Exploiting Task Relatedness for Multiple Task Learning

Shai Ben-David
Department of Computer Science
Technion, Haifa 32000, Israel
and Cornell University
Ithaca, NY 14853
shai@cs.cornell.edu

Reba Schuller
Department of Mathematics
Cornell University
Ithaca, NY 14853
ras51@cornell.edu

Abstract

The approach of learning of multiple "related" tasks simultaneously has proven quite successful in practice; however, theoretical justification for this success has remained elusive. The starting point of previous work on multiple task learning has been that the tasks to be learnt jointly are somehow "algorithmically related", in the sense that the *results* of applying a specific learning algorithm to these tasks are assumed to be similar. We take a logical step backwards and offer a data generating mechanism through which our notion of task-relatedness is defined.

We provide a formal framework for task relatedness that captures a certain sub-domain of the wide scope of issues in which one may apply a multiple task learning approach. Our notion of similarity between tasks is relevant to a variety of real life multi-task learning scenarios and allows the formal derivation of strong generalization bounds (bounds that are strictly stronger than the previously known bounds for both the learning-to-learn and the multi-task-learning scenarios). We provide general conditions under which our bounds guarantee smaller sample size per task than the known bounds for the single task learning approach.

1 Introduction

Most approaches to machine learning focus on the learning of a single isolated task. While great success has been achieved in this type of framework, it is clear that it neglects certain fundamental aspects of human learning. Human beings face each new learning task equipped with knowledge gained from previous learning tasks. There is no question that mankind would be seriously hindered if we simply threw away the knowledge gained from one learning task before commencing another, rather than using each learning task to become a better learner. Furthermore, human learning frequently involves approaching several learning tasks simultaneously; in particular, humans take advantage of the opportunity to compare and contrast similar categories in learning to classify entities into those categories. For example,

most of us probably learned the alphabet by learning several similar letters at the same time.

It is natural to attempt to apply these observations to machine learning–what kind of advantage is there in setting a learner to work on several tasks sequentially or simultaneously? Intuitively, there should certainly be some advantage, especially if the tasks are closely related in some way. And, indeed, much experimental work [1, 5, 6] has validated this intuition. However, thus far, there has been relatively little progress on any sort of theoretical justification for these results.

Relatedness of tasks is key to the multi task learning (MTL) approach. Obviously, one cannot expect that information gathered through the learning of a set of tasks will be relevant to the learning of another task that has nothing in common with the already learned set of tasks.

Previous work on MTL (or Learning to Learn) treated the notion of relatedness using a 'functional' approach. Consider for example Baxter's Learning To Learn work, e.g., [2] (which is, to our knowledge, the most systematic theoretical analysis of the simultaneous learning approach). In Baxter's work the similarity between jointly learned tasks is manifested solely through a model selection criterion, namely, the advantages of learning tasks together relies on the assumption that the tasks share a common optimal hypothesis class (or inductive bias).

We take a step backwards. We introduce a data generating framework through which a notion of task relatedness is defined. Not surprisingly, by limiting the discussion to problems that can be modelled by our data generating mechanism we leave many potential MTL scenarios outside the scope of our discussion. However, there are several interesting problems that can be treated within our framework. For these problems we can reap the benefits of having a mathematical notion of relatedness and prove sample size upper bounds for MTL learning that are far better than any previous proven bounds.

The rest of the paper is organized as follows: Section 2 formally introduces multiple task learning and describes our notion of task similarity, and. We state our generalization error bound for this framework in section 3 and in section 4, we compare these results for multiple task learning to the known bounds for the single task approach. We close with concluding remarks and directions for future work in section 5.

2 A Data Generation Model for Related Tasks

Formally, the typical classification learning problem is framed as follows: Given a domain $\mathcal{X}$ and a random sample S drawn from some unknown distribution P on $\mathcal{X} \times \{0,1\}$, find a hypothesis $h : \mathcal{X} \to \{0,1\}$ which approximates P (i.e., h such that for randomly drawn (x,b), with high probability $h(x) = b$). This problem is some times referred to as "statistical regressions".

The multiple task learning problem is the analogous problem for multiple distributions. That is, given domain $\mathcal{X}$ and sequence of random samples $S_1, \ldots, S_n$ drawn from some unknown distributions $P_1, \ldots P_n$, respectively, on $\mathcal{X} \times \{0,1\}$, find hypotheses $h_1, \ldots, h_n : \mathcal{X} \to \{0,1\}$ which approximate $P_1, \ldots, P_n$, respectively.

As we have mentioned previously, it is intuitive that the advantage of the multiple task approach depends on the "relatedness" between the different tasks. While there has been empirical success with sets of tasks related in various ways, thus far, no formal definition of "relatedness" has provided any theoretical results to this

effect.

2.1 Our Notion of Relatedness Between Learning Tasks

We define a data generation mechanism which serves to determine our notion of related tasks. Our data generation model is an extension of the *agnostic learning* framework.

