Robust Inference for Empirical Graphs: A Compression Approach

K. Pelckmans[(1)], **J.A.K. Suykens**[(1)], **B. De Moor** [(1)]
(1) K.U.Leuven - ESAT - SCD/SISTA,
Kasteelpark Arenberg 10, B-3001, Leuven (Heverlee), Belgium
kristiaan.pelckmans@esat.kuleuven.be

Abstract

Problems of learning over undirected weighted graphs often originate in empirical studies, implying almost inevitably uncertainty in the weights. This presentation argues fot the need of a dedicated learning algorithm when doing transductive inference for completing a partially observed labelling of the nodes. Two contributions are discussed: firstly, the concept of condensing and the role of compression are explored; and secondly, we indicate how the corresponding algorithm can be robustified against perturbations on the empirical weights of the graph under study. The ideas are illustrated by a realistic example application in predicting a user's interest in a set of books.

1 Introduction

This presentation considers the problem of transductive inference on a weighted undirected graph with weights which are estimated from observations. Transductive inference aims at completing (and correcting) an observed partial labeling. We found a key element in this discussion in the form of optimal condensing for weighted graphs, a problem which can be relaxed and robistified as a convex programming problem. The outcome of this approach is a subset of support nodes, compressing optimally the hypothesis consistent with the observations.

To illustrate the practical relevance of the ideas, we consider the following simple but challenging task in information processing: 'provided with a set of $n \in \mathbb{N}$ books indicated as $V = \{v_1, \ldots, v_n\}$, and suppose a subset $\mathcal{S} \subset \{1, \ldots, n\}$ of books which are either absolutely relevant ($y_i = 1$), either absolutely irrelevant ($y = -1$) for a certain person, how to predict the relevance of the remaining books?' Remark that such a question implies both the need for inferring a relation between the nodes (books), and simultaneously the inference of a particular labeling. It is common to apply a two-stage procedure: namely the unsupervised quest for a convenient organization of the nodes (e.g. by estimating the link weights), and consequently the application of a learning algorithm operating under assumption that the examples are organized in a deterministic way. However, the choice of the method to infer the link weights - and the graph topology - from the books is often arbitrary, and even worse, prone to uncertainty. E.g. in the context of the aforementioned example, it is often argued to take a simple bag of words rule for estimating the graph weights for the first stage. The message of this work is that the algorithm for solving the

overall inference problem should properly take care of the uncertainty induced by the graph construction.

2 Transductive Inference and Optimal Condensing

Let $\mathcal{G} = (V, E)$ where $V = \{v_1, \ldots, v_n\}$ and $E = \{x_{ij} \geq 0\}_{ij=1}^n$ denote the weighted undirected loopless graph such that $x_{ij} = x_{ji}$ and $x_{ii} = 0$ for all $i, j = 1, \ldots, n$. In [12, 2, 3, 10, 11] it was indicated how the performance of transductive inference techniques depend critically on the size of the hypothesis set, or the number of different predictions it could make. Formally, we adopt the view that the hypothesis set consists of all allowable labelings of the n given nodes, or $\mathbb{H} = \{q \in \{-1, 1\}^n\}$. It is now seen that one has to devise a proper restricted hypothesis space $\mathbb{H}'$ where $|\mathbb{H}'| \ll |\mathbb{H}|$. Consider the weighted neighbor predictor rule $r_q : V \to \{-1, 1\}$ as follows

$$r_q(v_i) = \text{sign}\left(\frac{1}{n}\sum_{j=1}^{n} x_{ij} q_j\right) = \text{sign}(q^T x_i), \tag{1}$$

which can be used to device a restricted set $\mathbb{H}'$ requiring that the rules prediction should be *consistent* with the actual label, or $q_i r_q(v_i) \geq 0$ for all $i = 1, \ldots, n$. In [11], it was argued that this argument together with a principle of maximal average margin results in the familiar MINCUT approach, and an effective characterization of the corresponding hypothesis space $|\mathbb{H}'_\rho|$ was given for fixed average margin $\rho > 0$. This argument was expressed in terms of the spectral properties of the graph, while it was seen that the problem could be solved efficiently as a MAX-FLOW problem with complexity $O(n^2)$.

