Transductive Learning over Graphs: Incremental Assessment

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

Graphs constitute a most natural way to represent problems involving finite or countable universes. This might be especially so in the context of bio-informatics (e.g. for *protein-interaction graphs*), collaborative filtering, the analysis of social networks and citation graphs, and to various problems in operations research in the context of incomplete information. A further argument for using graphs for characterizing learning problems was found in the connection it makes to the literature on network flow algorithms and other deep results of combinatorial optimization problems.

This short note reviews results obtained in [3], and extends results slightly towards an incremental setting by exploiting a subresult of [4]. The relevance for machine learning of this result can be seen e.g. in a context of bio-informatics. Assume one has 1000 genes organized in an observed graph. The results of this paper give probabilistic guarantees on hypotheses which are proposed *during* the course of gathering more label-information of the nodes. Suppose e.g. one performs experiments to inspect wether a gene is cancer-related or not. The result below quantifies the increase of confidence in the optimal hypotheses of the set of cancer-related genes at all times.

Transductive Learning on Weighted Graphs

Some notation is introduced. Let a weighted undirected graph $\mathcal{G}_n = (V, E)$ consist of $1 < n < \infty$ nodes $V = \{v_i\}_{i=1}^n$ with edges $E = \{w_{ij} \geq 0\}_{i \neq j}$ with w_{ij} connected to v_i and v_j for any $i \neq j = 1, \ldots, n$. Assume that no loops occur in the graph, i.e. $w_{ii} = 0$ for all $i = 1, \ldots, n$, and that the graph $\mathcal{G}$ is connected, i.e. there exists a path between any two nodes. This paper considers problems where each node has a fixed corresponding label $y_i \in \{-1, 1\}$ such that $\{(v_i, y_i)\}_{i=1}^n$, but only an index-subset $\mathcal{S}_m \subset \{1, \ldots, n\}$ with $|\mathcal{S}_m| = m$ of the labels is observed. The task in *transductive learning* is to predict the labels of the unlabeled nodes $\mathcal{S}_{-m} = \{1, \ldots, n\} \backslash \mathcal{S}_m$. This paper uses the notation $q \in \{-1, 1\}^n$ to denote a hypothesis $\{(v_i, q_i)\}_{i=1}^n$ of the true labeling $\{(v_i, y_i)\}_{i=1}^n$.

This research track is boosted by results [2] on transductive learning, and by e.g. [1] on graph cuts for learning (see [3] for a more complete literature review). Results are further complemented in the contribution [3] with the following results. The analysis there starts off by fixing a weighted neighborhood-rule $r_q : V \to \{-1, 1\}$ as

$$r_q(v_i) = \text{sign}\left(\sum_{j=1}^n q_j w_{ij}\right). \tag{1}$$

A specific hypothesis $q \in \{-1, 1\}^n$ is *plausible* if it is consistent with itself, i.e. $q_i \; r_q(v_i) = 1$ for all $i = 1 \ldots, n$. Let $r_q^n = (r_q(v_1), \ldots, r_q(v_n))^T$. The corresponding hypothesis space is defined for fixed $\rho \geq 0$ as

$$\mathcal{H}_\rho = \left\{ q \in \{-1, 1\}^n \;\middle|\; g\left(q, r_q^n\right) \geq \rho \right\}, \tag{2}$$

with $g : \{-1, 1\}^n \times \{-1, 1\}^n \to \mathbb{R}^+$ a function quantifying the *plausibility* of the hypothesis. Main contributions of [3] are **(i)** an explicit form of g in terms of the margin and average margin induced by the rule (1), and the relationship to the graph cut; **(ii)** an explicit characterization of this hypothesis space in terms of the eigenvalue spectrum of the graph Laplacian; **(iii)** an extension to the case where only positive samples are observed; and **(iv)** the proposal of an efficient relaxation of the corresponding problem in terms of a linear program. Recent results show further relations to network flow problems and graph cut algorithms.

Incremental Assessment for Transductive Learning

This section extends standard results to the incremental case where the graph $\mathcal{G}$ is completely known, and where an independent process (*nature*) gradually presents new label information for the task. Let the sequence $\Pi = \left(v_{\pi(1)}, \ldots, v_{\pi(n)}\right)$ which is followed in the process be a random permutation. With some slightly notational abuse, let $v_t = v_{\pi(t)}$ for all $t = 1, \ldots, m$ (t

indexes the nodes in the unknown but fixed sequence). The actual risk of a hypothesis $q \in \mathcal{H}_\rho$, and its empirical counterpart at timestep t is defined as

$$\mathcal{R}(q) = \frac{1}{n}\sum_{i=1}^{n} I(q_i y_i < 0), \quad \mathcal{R}_t(q) = \frac{1}{t}\sum_{i=1}^{t} I(q_i y_i < 0).$$

