Université de Montréal

Efficient Non-Parametric Function Induction in Semi-Supervised Learning

Olivier Delalleau
Yoshua Bengio
Nicolas Le Roux

Summary

- In the semi-supervised setting, we combine labeled and unlabeled data.
- With current non-parametric approaches, it is often unclear how to find labels for previously unseen examples without retraining the whole model (which typically requires $O(n^3)$ time, where n is the number of training points).
- We propose and justify a method to cheaply ($O(n)$ time) **perform function induction** in this context.
- This approach leads to **efficient approximations** of the original training algorithm, by writing all predictions in terms of a small subset of $m \ll n$ samples ($\Rightarrow O\left(\frac{n^2}{m^2}\right)$ faster)

An Optimization Framework for Semi-Supervised Learning

Several previously proposed methods can be cast into a transduction framework, where we learn a function $f(x)$ giving a *continuous* label on each point, such that:
(i) f is smooth (two neighbor samples are given similar labels)
(ii) f is coherent with already known labels
i.e. f minimizes

$$C_{K,D,D',\lambda}(f) = \frac{1}{2} \sum_{x_i,x_j \in U \cup L} K(x_i,x_j) D(f(x_i), f(x_j)) + \lambda \sum_{x_i \in L} D'(f(x_i), y_i) + R(f) \quad (1)$$

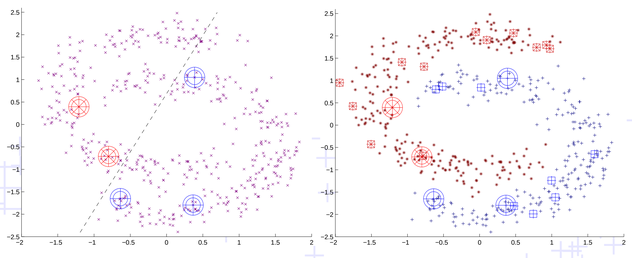
with
- U the unlabeled set
- L the labeled set
- x_i the input part of the i-th example
- y_i its target label
- $K(\cdot,\cdot)$ a similarity function (e.g. a Gaussian kernel)
- $D(\cdot,\cdot)$ and $D'(\cdot,\cdot)$ dissimilarity functions (typically, the Euclidean distance)
- $R(f)$ an optional additional regularization term.

See e.g. (Zhu, Ghahramani and Lafferty, 2003; Zhou et al., 2004; Belkin, Matveeva and Niyogi, 2004).

Induction: Extending to New Points

- Consider D, D' and R quadratic: minimizing (1) with respect to the $f(x_i)$ reduces to solving a linear system of size n.
- Given a new test point x, re-training will in general cost $O(n^3)$ time.
- Minimizing (1) with $f(x_i)$ ($i \leq n$) fixed (with D Euclidean and $R(f) = 0$) $\Rightarrow$ Parzen windows regressor (induction in $O(n)$ time)

$$f(x) = \frac{\sum_{j=1}^{n} K(x,x_j) f(x_j)}{\sum_{j=1}^{n} K(x,x_j)} \quad (2)$$



Left: a classifier trained only with the 5 labeled samples (in circles) completely overlooks the underlying structure of the data.
Right: classification learned from (1) on training data ($*$ and $+$), and tested with (2) (points in squares). Classification error on 10000 test points is 1.76%.

Efficient Approximation for Training

- Equation (2) suggests we can choose a subset S of $m \ll n$ samples and force $f(x_i)$ for $x_i \notin S$ to be expressed as a linear combination of the $f(x_j)$ with $x_j \in S$ as in (2).
- Minimizing (1) then reduces to solving a linear system with only m unknowns. However, to obtain this linear system, we still need to perform $O(m(n-m)^2)$ operations.
- To further improve the performance, we can choose to ignore in the total cost C the discarded points cross-terms $\frac{1}{2}\sum_{x_i,x_j \notin S} K(x_i,x_j) D(f(x_i), f(x_j))$, the most expensive to compute. Then *we only need $O(m^2(n-m))$ time and $O(m^2)$ memory, versus $O(n^3)$ and $O(n^2)$ for the original algorithm.*
- Smart selection of the subset S gives better results than random selection: we propose to greedily build S by iteratively choosing the point farther from S, i.e. x_i that minimizes $\sum_{j \in S} K(x_i,x_j)$. Additionally, a preliminary (fast) training is performed using only S in order to add more points near the decision surface.