Here we take a slightly different approach, which might be more suitable for other applications. Therefor, we adopt again the weighted neighborhood rule (1); then we seek the smallest number of nodes such that such that the observed labels are consistent with the prediction rule r_q based solely on the subset. A similar principle is also known in the context of K-nearest neighbors as *condensing*, see e.g. [5], chapter 19 and references. We refer to those nodes as to the *support nodes*, in analogy to the support vectors in the popular SVM technique.

$$\hat{q} = \underset{q \in \{-1,0,1\}^n}{\arg\min} \; \mathcal{J}^0(q) = \sum_{i=1}^{n} I(q_i \neq 0) \quad \text{s.t.} \quad y_i \left(\sum_{j=1}^{n} x_{ij} q_j\right) > 0, \quad \forall i \in \mathcal{S}. \tag{2}$$

It is clear that the fewer nodes one needs to compress the observed labels, the *easier* the learning task. This argument can be made precise through a compression argument as in [9, 6, 13] and refined in [7]. The integer program (2) can be relaxed as classically using a linear program by relaxing $q \in \{-1, 0, 1\}^n$ to $q \in [-1, 1]^n$, and by using the L_1 norm as a convex approach to the original L_0 norm:

$$\hat{q} = \underset{q \in [-1,1]^n}{\arg\min} \; \mathcal{J}^1(q) = \sum_{i=1}^{n} |q_i| \quad \text{s.t.} \quad y_i \left(\sum_{j=1}^{n} x_{ij} q_j\right) > 0, \quad \forall i \in \mathcal{S}. \tag{3}$$

Using this result as a initial value for a local search for integer solutions $q \in \{-1, 0, 1\}^n$ usually gives satisfying results.

3 A Robust Approach

At first we discuss a robust worst case approach. Let the link weights x_{ij} be estimated as $\hat{X}_{ij}$ having high probability that $|x_{ij} - \hat{X}_{ij}| \leq \sigma_{ij}$ for all $i < j = 1, \ldots, n$. The worst

case approach guarantees that the labeling is indeed compressed into a minimum number of support nodes, even if the weights are counteracting at worst. This problem can be solved as

$$\hat{q} = \underset{q\in\{-1,0,1\}^n}{\arg\min}\ \mathcal{J}_\sigma(q) = \sum_{i=1}^{n} I(q_i \neq 0) \quad \text{s.t.} \quad y_i\left(x_i^T q\right) > \sum_{j=1}^{n} |\sigma_{ij}|, \quad \forall i \in \mathcal{S}, \tag{4}$$

where $\sum_{j=1}^{n} |\sigma_{ij}|$ characterizes the worst case uncertainty. As in (3), a LP relaxation can be formulated. Secondly, we study an approach based on robust programming. This formulation relies on a parametric assumption on the weight sample distribution as follows

$$\hat{q} = \underset{q\in\{-1,0,1\}^n}{\arg\min}\ \mathcal{J}_{\gamma,\alpha}(q) = \sum_{i=1}^{n} |q_i| \quad \text{s.t.} \quad P\left(y_i\left(X_i^T q\right) > \gamma\right) < \alpha, \quad \forall i \in \mathcal{S}. \tag{5}$$

In case one assumes that X_i is sampled from a Gaussian distribution, it can be shown following [8, 4] that this problem can be cast as a second order cone programming problem (SOCP).

4 Information Retrieval Example

As an example we explore a particular problem in information processing. Given a set of nodes $\{v_1, \ldots, v_n\}$, representing different books, and provided a user expresses its taste by labeling a few $\mathcal{S} \in \{1, \ldots, n\}$ of them to be either interesting ($y_i = +1$) or not ($y_i = -1$). The task now is to predict the users appreciation of the remaining books. This gives implicitly answers to the question *'Customers who bought this item also bought...'*. This can be plainly cast as a transductive inference task over a graph, except for the additional work on estimating the graph weights. The latter task is even more complicated if only a little amount of resources may be used to infer the link weights from the different books.

A common principled assumption is that similar books in terms of literary content imply a similar users appreciation. To formalize this, the cosines rule is often used with a bag of word approach (see e.g. [1]): say $\mathcal{W}$ represent the word vocabularium, such that one can compute the link weights as

$$x_{ij} = x(v_i, v_j) = \frac{I_i^T I_j}{\|I_i\|_2 \|I_j\|_2}, \tag{6}$$

where $I_i \in \{0,1\}^{|\mathcal{W}|}$ denotes the binary vector with the kth element $I_{i,k}$ denoting whether the kth word of the vocabularium occurs in book v_i for any $i = 1, \ldots, n$. Now it can be argued that word-counts only do not capture the full correspondences between the books v_i and v_j. Therefor, we consider the universe of all consecutive sequences of words in the vocabularium (i.e. sentences) $\mathcal{W}$, yielding a space $\mathcal{W}^*$. Assume a fixed distribution $F_{\mathcal{W}^*}$ over all sequences $s \in \mathcal{W}^*$. The cosines rule can as such be written as

