Mutual Spectral Clustering: Microarray Experiments Versus Text Corpus

K. Pelckmans[1], S. Van Vooren[1], B. Coessens[1], J.A.K. Suykens[1], and B. De Moor[1]

K.U. Leuven, ESAT, SCD/SISTA, Kasteelpark Arenberg 10, B-3001, Belgium
e-mail: kristiaan.pelckmans@esat.kuleuven.be
WWW home page: http://www.esat.kuleuven.ac.be/scd/

Abstract. This work[1] studies a machine learning technique designed for exploring relations between microarray experiment data and the corpus of gene-related literature available via PubMed. The use of this task is found in that it provides better clusters of genes by fusing both information sources together, while it can also be used to guide the expert through the large corpus of gene-related literature based on insights into microarray experiments and vice versa. The learning technique addresses the unsupervised learning problem of finding meaningful clusters co-occurring in both knowledge-bases. Here, one is typically interested in whether the membership of an instance to one cluster in the former knowledge-base transduces to membership of the same instance to the corresponding cluster in the latter representation. This idea is described as an extended MINCUT problem and implemented using a spectral clustering technique possessing a well-defined out-of-sample extension.

1 STATEMENT OF THE LEARNING PROBLEM

In order to emphasize the peculiarity of the investigated learning setting, the problem is at first stated in an abstract way. Let $\{(X_i, Z_i)\}_{i=1}^n \subset \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ be iid sampled from the joint distribution F_{XZ}, for given $d_1, d_2, n \in \mathbb{N}$. Let $K < n$ be an appropriate constant. The following learning problem is studied: learn a mutual clustering $\mathcal{C}^{12,K} = \{(\mathcal{C}_k^1, \mathcal{C}_k^2)\}_{k=1}^K$ such that the following relation holds with high probability

$$(X, Z) \sim F_{XZ} : \ \mathcal{C}_k^1(X) \Leftrightarrow \mathcal{C}_k^2(Z), \ \ \forall k = 1, \ldots, K. \tag{1}$$

The relevance of this mutual clustering $\mathcal{C}^{12,K}$ is seen as follows: if one observes a new value $X_* \in \mathbb{R}^{d_2}$ which belongs to C_k^1, one can assert with high probability that this instance will belong to C_k^2 in the alternative representation (and vice versa). This method can be used for example to predict missing values based on an unsupervised dataset: if a random variable X_i is not observed due to reasons of independency, the membership of the observed Y_i can be used to infer partial knowledge - namely the membership to the corresponding cluster - in the latter representation. This question does not coincide with classification as it is symmetrically valid: the random variable X plays the role of labels as well as covariates for Y and vice versa, while the class assignments are not given a priori. The task differs from

[1] (KP): BOF PDM/05/161, FWO grant V4.090.05N; - (JS) is an associate professor and (BDM) is a full professor at K.U.Leuven Belgium, respectively. (SCD:) GOA AMBioRICS, CoE EF/05/006, (FWO): G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0321.06, G.0553.06, (ICCoS, ANMMM, MLDM); (IWT): GBOU (McKnow), Eureka-Flite2 IUAP P5/22,PODO-II,FP5-Quprodis; ERNSI;

clustering as it possesses an explicit objective. This problem has formal relations to the task of semi-supervised learning and transductive inference, see e.g. [5], while its use is situated in a purely exploratory data analysis setting useful for unsupervised data-mining problems.

2 MUTUAL SPECTRAL CLUSTERING

The discussion is cast in a context of graph cuts as the entities under study (genes in this case) are discrete by nature, and as it is not clear what an underlying distribution F_{XY} would mean. Given two graphs $\mathcal{G}^{(1)} = (\mathcal{N}, \mathcal{E}^{(1)})$ and $\mathcal{G}^{(2)} = (\mathcal{N}, \mathcal{E}^{(2)})$ which share the same nodes $\mathcal{N}$ (think of any node $\mathcal{N}_*$ as a representation of a single gene, e.g. 'P53'). Let the positive weights $\mathcal{E}^{(1)} = \{w_{ij}^{(1)}\}_{i \neq j}$ and $\mathcal{E}^{(2)} = \{w_{ij}^{(2)}\}_{i \neq j}$ be associated with $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ respectively based on the two different knowledge-bases. Let $w_{ii}^{(1)}$ and $w_{ii}^{(2)}$ be zero for all $i = 1, \ldots, n$. Let $\pi_1, \pi_2 > 0$ represent the relative importance or confidentiality of the two representations $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$. The following approach is based on an additive argument: the performance of a mutual clustering is essentially expressed as the sum of the performances of the clustering on both individual graphs. We start by explicitly defining a neighbor-based rule for deciding whether a node $\mathcal{N}_*$ (with edges $\{w_*^{(*j)}\}_{j=1}^n$ and $\{w_{*j}^{(2)}\}_{j=1}^n$) belongs to the former class (denoted as $q = -1$) or of the latter ($q = 1$): thus for $\mathcal{G}^{(1)}$, the decision rules become

