Predictive Mutual Cuts in Graphs: Learning in Bioinformatics

K. Pelckmans, J.A.K. Suykens, and B. De Moor

K.U. Leuven, ESAT, SCD/SISTA, Kasteelpark Arenberg 10, B-3001, Leuven, Belgium

Recent work of the authors[5] discussed a paradigm for learning correspondances between different graphs, emerging from a case-study in bioinformatics. Learning tasks in bioinformatics are often characterized by considerable uncertainty both in covariates as in the variable of interest, due to the biological origin of the data. One way to cope with uncertainty in data is to restrict attention to less precise statements, a notion which can be made formal as a set membership. We studied the task of finding sets (represented as clusters or equivalently as graph cuts) which have a one-to-one relation in different representations (mutual graph cuts). Specifically, the problem of mining for correspondances between microarray experiments and a relevant text corpus was studied. This study yields statements of the form *"if a gene is discussed in a specific cluster of texts -say $\mathcal{C}^1$- then it would probably possess an expression profile which fits in the corresponding cluster $\mathcal{C}^2$ of experiments"*, and vice versa.

More formally, let for all $i = 1, \ldots, n$ a gene v_i be represented as a couple (x_i, z_i) in G_1 and in G_2 respectively. Let those be sampled somehow iid from a joint distribution F_{XZ}. We now delve for a (nontrivial) couple of set indicators $(I_{\mathcal{C}^1}, I_{\mathcal{C}^2})$ based on G_1 and G_2 respectively, such that with high probability $I_{\mathcal{C}^1}(X) = I_{\mathcal{C}^2}(Z)$. This yields the risk function

$$\mathcal{R}(\mathcal{C}^1, \mathcal{C}^2) = \int \ell\left(I_{\mathcal{C}^1}(x) - I_{\mathcal{C}^2}(z)\right)\, dF_{XZ}(xz),$$

for an appropriate loss function ℓ. Note that this mechanism can be used to infer knowledge in case of missing information as in $(x, ?)$ or $(?, z)$. Statements of such problems - and various algorithms for solving them - appear to be omnipresent in many recent advanced machine learning applications, and various approaches to a formalization appear in recent publications, see e.g. [3, 4].

Here we discuss a formal framework to motivate such a statements, based on ideas from learning theory as described e.g. in [2]. This requires the formalization of various elements in the setup. First of all, a fixed but unknown stochastic mechanism is encoded through the choice of the weights of both graphs. A second necessary element is the definition of a predictor rule (out-of-sample extension operator) based on the labels of the given nodes and the connection weights. Thirdly, it is shown how the last implies the definition of an hypothesis space which can e.g. be bounded by an approach of maximal margin. Then it is indicated how the cardinality of this hypothesis space can be characterized in terms of properties of both graphs. The relation with Shannon capacity on the one hand, and VC dimension on the other is discussed. An approximate algorithm with polynomial time complexity is proposed and is related to classical results on max-flow problems [1].

References

1. R.K. Ahuja, T.L. Magnanti, and J.B. Orlin. *Network flows : theory, algorithms and applications*. Prentice Hall, 1993.
2. L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.
3. K. Fukumizu, F. Bach, and A. Gretton. Statistical convergence of kernel CCA. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 387–394. MIT Press, Cambridge, MA, 2006.
4. E. Krupka and N. Tishby. Generalization in clustering with unobserved features. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 683–690. MIT Press, Cambridge, MA, 2006.

5. K. Pelckmans, S. Van Vooren, B. Coessens, J.A.K. Suykens, and B. De Moor. Mutual spectral clustering: Microarray experiments versus text corpus. In *in Proc. of the workshop on Probabilistic Modeling and Machine Learning in Structural and Systems Biology*, pages 55–58. Helsinki University Printing House, Helsinki, Finland, 2006.