1

Explicit overall risk minimization transductive bound

Sergio Decherchi, Paolo Gastaldo, Sandro Ridella, Rodolfo Zunino

Dept. of Biophysical and Electronic Engineering (DIBE), Genoa University
Via Opera Pia 11a, 16145 Genoa, Italy

Summary. Aside classical inductive methods transduction has reached an always increasing attention from the scientific community because of its learning paradigm. Explicit error bounds for inductive methods are well established results and stem from Vapnik theory or Rademacher complexity. In this work we address the problem of building an explicit form of the transductive bound presented in Vapnik Overall Risk Minimization approach.

1 Introduction

In recent years, approaches alternatives to full induction have reached an always increasing attention from the machine learning research community [1]. Inductive methods find a global solution from empirical data and build general model applicable all over the population. Beside inductive learning schemes, exist the so called transductive learning: in this environment is not required generalization for every possible input, instead only achieving the best possible performance on a particular and known test data. This, intuitively, makes transduction simpler than induction, since what is request are values at given points [2] and not a global predictive function.

Transduction and semi-supervised learning are quite different concepts: in the first setting we are interested in finding values at given points and no more, in semi-supervised learning we are interested in producing a decision function by using labelled and unlabelled data: a transductive algorithm can perform predictions only on working set, a semi-supervised one can predict all over the population so it is completely inductive.

The importance of transduction is due to different reasons: one of them is its fundamental part over the inductive approach itself. Well known Vapnik classical bound, implicitly makes use of transduction when concerned with ghost set. In transductive setting, the ghost set is real and is the set of given points in which predictions are performed. Another reason stems from the possibility to take advantage of this new simpler setting to get tighter bounds on generalization error over a particular working set. In this work this second aspect will be studied: adapting the machinery of Theorem 4.2 [2] and a relatively recent result [3] an explicit formula will be obtained for overall risk minimization bound.

In the first part of the paper computational issues will be discussed over the numerical evaluation of transductive bound in its implicit original form and a closed form formula will be obtained; in the second one the result will be compared to other existing bounds. The same symbolic conventions of [2] will be used throughout the paper:

- $l + k$ is the total number of patterns; l are labelled and k are unlabelled
- ν_τ is the transductive error (the error over test or working sample); ν is the error on training set; ν_2 is the error on the ghost set; than we call $\nu_0 = \nu \frac{l}{l+k} + \nu_\tau \frac{k}{l+k}$, and $\nu_\alpha = \frac{\nu+\nu_2}{2}$
- m is the total numnber of errors and can be epxressed as $\nu l + \nu_\tau k$
- $G(l+k)$ is the Grow Function computed for $l+k$; $H^{\Lambda}_{ann}(2l)$ is the annealed Entropy
- $1 - \delta$ is the confidence level of the bound
- C^r_m is the binomial coefficient
- $\Gamma_{l,k}(\varepsilon, m)$ is a quantity derived from the hypergeometric distribution; than we call $\Gamma_{l,k}(\varepsilon) = \max_m \Gamma_{l,k}(\varepsilon, m)$; $E(\Gamma)$ is the expectation of the hypergeometric ; N_{l+k} is the finite number of equivalence classes

2 Overall Risk Minimization Transduction

The Overall Risk Minimization framework is part of the more general Statistical Learning Theory and is one of the possible approaches to transduction. In [2] transduction is introduced and two possible settings (Setting 1 and Setting 2) are exposed: it can be shown that both of them are equivalent [2]. The fundamental idea, on which ORM is built up, is that is useless solving a more difficult problem when a simpler one is needed to be solved. From a mathematical point of view in ORM we are endowed with a training labeled set, an unlabelled working set on which we want to perform predictions and it is allowed to use both of them during training. This fundamental theorem gives an implicit bound for transduction error

Theorem 2.1. *(Theorem 8.2 in [2]) Let the set of decision rules $f(x, \alpha)$, $\alpha \in \Lambda$ on the complete set of vectors have N_{l+k} equivalence classes. Then the probability that the*

relative size of deviation for at least one rule in $f(x,\alpha)$, $\alpha \in \Lambda$ exceeds ε is bounded by:

$$P\left\{\sup_{\alpha\in\Lambda}\frac{|\nu-\nu_\tau|}{\sqrt{\nu_0}}>\varepsilon\right\}<N_{l+k}\Gamma(\varepsilon) \tag{1}$$

Now, as said in the introductory section, we want to link the equipment of induction to Theorem 8.2 [2]: in Statistical Learning Theory the main inductive result is Theorem 4.1 [2]; the key part in which we are interested in, is Lemma 4.2 [2].

