Gene Ontology Semi-supervised Possibilistic Clustering of Gene Expression Data

Ioannis A. Maraziotis[1], George Dimitrakopoulos[1,2], and Anastasios Bezerianos[1]

[1] Department of Medical Physics, School of Medicine,
University of Patras, 26500, Patra, Hellas
[2] Department of Electrical and Computer Engineering,
University of Patras, 26500, Patra, Hellas
imaraziotis@gmail.com, {geodimitrak,bezer}@upatras.gr

Abstract. Clustering is one of the most important data analysis methods with applications of significant importance in many scientific fields. In computational biology, clustering of gene expression data from microarrays assists biologists to investigate uncharacterized genes by identifying biologically relevant groups of genes. Semi-supervised clustering algorithms have proven to bring substantial improvements in the results of standard clustering methods especially on datasets of increased complexity. In this paper we propose a semi-supervised possibilistic clustering algorithm (SSPCA) utilizing supervision via pair-wise constraints indicating whether a pair of patterns should belong to the same cluster or not. Furthermore we show how external sources of biological information like gene ontology data can provide constraints to guide the clustering process of SSPCA. Our results show that the proposed algorithm outperformed other well established standard and semi-supervised methodologies.

Keywords: possibilistic clustering, semi-supervision, constraints, gene ontology, gene expression.

1 Introduction

Clustering was, and still remains one of the most popular methods for the analysis of gene expression from microarray experiments, used to provide insight into the structure of the data and to aid at the discovery of biologically relevant groups of genes.

Initial computational efforts employed classical clustering techniques [1] for grouping genes according to their expression profile, based on the experimentally validated assumption that genes involved in the same biological process exhibit similar patterns of variation. In most of the cases however, certain peculiarities of the gene expressions at hand, like the large degree of complexity in the measured entities and the amount of inherent noise present in microarray experiments, prevent standard clustering methods to provide adequate results in terms of pattern similarity and biological correlation. Following several studies in the field of functional genomics showing the advantages of integrating different types of biological data [2], a solution in improving clustering results of microarray data would be to incorporate additional sources of

I. Maglogiannis, V. Plagianakos, and I. Vlahavas (Eds.): SETN 2012, LNAI 7297, pp. 262–269, 2012.

biological information [3]. An algorithmic family that could utilize prior knowledge on a certain field, is semi-supervised algorithms. Partially supervised clustering methods stand between purely unsupervised and fully supervised methods, benefiting from the advantages of both.

Algorithms performing semi-supervised clustering have recently received a significant amount of interest in the machine learning and data mining communities. It has been shown that even a relatively small amount of supervision significantly improves the accuracy of clustering [3]. Existing methods for semi-supervised clustering can be divided into two general categories known as constraint-based and metric-based approaches. In the metric-based approach an existing clustering algorithm is employed, but the measure of distortion used by this algorithm is first trained to satisfy the labels or constraints in the supervised data. On the other hand in constraint-based methods, the clustering algorithm itself is modified so as to integrate the user-provided labels or constraints, constituting this way a more suitable approach since its operation does not constitute of two steps but it is integrated in a single process.

Given the above considerations, and following the constraint-based approach we propose a novel Semi-Supervised Possibilistic Clustering Algorithm (SSPCA). SSPCA extends the operation of possibilistic clustering [4] in the semi-supervised field, considering sets of constraints either forcing patterns/genes to cluster together or assigning them to different clusters. We apply SSPCA on the intrinsic problem of gene expression clustering. Furthermore we show how external sources of information can guide the selection of constraints. While several types of biological data could serve as a source of external information; Gene Ontology (GO) Consortium currently serves as the dominant approach for machine-legible functional annotation.

Experimental results on real and artificial data prove not only the efficiency of the proposed SSPCA against other clustering algorithms but also the advantages of using external sources of biological information (i.e. GO) in clustering gene expression data.

