A Preliminary Study on Transductive Extreme Learning Machines

Scardapane S., Comminiello D., Scarpiniti M. and Uncini, A.

23rd Italian Workshop on Neural Networks (Vietri sul Mare)

intelligent signal processing and multimedia lab

SAPIENZA UNIVERSITÀ DI ROMA

May 23/25 2013

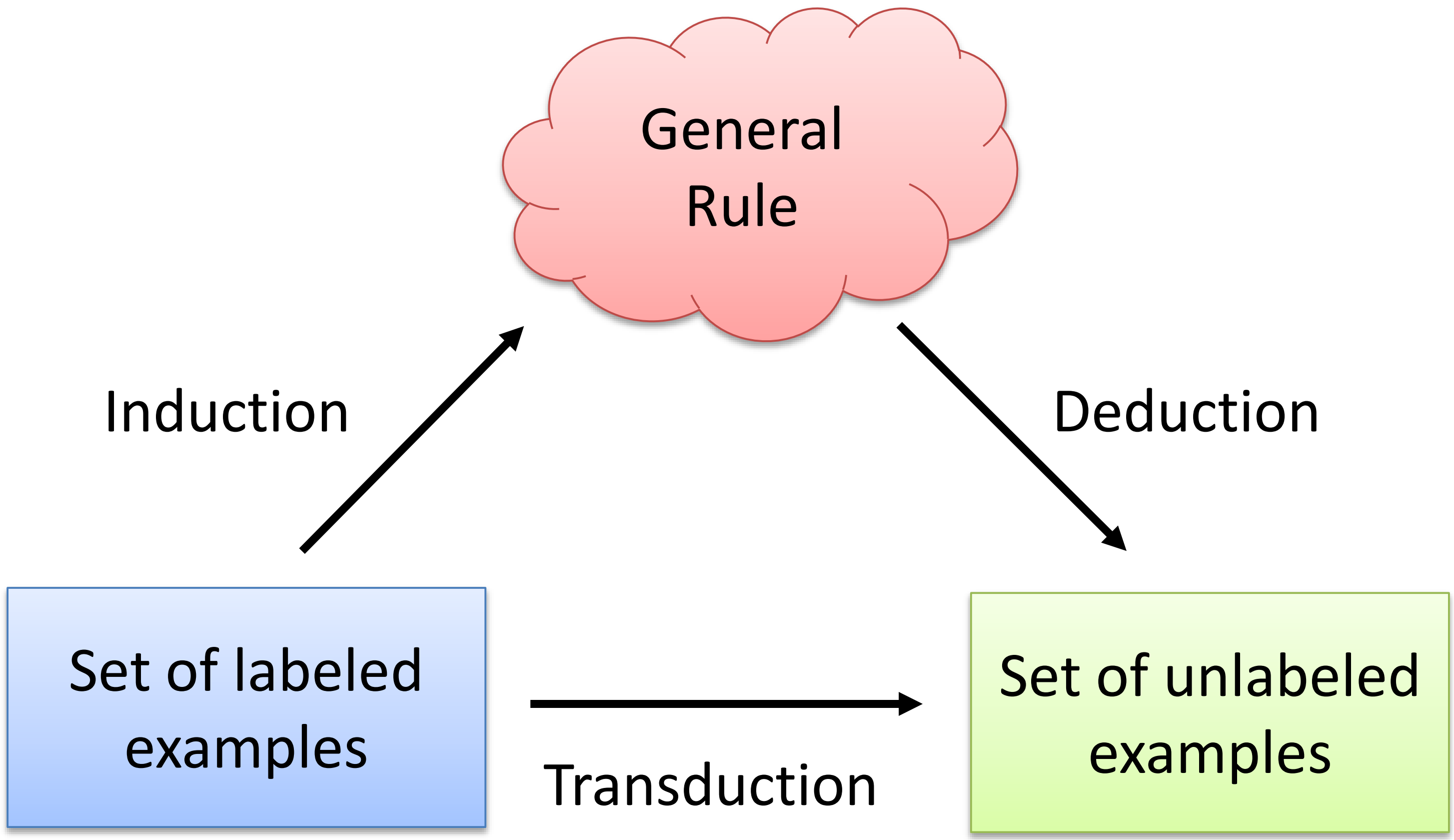
Contents

The Transductive Learning Problem

Transductive Extreme Learning Machines

Preliminary Results

Induction vs. Transduction



The Transductive Problem

- Vapnik argued that general induction may be unnecessary in some cases.

> *"[...] when solving a problem of interest, do not solve a more general problem as an intermediate step."*

- The knowledge of the actual testing points should improve the capabilities of the inference system.

- In the Transductive setting, the output is not a model but a set of predictions.