Lemma 1. *(Lemma 4.2 in [2]) For any $l > \varepsilon^{-2}$ is valid the following bound:*

$$P\left\{\sup_{\alpha\in\Lambda}\frac{|\nu_2-\nu_1|}{\sqrt{\nu_\alpha+1/(2l)}}>\varepsilon\right\}<\exp\left\{H^{\Lambda}_{ann}(2l)-\frac{\varepsilon^2 l}{4}\right\} \tag{2}$$

further by using the property $H^{\Lambda}_{ann}(2l) < G(2l)$ for right hand side we get $\exp\left\{G(2l)-\frac{\varepsilon^2 l}{4}\right\}$. To make ORM approach consistent with machinery of Lemma 4.2 [2] we need to replace original Vapnik gamma function argument $\varepsilon\sqrt{\frac{m}{l+k}}$ with $\varepsilon\sqrt{\frac{m+1}{l+k}}$ (as in Lemma 4.2 [2] proof where we have $\varepsilon\sqrt{\frac{m+1}{2l}}$). This marginal modification leads also to modify (1) into

$$P\left\{\sup_{\alpha\in\Lambda}\frac{|\nu-\nu_\tau|}{\sqrt{\nu_0+\frac{1}{l+k}}}>\varepsilon\right\}<N_{l+k}\Gamma(\varepsilon) \tag{3}$$

and the final explicit formula becomes:

$$\nu_\tau \le \nu+\frac{k\varepsilon^2}{2(l+k)}+\varepsilon\sqrt{\left[\frac{k\varepsilon}{2(l+k)}\right]^2+\nu+\frac{1}{l+k}} \tag{4}$$

This modification makes Theorem 8.2 [2] consisten with Lemma 4.2 [2] at price of adding the term $1/(l+k)$ over the original formulation; as will be seen later this adaptation open the possibility to build up a plain proof that makes ε term explicit.

After this simple variation we tried to appraise the implicit bound derived in Theorem 8.2 [2]. For its evaluation one has to find the smallest solution of $\ln N_{l+k} + \ln\Gamma(\varepsilon) < \ln\delta$ and plug it into the bound; so one has to solve this equation for trials performing discretization on ε. Before proceeding it is necessary to explicitly compute the number of equivalence classes: for this purpose we used the fact that $\ln N_{l+k} < G(l+k)$. This approach presents some performance problems when the number of patterns (l or k) is over $1e3$. The main issue consists in the explicit evaluation of the gamma function: its calculation plans the evaluation of three binomial coefficients. We implemented this computation using Stirling and Ramanujan formulas and logarithmic representations, but despite this, the execution time is quite high when dealing with data mining problems. As suggested by Vapnik itself, gamma function can be tabulated but it should be preferable having a simpler and explicit way of computing the bound. Although these concerns, evaluation of the bound has been possible via iterative search of the solution. For the exposed reason a more practical solution consists in deriving an explicit bound.

3 Bound derivation

The subsequent theorem is the central result of this work: it follows Vapnik demonstration for the classical inductive bound (Theorem 4.2 [2], Lemma 4.2 [2]) and readapts it to the transductive issue (Theorem 8.2 [2]) using a quite recent statistical result.

Theorem 3.1. *(Explicit bound). Assured that* $G(l+k) - \ln\delta > 6$*, and setting* $\varepsilon^2 = \frac{G(l+k)-\ln\delta+2}{2(lk)^2}(l+k)^3$*, with probability* $1-\delta$*, the bound in (4) is valid.*

Proof. Suppose having a population of $l+k$ patterns in which there are m misclassified patterns. We select randomly l of them. The probability that among the selected patterns there are r errors equals $\frac{C_m^r C_{l+k-m}^{l-r}}{C_{l+k}^k}$. The probability that the frequency of misclassified patterns in the first group (l) deviates from the frequency of errors in the second group (k) by the amount exceeding $\bar{\varepsilon}$ equals:

$$P\left\{\left|\frac{r}{l} - \frac{m-r}{k}\right| > \bar{\varepsilon}\right\} = \sum_r \frac{C_m^r C_{l+k-m}^{l-r}}{C_{l+k}^k} = \Gamma_{l,k}(\bar{\varepsilon}, m) \tag{5}$$

Where the sum is taken over the value of r such that:

$$\max(0, m-k) \le r \le \min(l, m), \left|r - \frac{lm}{l+k}\right| > \bar{\varepsilon}\frac{lk}{l+k} \tag{6}$$