2 Methods

2.1 Constraints and Semi-supervision

In the proposed methodology additional information (or prior knowledge) on a specific domain, is given on sets of either must-link or cannot-link constraints or both. Let E be the set of must-link constraints to be given in pairs $(x_i, x_j) \varepsilon E$ where the instances x_i and x_j should be assigned to the same cluster, while cannot-link constraints in pairs $(x_i, x_j) \varepsilon \Delta$ where Δ is the set of cannot-link constraints and x_i, x_j should be assigned to different clusters. In the approach we are adapting a specific gene can be associated with more than one pair and kind of constraints. We could for example have three constraints that would impose a gene to be in the same group/class with some other three genes, while at the same time to belong to different classes with another pair of genes. Therefore we will insert a new metric that will calculate the number of

constraints that are retained for a specific gene *j* and the number of violations over the constraints within a certain cluster *i*:

$$\beta_{ij} = \frac{V_{ij} - R_{ij}}{T_j} \tag{1}$$

where T_j is the total number of given constraints concerning the element or gene *j*, R_{ij} are the pairs of constraints that are preserved within the cluster, while V_{ij} is the number of constraints that are violated within the same cluster concerning sample *j*. As we can determine from eq. (1), the range of the score of every member of the cluster is within -1 and 1. Specifically concerning the range of value for β_{ij}, it is:

$$\beta_{ij} = \begin{cases} 1, & \{R_{ij} \to 0 \wedge V_{ij} \to T_j\} \\ 0, & \{R_{ij} = V_{ij} \vee T_j = 0\} \\ -1, & \{R_{ij} \to T_j \wedge V_{ij} \to 0\} \end{cases} \tag{2}$$

When β_{ij} approaches 1, then most of the constraints regarding the specific pattern are violated within the cluster, while when the score approaches -1, most of the constraints are preserved. At this point we should note that the score becomes zero, if there are no constraints regarding a specific pattern at the dataset of supervision and that approaches zero in the case that the percentage of constraints that are violated equals the number of constraints that are retained. We will expand now the metric proposed in (1), in order to account for the validity of a cluster in terms of retained constraints for all of its members:

$$\mathrm{B}_i = \frac{1}{2}\left(1 - \frac{1}{N_C} \cdot \sum_{j=1}^{N_C} \frac{R_{ij} - V_{ij}}{T_j}\right) \tag{3}$$

where N_c is the number of the members of a cluster *i*. In contrast to the previously proposed metric, the score of a cluster ranges from 0 to 1, where 1 is the case for which there is no violation for any of the members of the group/cluster regarding the constraints known for it, while 0 is the exact opposite case. As we will see later in the analysis of those two metrics will play a central role in SSPCA algorithm.

2.2 Gene Ontology and Constraints Selection

We present a framework (Fig. 1) for selecting constraints from gene ontology terms that will be used as input to SSPCA to guide the clustering process of gene expression data. The Gene Ontology Consortium is one of the most widely used database concerning annotations of gene functions. The number of genes associated with a certain annotation term indicates how specific that term is, therefore based on this criterion we could discriminate between general and more specific terms. Therefore two genes sharing a more specific term are more likely to interact than genes that share a general term. While, there are many GO measures in the literature that provide a quantitative degree of similarity between two specific genes in respect to their GO terms, Resnick's [5] is one of the most widely used.

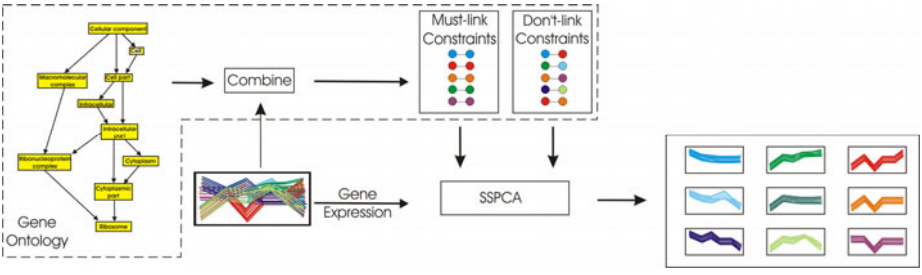

Fig. 1. Schematic representation of the framework we adapt to extract constraints, based on gene ontology terms, used to guide the clustering process of SSPCA on gene expression data