$$x^*(v_i, v_j) = \frac{\int I(s \in v_i) I(s \in v_j) dF_{\mathcal{W}^*}(s)}{\sqrt{\int I(s \in v_i) dF_{\mathcal{W}^*}(s)} \sqrt{\int I(s \in v_j) dF_{\mathcal{W}^*}(s)}}, \tag{7}$$

where the indicator $I(s \in v_i) \in \{0,1\}$ denotes whether the sentence s occurs in book v_i. As it is clear that this rule is not affected by any word sequences occurring in either v_i nor v_j, a simple (but computational heavy) approach is to check any sentence in v_i, and checking whether it occurs in v_j, and vice versa. For the sake of argument, take $F_{\mathcal{W}^*}$ a uniform distribution over all sequences s consisting of a few different words. Unfortunately, it is clearly infeasible to compute all the distances $x^*(v_i, v_j)$ in a reasonable time.

The main argument now is to approximate the distance $x^*(v_i, v_j)$ by random sampling sequences s of length three from either v_i and v_j, and to check whether they occur in the

latter book. By repeating this operation a few times (Monte-Carlo approximation), one get an estimate of the uncertainty of the location estimate, which can be fed into the robust approach (4) or (5). This approach shortcuts the dire necessity for computing the rule (7) exactly. Some small-scale experiments are used to argument the advantage of this approach over a conventional transductive inference approach.

5 Conclusion

This presentation elaborates on two novel issues, namely the handling of empirical graphs, and the use of optimal condensing as a regularization method. Both a worst-case and a robust programming approach are discussed for implementing the resulting task. A specific problem in information retrieval and user modelling is described for illustrating ideas.

Acknowledgments. Research supported by BOF PDM/05/161, FWO grant V 4.090.05N, IPSI Fraunhofer FgS, Darmstadt, Germany. (Research Council KUL): GOA AMBioRICS, CoE EF/05/006 Optimization in Engineering, several PhD/postdoc & fellow grants; (Flemish Government): (FWO): PhD/postdoc grants, projects, 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, G.0302.07. research communities (ICCoS, ANMMM, MLDM); (IWT): PhD Grants,GBOU (McKnow), Eureka-Flite2 - Belgian Federal Science Policy Office: IUAP P5/22,PODO-II,- EU: FP5-Quprodis; ERNSI; - Contract Research/agreements: ISMC/IPCOS, Data4s, TML, Elia, LMS, Mastercard. JS is a professor and BDM is a full professor at K.U.Leuven Belgium.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[2] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, pages 19–26. Morgan Kaufmann Publishers, 2001.

[3] A. Blum, J. Lafferty, M.R. Rwebangaria, and R. Reddy. Semi-supervised learning using randomized mincuts. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*. Morgan Kaufmann Publishers, 2004.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.

[6] S. Floyd and M.K. Warmuth. Sample compression, learnability and the VC dimension. *Machine Learning*, 21(3):269–304, 1995.

[7] T. Graepel, R. Herbrich, and J. Shawe-Taylor. PAC-bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55 – 76, 2005.

[8] M.S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret. Applications of second order programming. *Linear Algebra and its Applications*, 284:193–228, 1998.

[9] Littlestone N. and Warmuth M. Relating data compression and learnability. *Technical Report University of California, Santa-Cruz*, 1986.

[10] R. El-Yaniv P. Derbeko and R. Meir. Explicit learning curves for transduction and application to clustering and compression algorithms. *Journal of Artificial Intelligence Research*, 22:117–142, 2004.

[11] K. Pelckmans, J. Shawe-Taylor, J.A.K. Suykens, and B. De Moor. Margin based transductive graph cuts using linear programming. *Internal Report 06-164, ESAT-SISTA, K.U.Leuven (Leuven, Belgium), submitted for publication*, 2006.

[12] V.N. Vapnik. *Statistical Learning Theory*. Wiley and Sons, 1998.

[13] U. von Luxburg, O. Bousquet, and B. Schölkopf. A compression approach to support vector model selection. *Journal of Machine Learning Research*, 5:293–323, 2004.