained over 10 runs. With respect to our approach it is interesting to note that Relational K-Means converges to a partitioning solution with a steady average Pearson Correlation Coefficient. In this case the clustering results of our approach are independent to the random choice of the initial representative elements, proving that the relational constraints are fundamental to reach the convergence that ensures a good overall quality. Results show that considering relational constraints brings to the definition of clusters that are homogeneous both in terms of gene expression and drug activity.

5. Conclusion

In this paper, the NCI60 dataset has been analyzed for the molecular pharmacology of cancer. The experimental results show that the proposed method outperforms the state-of-the-art methods. In particular, our clustering algorithm brings to the definition of clusters that are homogeneous both in terms of gene expression and drug activity profiles. The defined model could help in selecting a set of most responsive drugs given the gene expression profile related to a specific tumor tissue.

6. Reference

[1] U. Scherf, D.T. Ross, M. Waltham, L.H. Smith, J.K. Lee, L. Tanabe, K.W. Kohn, W.C. Reinhold, T.G. Myers, D.T. Andrews, D.A. Scudiero, M.B. Eisen, E.A. Sausville, Y. Pommier, D. Botstein, P.O. Brown, J.N. Weinstein: A gene expression database for the molecular pharmacology of cancer. Nat Genet. 2000; 24:236–244. doi: 10.1038/73439.

[2] J.H. Chang, , K.B. Hwang, B.T. Zhang: Analysis of gene expression profiles and drug activity patterns by clustering and Bayesian network learning. Methods of Microarray Data Analysis II, chapter 11, pp. 169-184, Kluwer Academic Publishers, 200.

[3] J. B. MacQueen: Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297.

[4] P. Sneath, R. Sokal. Numerical Taxonomy: the principles and practice of numerical classification. W. H. Freeman, San Francisco, 1973.

[5] S. Basu, M. Bilenko, R. J. Mooney: A probabilistic framework for semi-supervised clustering. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), pp. 59-68, Seattle, WA, August 2004.

[6] F, Archetti, P. Campanelli, E. Fersini, E. Messina: A Hierarchical Document Clustering Environment Based on the Induced Bisecting k-Means. FQAS, (2006) 257-269.