However, the main goal of the proposed method is clustering, hence along with the GO information we must also consider the similarity of the expression profiles for a certain couple of genes. Hence we will insert a measure that will take under consideration both of the aforementioned criteria:

$$G_{ij} = \frac{d_{ij}^2 / s_{ij}}{\sum_{k=1}^{N} d_{ik}^2 / s_{ik}}, \quad \forall j \in [1 \ldots N] \tag{4}$$

where i and j correspond to certain genes, d_{ij} is the Euclidean distance between the expression profiles of i and j while s_{ik} is Resnik's similarity measure. As we can depict from (4) the reverse GO similarity between two genes is weighted by their corresponding euclidean expressional distance, since to determine a certain constraint we take under consideration gene expression as well. We finally normalize by the total sum of these scores of i against all other genes. While s_{ij} ranges from 0 to a maximum value, having 0 as worst case the opposite occurs for Euclidean distance. Hence the range of the proposed measure ranges from 0 to a maximum value, having 0 as the best case.

In order to extract the necessary constraints from a given data set, we adapt a methodology where every one of the genes present in the dataset under study, is cross-checked against all others. Each one of these pairs is given a similarity degree based on (4). After the process is concluded for all genes, we sort the similarities degrees of the corresponding pairs. A specific fraction of the pairs that have achieved a minimum score will be used as must-link constraints, while the ones that have the largest values will be used as cannot-link constraints. The exact percentage of the constraints is given as input to the algorithm.

2.3 SSPCA

In this section we will describe the operation of the proposed SSPCA algorithm. SSPCA through its objective function tries to minimize the distance among the patterns and the corresponding centroids of the clusters while at the same time is guided by the pairs of constraints towards the determination of more concise clusters. The

mathematical description of the SSPCA objective function expanding the operation of possibilistic clustering in the semi-supervised field is:

$$J_1(U,V;X,\Pi)=\sum_{i=1}^{C}\sum_{j=1}^{N}a_{ij}^m d_{ij}^2+\sum_{i=1}^{C}\sum_{j=1}^{N}\left(1-a_{ij}\right)^m\left(\gamma_i+D_i d_{ij}^2\right) \tag{5}$$

where:

$$D_i=(1-n)\frac{N_T}{N_C}B_i \quad \text{and} \quad \gamma_i=n\frac{\sum_{k=1}^{N}\alpha_{ik}^m d_{ik}^2}{\sum_{k=1}^{N}\alpha_{ik}^m} \tag{6}$$

N_T represents the number of patterns in a specific cluster and N_C the number of patterns that are part of both the cluster and the constraints data set. The term N_T over N_C ensures that if a small number of constraints is provided in comparison to the total number of patterns in the dataset, these constraints will not dominate the overall clustering process. We will discuss about the effect of the parameters D_i later in the text, while and γ_i is in accordance to the corresponding value described in [4]. The variable a_{ij} is a function of u_{ij} and β_{ij}, assuming a small positive number n, ranging from zero to unity, we set:

$$a_{ij}=n\cdot u_{ij}-(1-n)\cdot\beta_{\phi(ij)} \tag{7}$$

In every iteration the membership value of a_{ij} depends not only on the distance x_j from v_i but also from the number of constraints that are retained or violated concerning a specific pattern j. As we can depict from (6) n is a parameter controlling the degree that constraints will be taken into account in the overall clustering process and can either have a fixed value throughout all the clustering process or can vary during the iterations steps. In cases that we are confident for the accuracy of the constraints and also have a satisfying number of constraints for the majority of the dataset then n can be viewed as a constant whose value can have a small value (i.e. ranging from 0.4 to 0.6) that reflects the quantity and confidence of the semi-supervised information we have. In the case however where either we do not have a satisfying number of constraints for the data set and/or we have either a minimum amount of confidence or even uncertainty concerning the accuracy of the constraints, parameter n could be regarded as a time/iterations dependent variable. In this paper we will study the first case only.