The incremental procedure goes for all $t = 2, \ldots, m$:

$$\begin{cases} 1. \text{ Estimate } q^{(t)} \in \mathcal{H}_\rho \text{ based on } \mathcal{G}, \{y_1, \ldots, y_{t-1}\} \\ 2. \text{ Nature asks for } \textit{randomly chosen} \text{ node } v_j \in V \\ 3. \text{ The algorithm presents } q_j^{(t)} \text{ with confidence } \mathcal{R}_t(q^{(t)}) \\ 4. \text{ A new experiment reveals } y_t \in \{-1, 1\}. \end{cases}$$

Remark that one can do better when $j < t$ by returning y_j, but as this occurs not too often if $m \ll n$, we proceed as such for the moment being. Now one can analyze how well the estimate of the risk correspond with the actual risk. we use a result by Serfling [4] to give a generalization bound in each stage of the incremental process, which is surprisingly as tight as in the batch case. The first result states that the difference of the estimated risk of a fixed hypothesis q will converge to the true risk during the incremental process where one receives gradually new labels.

Theorem 1 (Incremental Serfling Bound) *Let $\mathcal{G}$ be fixed and observed, and let $q \in \{-1, 1\}^n$ be a fixed hypothesis. The risk $\mathcal{R}(q)$ is defined as before, but the empirical counterparts now become $\{\mathcal{R}_t(q)\}_{1 \le t \le m}$. Let $C = \frac{m}{n-m}$. With probability $1 - \delta < 1$, the following inequality holds for all $1 \le t \le m$:*

$$\mathcal{R}(q) \le \mathcal{R}_t(q) + \left(\frac{n-t}{t}\right) C \sqrt{\frac{2(n-m+1)}{nm} \log\left(\frac{1}{\delta}\right)}.$$

Proof: This results immediately from a sub-result in Serfling's seminal paper [4], Corollary 1.1 and its proof. Specifically, the martingale strategy used to proof Serfling's inequality uses the quantity

$$U_n(\epsilon; q) = P\left(\max_{1 \le t \le m} \frac{t\mathcal{R}_t(q) - t\mathcal{R}(q)}{n-t} \le \left(\frac{m}{n-m}\right)\epsilon\right), \tag{3}$$

which is proven to be smaller than $\exp\left(-\frac{1}{2} m\epsilon^2 \frac{n}{n-m+1}\right)$. By reshuffling variables n, m, t in (3), the following inequality follows $P\left(\max_{1 \le t \le m} \mathcal{R}(q) \ge \mathcal{R}_t(q) + \epsilon C \frac{n-t}{t}\right) \le e^{-\left(\frac{n}{2(n-m+1)}\ m\epsilon^2\right)}$,with $C = \frac{m}{n-m}$. Inverting the statement proves the result. □

This result is especially convenient as it states a result on a set of tests $\{\mathcal{R}(q) - \mathcal{R}_t(q)\}_{1 \le t \le m}$ without having to resort to an (often pessimistic) union bound technique. It states that in an incremental scenario, the uncertainty decreases as $\mathcal{O}\left(\frac{n-t}{t}\right)$. Taking the limit $\lim_{n\to\infty} \frac{2(n-m+1)}{n} = 2$ and $\left(\frac{n-t}{t}\right)\left(\frac{m}{n-m}\right) \le \frac{m}{t}$ one gets an expression for a graph with an infinite number of nodes. The following practical expression is immediate.

Corollary 1 (Incremental PAC Bound) *With probability $0 < 1 - \delta < 1$, the following inequality holds for all $1 \le t \le m$ and for any $q \in \mathcal{H}_\rho$*

$$\mathcal{R}(q) \le \mathcal{R}_t(q) + \left(\frac{n-t}{t}\right) C \cdot \sqrt{\frac{2(n-m+1)}{nm}\left(\log(|\mathcal{H}_\rho|) + \log\left(\frac{1}{\delta}\right)\right)}.$$

This result follows as one can switch '$\max_{1 \le t \le m} \sup_{q \in \mathcal{H}_\rho}$' to '$\max_{q \in \mathcal{H}_\rho} \max_{1 \le t \le m}$', as both domains $1 \le t \le m$ and $\mathcal{H}_\rho$ are finite.

It becomes clear that those results open up new possibilities for research in the context of transductive learning. In particular, it can be expected to help in bridging the gap between the analysis of (deterministic) mistake bounds (e.g. for the perceptron and weighted majority rule) and the stochastic setting of empirical risk minimization. A second interesting implication can be found in the analysis of experimental designs.

Acknowledgments. Research supported by GOA AMBioRICS, CoE EF/05/006; (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. (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. This publication only reflects the authors' views.

References

[1] 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.

[2] 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.

[3] K. Pelckmans, J. Shawe-Taylor, J.A.K. Suykens, and B. De Moor. Margin based transductive graph cuts using linear programming. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, San Juan, Puerto Rico, 2007.

[4] R.J. Serfling. Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, 1:39–48, 1974.