$$\begin{cases} R^1(\mathcal{N}_*; q) = \operatorname{sign}\left(\sum_{j=1}^n q_j w_{*j}^{(1)}\right) \\ R^2(\mathcal{N}_*; q) = \operatorname{sign}\left(\sum_{j=1}^n q_j w_{*j}^{(2)}\right). \end{cases} \tag{2}$$

Now it can be proven that the MINCUT results in a vector $q \in \{-1, 1\}^n$ which yields decisions using the above rules which are maximally consistent with the labeling itself. This argument can be made precise, but for clarity of explanation we give only the resulting learning problem and its spectral approximation.

Proposition 1 (Mutual Spectral Clustering) *Let $q_i = 1$ if the i-th node of $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ belongs to a cluster $(\mathcal{C}_k^{(1)}, \mathcal{C}_k^{(2)})$ for fixed k, and $q_i = -1$ otherwise. The size of the cut in both $\mathcal{G}^{(1)}$ and $\mathcal{G}^{(2)}$ corresponding with the assignment $q \in \{-1, 1\}^n$ is then minimized by*

$$\min_{q \in \{-1,1\}^n} \mathcal{J}_{\pi_1,\pi_2}(q) = \frac{\pi_1}{4} \sum_{i \neq j} w_{ij}^{(1)} (q_i - q_j)^2 + \frac{\pi_2}{4} \sum_{i \neq j} w_{ij}^{(2)} (q_i - q_j)^2. \tag{3}$$

Let the extended Laplacian be defined as $L_{\pi_1,\pi_2} = (\pi_1 D^{(1)} + \pi_2 D^{(2)}) - (\pi_1 W^{(1)} + \pi_2 W^{(2)}) \in \mathbb{R}^{n \times n}$ possessing the same properties as the individual Laplacians $D^{(1)} - W^{(1)}$ and $D^{(2)} - W^{(2)}$. This combinatorial optimization problem can be approximated by the spectral problem

$$L_{\pi_1,\pi_2} q = \lambda q, \tag{4}$$

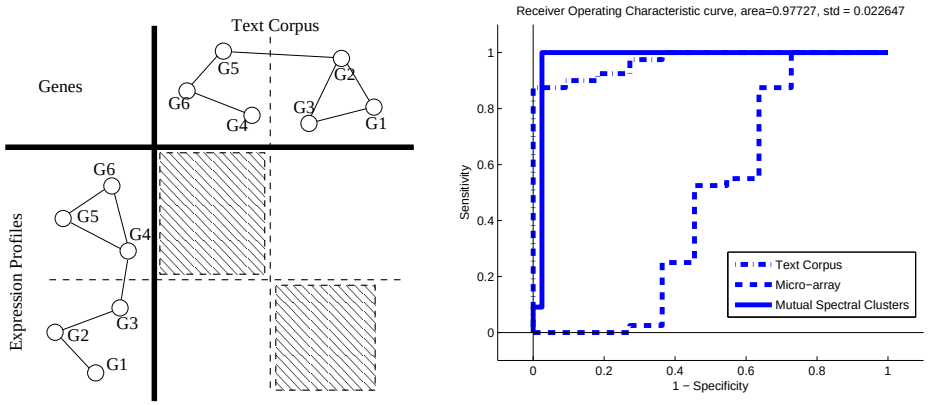
where $\lambda \in \mathbb{R}^+$ is the associated Lagrange multiplier.