Experiments

- Comparison between *Laplacian* (Belkin and Niyogi, 2003), *WholeSet* in transduction and *WholeSet* in induction on the MNIST Database.

Labeled	50	100	500	1000	5000
Total: 1000					
Laplacian	29.3	19.6	11.5		
$WholeSet_{trans}$	25.4	17.3	9.5		
$WholeSet_{ind}$	26.3	18.8	11.3		
Total: 10000					
Laplacian	25.5	10.7	6.2	5.7	4.2
$WholeSet_{trans}$	25.1	11.3	5.3	5.2	3.5
$WholeSet_{ind}$	25.1	11.3	5.7	5.1	4.2

- Comparison (induction) between
  - *WholeSet*: uses all unlabeled data (no approximation)
  - $RSub_{subOnly}$: uses only a random subset of the unlabeled data (no approximation)
  - $RSub_{noRR}$: uses a random subset of the unlabeled data to approximate for all data
  - $SSub_{noRR}$: uses a selected subset of the unlabeled data to approximate for all data

% labeled	LETTERS	MNIST	COVTYPE
1%			
WholeSet	56.0 ± 0.4	35.8 ± 1.0	47.3 ± 1.1
$RSub_{subOnly}$	59.8 ± 0.3	29.6 ± 0.4	$\mathbf{44.8} \pm 0.4$
$RSub_{noRR}$	57.4 ± 0.4	27.7 ± 0.6	75.7 ± 2.5
$SSub_{noRR}$	$\mathbf{55.8} \pm 0.3$	$\mathbf{24.4} \pm 0.3$	45.0 ± 0.4
5%			
WholeSet	$\mathbf{27.1} \pm 0.4$	12.8 ± 0.2	37.1 ± 0.2
$RSub_{subOnly}$	32.1 ± 0.2	14.9 ± 0.1	$\mathbf{35.4} \pm 0.2$
$RSub_{noRR}$	29.1 ± 0.2	12.6 ± 0.1	70.6 ± 3.2
$SSub_{noRR}$	28.5 ± 0.2	$\mathbf{12.3} \pm 0.1$	35.8 ± 0.2
10%			
WholeSet	$\mathbf{18.8} \pm 0.3$	$\mathbf{9.5} \pm 0.1$	34.7 ± 0.1
$RSub_{subOnly}$	22.5 ± 0.1	11.4 ± 0.1	$\mathbf{32.4} \pm 0.1$
$RSub_{noRR}$	20.3 ± 0.1	9.7 ± 0.1	64.7 ± 3.6
$SSub_{noRR}$	19.8 ± 0.1	$\mathbf{9.5} \pm 0.1$	33.4 ± 0.1

- Comparison between $RSub_{noRR}$ and $SSub_{noRR}$: 10% of the data is labeled, and we use 10% of unlabeled data as a subset for approximation (5% for ADULT).

	USPS	IMAGE	ISOLET	SATIMAGE	NURSERY	PENDIGITS	ADULT	SPAMBASE
$RSub_{noRR}$	9.8	17.0	24.8	13.9	18.6	19.6	19.3	30.5
$SSub_{noRR}$	**8.6**	**16.1**	**22.9**	13.8	18.4	**17.1**	**18.6**	**28.3**

- **CONCLUSION**
  - **fast induction with little loss w.r.t. transduction**
  - **fast training when choosing a subset of unlabeled data to approximate the cost**
  - **smart subset selection > random selection**

References

Belkin, M., Matveeva, I., and Niyogi, P. (2004). Regularization and semi-supervised learning on large graphs. In Shawe-Taylor, J. and Singer, Y., editors, *COLT'2004*. Springer.

Belkin, M. and Niyogi, P. (2003). Using manifold structure for partially labeled classification. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, Cambridge, MA. MIT Press.

Zhou, D., Bousquet, O., Navin Lal, T., Weston, J., and Schölkopf, B. (2004). Learning with local and global consistency. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16*, Cambridge, MA. MIT Press.

Zhu, X., Ghahramani, Z., and Lafferty, J. (2003). Semi-supervised learning using gaussian fields and harmonic functions. In *ICML'2003*.