Note that both sides are always greater than 0. From [3] is known that: if

$$r - E(\Gamma) \ge 2 \tag{7}$$

is true, than:

$$\ln \Gamma_{l,k}(\bar{\varepsilon}, m) < -2\alpha((r - E(\Gamma))^2 - 1) \tag{8}$$

where $\alpha = \max\left(\frac{1}{l+1} + \frac{1}{k+1}, \frac{1}{m+1} + \frac{1}{l+k-m+1}\right)$. Knowing that $E(\Gamma) = \frac{ml}{l+k}$ we get:

$$\Gamma_{l,k}(\bar{\varepsilon}, m) < \exp\left(-2\alpha\left(r - \frac{ml}{l+k}\right)^2 - 1\right) \tag{9}$$

Expressing $\bar{\varepsilon} = \varepsilon\sqrt{\frac{m+1}{l+k}}$ we get: $\left|r - \frac{ml}{l+k}\right| > \varepsilon\frac{lk}{l+k}\sqrt{\frac{m+1}{l+k}}$. Note that beacause we need the square in (9), we can observe that by using (6) $\left(r - \frac{lm}{l+k}\right)^2 > \left(\bar{\varepsilon}\frac{lk}{l+k}\right)^2$ holds. For proceeding we have to assure that the hypothesis on hypergeometric bound (7) and (6) both hold. A simple way for achieving this goal is to request that: $\left(r - \frac{lm}{l+k}\right)^2 > \max\left(\left(\bar{\varepsilon}\frac{lk}{l+k}\right)^2, 2^2\right)$. Now observe that asking $\bar{\varepsilon}\frac{lk}{l+k} > 2$ is a sufficient condition to resort to the only $\left(r - \frac{lm}{l+k}\right)^2 > \left(\bar{\varepsilon}\frac{lk}{l+k}\right)^2$ original condition; in this way, at the end of the proof, we will have to check for what values the expression $\bar{\varepsilon}\frac{lk}{l+k} > 2$ is true. With these hypothesis we can bound the hypergeometric on (3) getting: $P\left\{\sup_{\alpha\in\Lambda}\frac{|\nu-\nu_\tau|}{\sqrt{\nu_0+\frac{1}{l+k}}} > \varepsilon\right\} < N_{l+k}\max_m\left\{\exp\left\{-2\alpha\left(\left(r - \frac{ml}{l+k}\right)^2 - 1\right)\right\}\right\}$ than we get:

$$P\left\{\sup_{\alpha\in\Lambda}\frac{|\nu-\nu_\tau|}{\sqrt{\nu_0+\frac{1}{l+k}}}>\varepsilon\right\}<N_{l+k}\max_m\left\{\exp\left\{-2\alpha\left(\left(\varepsilon\frac{lk}{l+k}\sqrt{\frac{m+1}{l+k}}\right)^2-1\right)\right\}\right\} \tag{10}$$

It can be easily shown, e.g. by plotting the function for different m, l, k values, (see fig.1) that hypergeometric dependent part of the previous formula is maximized for $m=0$ (as happens in Vapnik inductive proof). This fact ($m=0$) makes $\alpha=\max\left(\frac{1}{l+1}+\frac{1}{k+1},1+\frac{1}{l+k+1}\right)$, that is the same that saying that $\alpha>1$; for this reason we can replace α with 1. Observe that this operation on α slightly affects the quality of the bound, in facts in almost all real world problems (e.g $l,k>10$) $\alpha\simeq 1$ holds.

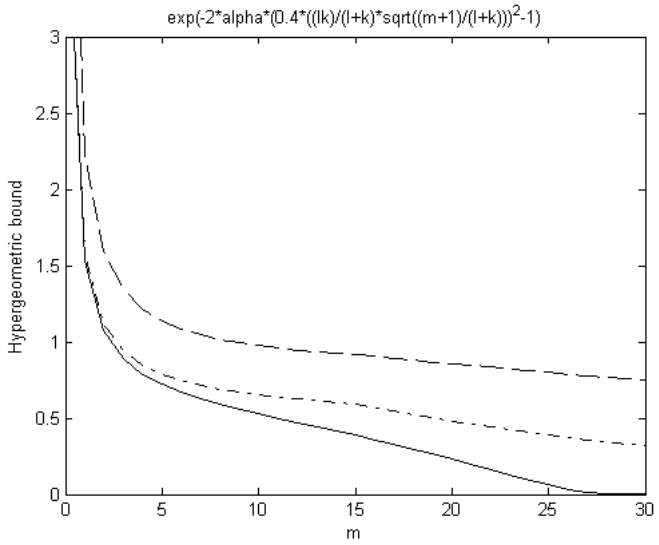

Fig. 1 Exponential part of the bound for different values of l and k with m variable and $\varepsilon=0.4$. Solid line represents the case $k=l$.