Proof: The derivation follows [2]. Let the degrees be defined as $d_i^{(1)} = \sum_{j=1}^n w_{ij}^{(1)}$ and let $d_i^{(1)} = \sum_{j=1}^n w_{ij}^{(1)}$. Problem (3) can be written equivalently as

$$\min_{q \in (-1,1)^n} \mathcal{J}_{\pi_1,\pi_2}(q) = \frac{1}{4} q^T \left((\pi_1 D^{(1)} + \pi_2 D^{(2)}) - (\pi_1 W^{(1)} + \pi_2 W^{(2)}) \right) q \tag{5}$$

subject to $q \in \{-1, 1\}^n$. Replacing the integer constraint by the norm constraint $q^T q = 1$ yields the familiar spectral formulation (4). where the eigenvector associated with the lowest eigenvalue is the trivial $q = c(1, \ldots, 1)^T \in \mathbb{R}^n$ with constant $c = \pm\sqrt{n}$. Then it is known that the lowest nontrivial eigenvector $q(2)$ associated with the second lowest eigenvalue $\lambda_{(2)}$ is a continuous approximation to (3). ■

This reasoning provides a complete answer how to extend the mutual clustering to label out-of-sample examples using (2). A refinement of the method and a discussion of a normalized cut method as in [6] is under investigation. A clearcut analyzes of the learning algorithm is possible due to the clear definition of a criterion, and the definition of a rule underlying the analysis which describes extensions to new nodes.



(a) (b)

Fig. 1. (a) *Graphical presentation of mutual clustering. The objective is to find a clustering of the shared nodes which is consistent with both representations at the same time. This application shows the genes $G_1, \ldots, G_6$ represented using information extracted from microarray experiments (vertical) and using information retrieval techniques based on the PubMed corpus (horizontal). This example indicates that the mutual clustering can improve the min cut by fusing different data sources together.* **(b)** *Validation of the clustering based on the text corpus only, the microarray experiments only and the proposed technique for data fusion. The plot shows the ROC curves of the predicted cluster-membership of the different clustering methods versus the labels given in the gene ontology.*

3 MICROARRAY EXPERIMENTS VERSUS TEXT CORPUS

Both knowledge bases have a one-to-one correspondence: textual information and microarray experiments can be used to construct a gene graph. The text corpus can be organized in a gene graph as follows: graph $\mathcal{G}^{(1)}$ encodes the relation between genes based on the abstracts concerning this gene. The relation between genes is based on the distance between genes in

a classical term based vector space model [3]. Specifically, a gene is represented as the average term vector of the different citing abstracts. The graph weights are determined using the cosine rule applied to those terms. Graph $\mathcal{G}^{(2)}$ encodes the similarity between genes using information obtained from a series of microarray experiments [4]. To estimate the relations between genes based on the different experimental conditions, an RBF based scheme is used. Some preliminary experiments are conducted on a database of 51 different genes [7] concerning motor activity and visual perception. Figure 1.b shows the performance of a spectral clustering method using text data only, using microarray data only, and using the technique for integrating both knowledge-bases. The performance is expressed as a ROC curve measuring the correspondence of the predicted membership via the nearest neighbor rule (2), versus the labeling as given in gene ontology. This plot indicates that the proposed mutual clustering method can indeed improve the use of the learned clusters.

4 FURTHER ISSUES

Several further important issues need to be addressed. Important from a practical point of view is how to zoom in on small but coherent mutual clusters effectively representing functionally related genes. Further, it is important to extend the method of mutual spectral clustering based on neighborhood rules to multiple (overlapping) clusters. A related issue is how to validate the obtained clustering using biological experience as encoded in the gene ontology [1]. Moreover, the example described in the previous section indicates that the practitioner should bear the influence of weakly connected nodes in mind. It also emphasizes the importance of the choice of a proper method to infer a graph based on the observations. Important from a methodological point of view is a quantification of the probabilistic confidence in a learned mutual rule. Extensions to the data integration of multiple sources is straightforward in this setting, while large scale versions can straightforwardly incorporate results described in the large literature on large scale eigenvalue decompositions.

References

1. Gene Ontology Consortium. The Gene Ontology (GO) project in 2006. *Nucleic Acids Res*, 34(Database issue):322–326, Jan 2006.
2. M. Fiedler. A property of eigenvectors of nonegative symmetric matrices and its application to graph theory. *Czech. Math. J.*, 25(100):619–633, 1975.
3. G. Salton, A. Wong, and C.S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, 1975.
4. M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, Oct 1995.
5. J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.
6. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 22(8), aug. 2000.
7. A. I. Su, T. Wiltshire, S. Batalov, H. Lapp, K. A. Ching, D. Block, J. Zhang, R. Soden, M. Hayakawa, G. Kreiman, M. P. Cooke, J. R. Walker, and J. B. Hogenesch. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA*, 101(16):6062–6067, Apr 2004.