The basic ingredient in our definition is a set $\mathcal{F}$ of transformations $f : \mathcal{X} \to \mathcal{X}$. We say that tasks are $\mathcal{F}$- related if, for some fixed probability distribution over $\mathcal{X} \times \{0,1\}$, the data in each of these tasks is generated by applying some $f \in \mathcal{F}$ to this fixed distribution.

Definition 2.1 *Let $\mathcal{F}$ be a set of transformations $f : \mathcal{X} \to \mathcal{X}$, and let P_1, P_2 be probability distributions over $\mathcal{X} \times \{0,1\}$. We say that P_1, P_2 are $\mathcal{F}$-related distributions if there exists some $f \in \mathcal{F}$ such that for any $T \subseteq \mathcal{X} \times \{0,1\}$, T is P_1-measurable iff $f[T] = \{(f(x), b) \mid (x, b) \in T)\}$ is P_2 - measurable and $P_1(T) = P_2(f[T])$.*

Note that the strength of this definition depends of the richness of the family of transformations $\mathcal{F}$. The larger this set gets, the loser is the notion of $\mathcal{F}$- relatedness. Clearly there are many examples of potential applications of simultaneous learning that do not fit into this model of relatedness. However, there are various interesting examples where this notion seems to provide a satisfactory mathematical model of the similarity between the tasks of a set of related learning problems.

Typically our notion of $\mathcal{F}$ relatedness arises in situations where many different sensors collect data for the same classification problem. For example, consider a set of cameras located in different locations in the lobby of some high security building. Assume that they are all used to automatically detect unauthorized visitors, based on the images they record. Clearly, each of these cameras has its own bias, due to a different hight, light conditions, angle, etc. While it may be difficult to determine the exact bias of each camera, it may be feasible to define mathematically a set of image transformations $\mathcal{F}$ such that the data distributions of images collected by of all these recorders are $\mathcal{F}$- related.

Another area in which such a notion of similarity is applicable is that of database integration. Suppose there are several databases available, each of which obtains its information from the same data pool, yet represents its information with a different database schema. For the purpose of classification prediction, our results in the next section eliminate the need for the difficult undertaking of database integration, treating each database as one task in a multiple task learning problem.

In the following section, we show that our notion of relatedness yields a bound on the generalization error for *each* task.

3 Learning $\mathcal{F}$-similar Tasks

In this section, we analyze multiple task learning for $\mathcal{F}$-related tasks. Formally, given domain $\mathcal{X}$, hypothesis space $\mathbb{H}$ on $\mathcal{X}$, set of transformations $\mathcal{F}$ on $\mathcal{X}$, and sequence $S_1, \ldots, S_n$ of samples from some sequence of $\mathcal{F}$-similar distributions, $P_1, \ldots, P_n$ on $\mathcal{X} \times \{0,1\}$, select $h_1, \ldots, h_n \in \mathbb{H}$ which make good classification predictions for $P_1, \ldots, P_n$.

We proceed by addressing this problem in two phases. The first phase makes use of all of the samples to reduce the size of our hypothesis space by selecting a subspace of $\mathbb{H}$. The second phase then uses standard learning techniques to select a hypothesis

from this subspace for each task separately. The advantage of this approach is that, generally, the first phase reduces the complexity (e.g. VC-dimension) of the hypothesis space, thus reducing the sample complexity of the second phase.

In general, our approach to the first phase is as follows: Define a partition $\mathbb{H}$ into a family of subspaces. Then choose a subspace, H, to minimize the average empirical error over the different tasks.

Our main result is a bound on the generalization error for this approach.

We now describe explicitly our method for partitioning $\mathbb{H}$. From the given hypothesis space, $\mathbb{H}$, we create a family, $\mathbf{H}$, of hypothesis spaces consisting of sets of hypothesis in $\mathbb{H}$ which are equivalent up to transformations in $\mathcal{F}$. Without loss of generality, we assume that $\mathcal{F}$ forms a group under function composition and that $\mathbb{H}$ is closed under the action of $\mathcal{F}$ (i.e., for each $f \in \mathcal{F}$ and each $h \in \mathbb{H}$, we have $h \circ f \in \mathbb{H}$).

Definition 3.1 *Given hypothesis space $\mathbb{H}$ and transformation family $\mathcal{F}$ on domain $\mathcal{X}$, define equivalence relation $\sim_{\mathcal{F}}$ on $\mathbb{H}$ by:*

$$h_1 \sim_{\mathcal{F}} h_2 \text{ iff there exists } f \in \mathcal{F} \text{ such that } h_2 = h_1 \circ f.$$

Now, we let our family of hypothesis spaces, $\mathbf{H}$, be the family of all equivalence classes of $\mathbb{H}$ under $\sim_{\mathcal{F}}$, i.e., $\mathbf{H} = \mathbb{H}/\sim_{\mathcal{F}}$. In this scenario,

$$Er^{P_j}(h \circ f_j^{-1}) = Er^{P}(h) \text{ for any } h \text{ and any } 1 \leq j \leq n, \tag{1}$$

where P is the common underlying distribution from the definition of $\mathcal{F}$-similar (definition 2.1). Using this fact, we can deduce that the equivalence classes of $\mathbb{H}$ perform equally well on the different tasks in the following sense.