Setting $m=0$ and resembling that $\ln N_{l+k}<G(l+k)$ we obtain:

$$P\left\{\sup_{\alpha\in\Lambda}\frac{|\nu-\nu_\tau|}{\sqrt{\nu_0+\frac{1}{l+k}}}>\varepsilon\right\}<\exp\left(G(l+k)-2\left(\varepsilon^2\frac{(lk)^2}{(l+k)^3}-1\right)\right) \tag{11}$$

Remember that the right part of the above inequality is δ. So expressing all in terms of ε^2, we get: $\varepsilon^2=\frac{G(l+k)-\ln\delta+2}{2(lk)^2}(l+k)^3$. Pluggin this formula in (4) we get the final expression of the bound where ε^2 is explicit.

Finally we have to check the correctness of $\bar{\varepsilon}\frac{lk}{l+k}>2$ hypothesis. In other terms we have to verify that $\varepsilon^2>4\frac{(l+k)^3}{(lk)^2}$. So we get $\varepsilon^2=\frac{G(l+k)-\ln\delta+2}{2(lk)^2}(l+k)^3>4\frac{(l+k)^3}{(lk)^2}$ that produces the condition of the theorem: $G(l+k)-\ln\delta>6$.

4 Valuation and experimental results

The obtained result is valid when the Grow function is explicitly known. If Grow Function is not exactly known, Sauer lemma can be used to get a bound in terms of Vapnik-Chervonenkis dimension d_{vc} that leads to:

$$\nu_\tau \leq \nu + \frac{\beta}{4(l^2k)}(l+k)^2 + \sqrt{\frac{\beta}{2(lk)^2}(l+k)^3}\sqrt{\nu + k^2\frac{\beta}{8(lk)^2}(l+k)} \tag{12}$$

where $\beta = d_{vc}\left(1 + \ln\frac{l+k}{d_{vc}}\right) - \ln\delta + 2$.

Note also that a when typical confidence value of .95 is used, the theorem hypothesis is $G(l+k) > 3$, that is very likely to happen in practice.

There are others aspects that need analysis: first of all it is appropriate to observe that for $k = l$ the bound becomes:

$$\nu_\tau \leq \nu + \frac{G(2l) - \ln\delta + 2}{l} + 2\sqrt{\frac{G(2l) - \ln\delta + 2}{l}}\sqrt{\nu + \frac{G(2l) - \ln\delta + 2}{4l}} \tag{13}$$

From a cognitive and mathematical point of view keeping $k = l$ and requesting l, k big enough makes the above bound quite similar to original Vapnik inductive bound; these bounds became very similar when the Grow Function is far less, in absolute value, than the number of patterns (e.g. this can happen in clustering based classifiers). For completeness of information original Vapnik formulation was:

$$\pi \leq \nu + 2\frac{G(2l) - \ln\delta + 2\ln 2}{l} + 2\sqrt{\frac{G(2l) - \ln\delta + 2\ln 2}{l}}\sqrt{\nu + \frac{G(2l) - \ln\delta + 2ln2}{l}} \tag{14}$$

It is important to note down that in this case ($k = l$) the obtained bound is always convenient over induction (see figure2). Roughly speaking explicit transduction bound is convenient over induction in this case because we did not pay the price of the ghost set, because ghost set in this setting exists and it is represented by k patterns. When k and l are unbalanced this advantage is lost due to the behaviour of the hypergeometric distribution.

Now we want to present a comparison of the obtained result respect the bounds obtained in [4] via PAC-Bayesian arguments. In our experimental environment we always choose $1 - \delta = .95$ and $p(h) = 1/3$ so getting $N_{l+k} = \exp(3)$ for the number of possible functions to satisfy hypothesis of our theorem.

The bounds that we are going to compare are: the bound of this work, corollary 23 bound [4] and Serfling bound [4]. We propose the same experiments performed on figure 1 in [4].

As can be seen the obtained bound is tigther respect those obtained in [4] in 3 cases over 4 and this confirms the valuability of the result.