In possibilistic and fuzzy clustering, each pattern is a member of all existing clusters up to a certain degree indicated by the corresponding membership value and hence all pairs of patterns constraints should be checked for violations throughout all clusters. We can however consider that a pattern is part of the cluster for which it has the maximum membership value. Hence, we can check for pattern violations, concerning one cluster per pattern. This is accomplished using a function, which rewards or penalizes membership values only in the case of cluster members, while in the opposite case eliminates the influence of the constraints. Using this technique we

reduce the computational complexity, contribute to the faster convergence of the algorithm and at the same time not harming the generality of the solution. Given all the above, we introduce φ as a function returning the set of patterns that have their maximum membership value within a certain cluster (i.e. compared to other clusters) at every iteration. The mathematical interpretation of function φ is:

$$\phi(i)=\begin{cases}0, & \text{if } \left\{\arg\left(\max_k u_{k\rho}\right)\equiv i \,\middle|\, \forall\rho\in \mathrm{D}\right\}\\ 1, & \text{else}\end{cases} \tag{8}$$

$\varphi(i)$ is a function returning the set of patterns that have their maximum membership value is within the i-th cluster at every iteration. Given φ we have that:

$$\beta_{\phi(i,j)}=\begin{cases}\beta_{\phi(i)j}, & \mathrm{j}\in \mathrm{C}\\ 0, & \text{else}\end{cases} \tag{9}$$

Using partial derivatives and the method of Largange multipliers we solve (5) in respect to membership values u_{ij} and the centroids v_i, hence:

$$u_{ij}=\frac{1/n}{1+\left(\frac{d_{ij}^2}{D_i d_{ij}^2+\gamma_i}\right)^{\frac{1}{m-1}}}+\frac{1-n}{n}\beta_{\varphi(i)j} \tag{10}$$

while for the centroids we have:

$$v_i=\frac{\sum_{j=1}^{N}\left[a_{ij}^m+D_i\left(1-a_{ij}\right)^m\right]x_j}{\sum_{j=1}^{N}\left[a_{ij}^m+D_i\left(1-a_{ij}\right)^m\right]} \tag{11}$$

As we can depict from (9) the value of u_{ij} is highly influenced by the value of β_{ij} and D_i. Based on the definition given in (1) and range in (2) we can depict that in order for the value of u_{ij} to be increased the majority of the constraints regarding the j-th pattern in i-th cluster must be retained. Also the value of constraints has a zero effect when the number of violations equals the number of retentions. On the other hand if D_i is high then u_{ij} will be high, if D_i is low then u_{ij} will be low. Indeed as we discussed in the previous section based on (3) the value of D_i (from 0 to unity) increases as the number of patterns for which the majority of constraints is retained increases within the a certain cluster constraints that are retained increases.

3 Results

In this section we will describe the experiments we conducted to test the validity of our approach, based on both artificial and real data sets. In the followings, we com pared the apodosis of SSPCA against two standard clustering techniques PK-means [4] and K-means as well as a semi-supervised method CPK-means [6].

Table 1. Results on the apodosis of the standard clustering techniques PK-means, K-means and the semi-supervised method CPK-means in comparison to the proposed SSPCA

Algorithm	Supervision	ARI	
		DS1	DS2
SSPCA	0	0.39	0.30
	10	0.57	0.50
	30	0.84	0.69
CPK-means	30	0.580	0.517
PK-means	-	0.39	0.30
K-means	-	0.460	0.362

In order to test our method under more controlled conditions we resorted to artificial data (hereafter DS1). This dataset has been artificially created initiating from real data as described in [7]. It consists of 400 patterns across 10 different experimental conditions. The dataset has 10 clusters. The second dataset was based on an experimental study published in [8], consisting of the expression levels of more than 6000 genes measured across 17 time points during two cell cycles of Saccharomyces cerevisiae (SS). From this study we have used a subset of 384 annotated genes (DS2) visually identified as five distinct time points, each one representing a phase of the SS cell cycle. The expression levels of each gene were normalized to zero mean and unity standard deviation.