A Theoretical Perspective

- To appreciate the theoretical difference, consider the set of possible hypotheses H.
- In the transductive case, H is necessarily *finite*.
- This leads to an extension of inductive statistical learning theory, resulting in the following "advice":

> *Minimize the error on both training and testing set while maximizing the margin.*

Notation

- The **training set** is $S = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N}$, and we restrict to the binary classification case $y_i = \{0,1\}$.
- The **testing set** is $U = \{\boldsymbol{x}_i\}_{i=N+1}^{N+M}$. A possible labelling is $\boldsymbol{y}^* = [y_{N+1} \dots y_{N+M}]^T$.
- Minimization is done over a generic *Reproducing Kernel Hilbert Space* $\mathcal{H}$ with norm $\|\cdot\|_{\mathcal{H}}$.
- $k(\cdot,\cdot)$ is the kernel associated to $\mathcal{H}$.

Transductive SVM

$$\min_{f,\boldsymbol{y}^*} \underbrace{\frac{1}{2}\|f\|_{\mathcal{H}}^2 + C_S \sum_{i=1}^{N} \zeta_i}_{\text{Standard Terms}} + \underbrace{C_U \sum_{i=N+1}^{N+M} \zeta_i}_{\text{Transductive Term}}$$

$$s.t. \quad \begin{array}{lll} y_i f(\boldsymbol{x}_i) \geq 1 - \zeta_i, & \zeta_i \geq 0, & i = 1, \dots, N \\ y_i^* f(\boldsymbol{x}_i) \geq 1 - \zeta_i, & \zeta_i \geq 0, & i = N+1, \dots, N+M \end{array}$$

- ζ_i are *slack variables* controlling the error.
- C_s and C_U are regularization parameters.

T-SVM Learning

- T-SVM training results in a partly combinatorial problem, due to the presence of the unknown labels.
- Several algorithms have been devised for its efficient solution, depending on the simplifications that are made.
- A good discussion can be found in:

[1] O. Chapelle, V. Sindhwani, and S. Keerthi, “Optimization techniques for semi-supervised support vector machines,” *Journal of Machine Learning Research*, vol. 9, pp. 203–233, 2008.

Extreme Learning Machine

- An **Extreme Learning Machine** (ELM) is a model of the form:

$$f(\boldsymbol{x}) = \sum_{i=1}^{L} h_i(\boldsymbol{x})\beta_i = \boldsymbol{h}(\boldsymbol{x})^T\boldsymbol{\beta}$$

- The hidden layer $\boldsymbol{h}(\boldsymbol{x})^T$ is fixed before observing the data.

- Typically, it is constructed by randomizing over a known function $g(\boldsymbol{x}, \boldsymbol{\theta})$.

ELM Training

- The weights are found by a L_2-regularized linear regression:

$$\min_{\boldsymbol{\beta}} \ \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \frac{C_S}{2}\sum_{i=1}^{N}\zeta_i^2$$

$$s.t. \qquad \boldsymbol{h}(\boldsymbol{x}_i)^T\boldsymbol{\beta} \geq y_i - \zeta_i, \ \ \zeta_i \geq 0, \qquad i = 1, \dots, N$$

- A possible solution is given by:

$$\boldsymbol{\beta} = \boldsymbol{H}^T\left(\frac{1}{C_S}\boldsymbol{I}_N + \boldsymbol{H}\boldsymbol{H}^T\right)^{-1}\boldsymbol{y}$$

Transductive ELM

- We propose the following transductive model:

$$\min_{\boldsymbol{\beta},\boldsymbol{y}^*} \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \frac{C_S}{2}\sum_{i=1}^{N}\zeta_i^2 + \frac{C_U}{2}\sum_{i=N+1}^{N+M}\zeta_i^2$$

$$\begin{aligned} s.t. \quad & \boldsymbol{h}(\boldsymbol{x}_i)^T\boldsymbol{\beta} \geq y_i - \zeta_i, \;\; \zeta_i \geq 0, \qquad i = 1, \dots, N \\ & \boldsymbol{h}(\boldsymbol{x}_i)^T\boldsymbol{\beta} \geq y_i^* - \zeta_i, \;\; \zeta_i \geq 0, \qquad i = N+1, \dots, M \end{aligned}$$

- Back-substituting the solution for $\boldsymbol{\beta}$ we obtain a *fully combinatorial* problem.

T-ELM Training

- A simplification to the minimization problem can be derived:

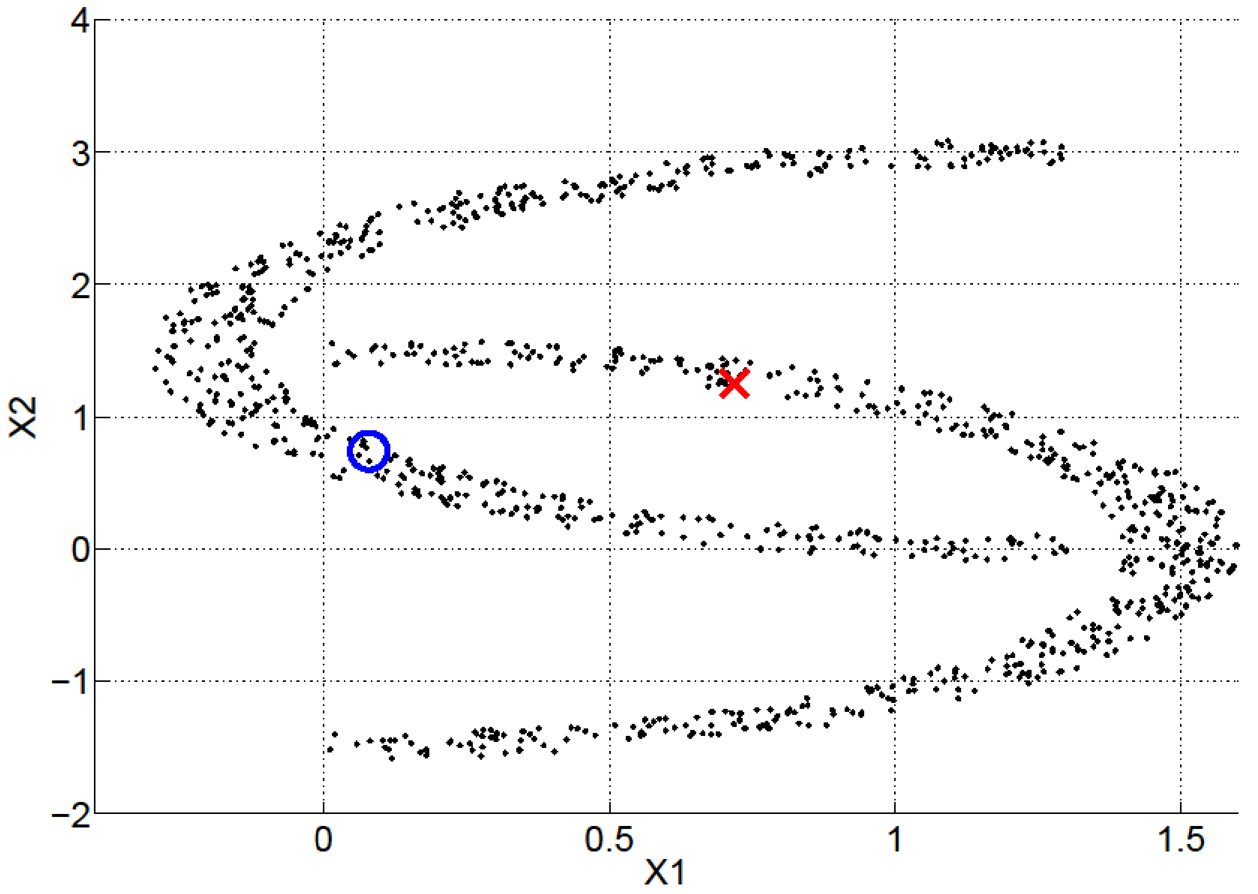
$$\boldsymbol{\beta} = \boldsymbol{H}^T(\boldsymbol{C}^{-1}\boldsymbol{I} + \boldsymbol{H}\boldsymbol{H}^T)^{-1}\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{y}^*\end{bmatrix}$$

$$\widehat{\boldsymbol{H}} = \boldsymbol{H}^T(\boldsymbol{C}^{-1}\boldsymbol{I} + \boldsymbol{H}\boldsymbol{H}^T)^{-1} = [\widehat{\boldsymbol{H}_1} \;\; \widehat{\boldsymbol{H}_2}]$$

$$\downarrow$$

$$\boldsymbol{\beta} = \widehat{\boldsymbol{H}_1}\boldsymbol{y} + \widehat{\boldsymbol{H}_2}\boldsymbol{y}^*$$

Two Moons Dataset



Setting

- ELM and T-ELM feature vectors are found by randomizing over:

$$g(\boldsymbol{x}, \boldsymbol{a}, b) = \frac{1}{1 + \exp\{-(\boldsymbol{a}^{\boldsymbol{T}} \boldsymbol{x} + b)\}}$$

- T-ELM is solved with a standard Genetic Algorithm.

- Parameters are found by cross-validating over an independent validation set.

Results

N	M	SVM	ELM	T-ELM
100	200	**0,89**	0,81	0,843
2	200	0,47	0,5	**0,58**

- The T-ELM model does not improve in the normal situation.
- However, it gives a substantial improvement in the harder situation.
- We hypothesize the first situation is due to poor performance of the GA.

Open Problems

1. Our formulation is not trivially extended to regression. See for example:

   [1] C. Cortes and M. Mohri, "On transductive regression," *Advances in Neural Information Processing Systems*, 2007.

2. There is the need of a specialized solver for the minimization problem.

Conclusions

1. We proposed a transductive model which is simpler than T-SVM.

2. Some preliminary results showed good results with unbalanced datasets.

3. Further work is needed for a realistic implementation.

Thanks for your attention!

Any Questions?