5 Conclusions

Here we presented a possible approach for building an explicit and simple to use transductive bound by using only Vapnik theory and without requiring any bayesian

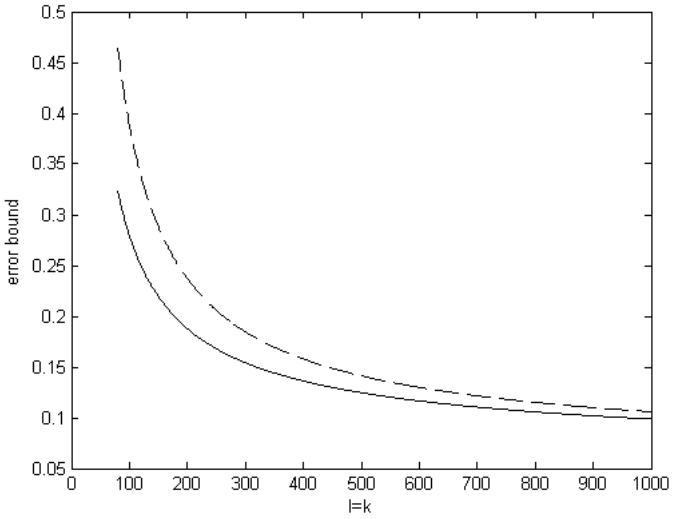

Fig. 2 Experiment for $k = l$ variable. Note the advantage of transductive bound (solid line) over induction (dashed line)

approach. We don't claim that is the best possible transductive bound, instead we want to underline the feasibility of the result respect to the transductive issue and respect more complicated arguments. Other improvements are possible in two directions: by using better concentration inequalities of the hypergeometric distribution or by using a Rademacher complexity approach, that at our knowledge, for now is only conceived in induction problems. Another interesting direction of research is trying to reproduce Vapnik machinery over the problem of building semi-supervised generalization error bounds: this aspect is a completely open problem and much theory lacks for a broad understanding.

References

1. Chapelle, O., B.Scholkopf, A.Zien: A Discussion of Semi-Supervised Learning and Transduction. In: Semi-Supervised Learning. MIT Press (2006)
2. Vapnik, V.: Estimation of Dependences Based on Empirical Data. Springer-Verlag (1982)
3. D.Hush, C.Scovel: Concentration of the hypergeometric distribution. Statistics Probability Letters **75** (2005) 127–132
4. P.Derbeko, R.El-Yaniv, R.Meir: Explicit learning curves for transduction and application to clustering and compression algorithms. Journal of Artificial Intelligence Research **22** (2004) 117–142

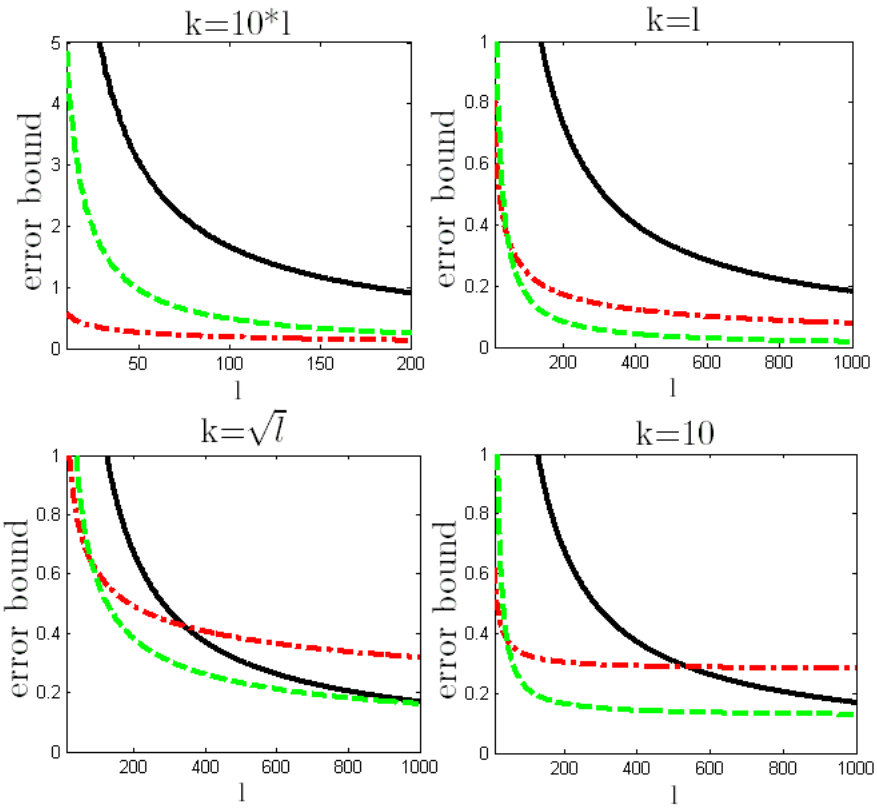

Fig. 3 Same experiments as in [4]. Note the advantage of the obtained bound (dashed line) over the other bounds (3 cases over 4).