The constraints for DS1 were extracted by considering the known labels of the patterns and following the methodology described in [3], while for DS2 we acquired constraints as described in previous section. In this study we have used information on the GO domain: molecular function, for the SS micro-organism. Given that the labels in both datasets considered are a priori known we have the adjusted rand index (ARI) metric [9] to measure the efficiency of the considered algorithms. A value of ARI equal to 1 indicate a perfect clustering according to the provided pattern labels while a value of zero the opposite. Clustering was repeated 10 times for the data sets under consideration, by all the algorithms checked and the mean values of the results in terms of adapted metric were used. As we have already mention the key parameter in the operation of SSPCA is n that controls the influence of the provided constraints in the overall clustering process. We have repeatedly executed SSPCA for the following range of n values: 0.4, 0.45, 0.50, 0.55, 0.6. The best results, reported on Table 1, were acquired for a value of n equal to 0.6.

As we can depict from Table 1, the results of SSPCA and PK-means is the same when the percentage of provided constraints equals zero, since PK-means is a special case of SSPCA in the non-supervised field. As we can see on the table the proposed algorithm outperformed both of the considered unsupervised algorithms for a small percentage of supervision (10%). Finally for the same percentage of provided constraints, SSPCA had more than 25% and 30% improved apodosis in DS2 and DS1, respectively than the semi-supervised algorithm CPK-means.

The reported results not only demonstrate the efficiency of SSPCA and the benefits of semi-supervised over standard clustering methods but also indicate the advantages of using external sources of biological information to guide the clustering of gene expression data.

4 Conclusions

In this work we presented a semi-supervised possibilistic clustering algorithm incorporating prior-knowledge in the form pair-wise constraints. Initial results suggested that the proposed algorithm outperformed other crisp and fuzzy methods. Furthermore we showed that under the semi-supervised framework adapting external sources of biological information, such as GO, for constraints selection, can significantly improve the clustering results.

We are working in methodologies that will extend the operation of SSPCA by allowing the algorithm to automatically extract a meaningful number of clusters. Additionally, we are performing a wide range of additional simulations from gene expression data arriving from several organisms and other sources of biological knowledge (i.e. protein-protein interactions) to further validate the findings of this study.

Acknowledgments. This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALES. Investing in knowledge society through the European Social Fund.

References

1. Wu, L.F., et al.: Large scale prediction of Saccharomyces cerevisiae gene function using overlapping transcriptional clusters. Nat. Genetics 31, 255–265 (2005)
2. Maraziotis, I.A., Dimitrakopoulou, K., Bezerianos, A.: An in silico method for detecting overlapping functional modules from composite biological networks. BMC Systems Biology 2, 93 (2008)
3. Maraziotis, I.A.: A Semi-supervised algorithm applied on gene expression data. Pattern Recognition 45(1), 637–648 (2012)
4. Krishnapuram, R., Keller, J.M.: A possibilistic approach to clustering. IEEE Trans. on Fuzzy Systems 1(2) (1993)
5. Resnik, P.: Using information content to evaluate semantic similarity in taxonomy. In: Proc. of Int. Joint Conf. on Artificial Intelligence, pp. 448–453 (1995)
6. Wagstaff, K., Cardie, C., Rogers, S., Schroedl, S.: Constrained K-Means clustering with background knowledge. In: Proceedings of 18th International Conference on Machine Learning, pp. 577–584 (2001)
7. Yeung, K.Y., Haynor, D.R., Ruzzo, W.L.: Validating clustering for gene expression data. Bioinformatics 17, 309–318 (2001)
8. Cho, R.J., et al.: A genome-wide transcriptional analysis of the mitotic cell cycle. Molecular Cell 2, 65–73 (1998)
9. Yeung, K.Y., Ruzzo, W.L.: An empirical study on principal component analysis for clustering gene expression data. Bioinformatics 17, 763–774 (2001)