Definition 3.2 *For any hypothesis space, H, define*

$$Er^{P}(H) = \inf_{h \in H} Er^{P}(h).$$

Lemma 3.3 *Let P_1, P_2 be $\mathcal{F}$-similar tasks, and $\mathcal{F}$ be a group under function composition. If H is closed under the action of $\mathcal{F}$ then $Er^{P_1}(H) = Er^{P_2}(H)$.*

Proof: We need to show that

$$\inf_{h \in H} Er^{P_1}(h) = \inf_{h \in H} Er^{P_2}(h).$$

It suffices to show that for every $h \in H$ there exist $h', h'' \in H$ such that $Er^{P_2}(h') \leq Er^{P_1}(h)$ and $Er^{P_1}(h'') \leq Er^{P_2}(h)$.

Let P, f_1, f_2 be as in the definition of $\mathcal{F}$-similar distributions (definition 2.1), i.e., P_i corresponds to P modified according to f_i. Given h, let $h' = h \circ f_1 \circ f_2^{-1}$ and $h'' = h \circ f_2 \circ f_1^{-1}$. By equation 1, $Er^{P_2}(h') = Er^{P_1}(h)$ and $Er^{P_1}(h'') = Er^{P_2}(h)$, so we are done. ∎

Corollary 3.4 *For any $h \in \mathbb{H}$ and any $1 \leq j \leq n$,*

$$Er^{P_j}([h]_{\sim_{\mathcal{F}}}) = \inf_{h_1, \ldots, h_n \in [h]_{\sim_{\mathcal{F}}}} \frac{1}{n} \sum_{i=1}^{n} Er^{P_i}(h_i).$$

We are now ready to state and prove our main result, which gives an upper bound on the sample complexity of finding a $\sim_{\mathcal{F}}$-equivalence class which is near-optimal for *each* of the tasks.

Theorem 3.5 *For any $0 \leq \epsilon, \delta \leq 1$ and $h \in \mathbb{H}$, if $S_1, \ldots, S_n$ is an $\mathcal{F}$-similar sequence of samples corresponding to $P_1, \ldots, P_n$, with $|S_i| \geq m$ for all i, where*

$$m \geq \frac{88}{\epsilon^2}\left[2d_{\mathbf{H}}(n)\log\frac{22}{\epsilon} + \frac{1}{n}\log\frac{4}{\delta}\right],$$

then with probability at least $1-\delta$, for any $1 \leq j \leq n$,

$$\left| Er^{P_j}([h]_{\sim_{\mathcal{F}}}) - \inf_{h_1,\ldots,h_n \in [h]_{\sim_{\mathcal{F}}}} \frac{1}{n}\sum_{i=1}^{n} \hat{Er}^{S_i}(h_i) \right| \leq \epsilon.$$

Theorem 3.5 follows almost directly from theorem 3.8 and corollary 3.4.

Note that combining the standard generalization error result for single task learning with 3.5 gives an information complexity bound for these two phases together. We state this bound explicitly in the next section.

3.1 Background from Baxter [2]

Baxter's generalization error bound for inductive bias depends on the following notion of generalized VC-dimension for families of hypothesis spaces.

Notation: For function $g : Y \to Z$ and $\overline{y} = (y_1, \ldots, y_n) \in Y^n$, $\overline{g}(\overline{y})$ will denote $(g(y_1), \ldots, g(y_n)) \in Z^n$.

Definition 3.6 *For family $\mathbf{H}$ of hypothesis spaces, and $n, m \in \mathbb{Z}^+$,*

$$\Pi_{\mathbf{H}}(n,m) = \max_{\overline{x}_1,\ldots,\overline{x}_n \in X^m} \left| \left\{ \begin{bmatrix} \overline{h_1}(\overline{x_1}) \\ \vdots \\ \overline{h_l}(\overline{x_n}) \end{bmatrix} : \exists H \in \mathbf{H} \text{ with } h_1, \ldots, h_n \in H \right\} \right|.$$

Definition 3.7 $d_{\mathbf{H}}(n) = \max\{m \ : \ \Pi_{\mathbf{H}}(n,m) = 2^{nm}\}$.

We can now state the necessary result from [2] on multitask learning, which appears as corollary 13 in [2]. [1]

Theorem 3.8 *Let $\mathbf{H}$ be any permissible boolean hypothesis space family[2], and let $S_1, \ldots, S_n$ be a sequence of random samples from distributions $P_1, \ldots, P_n$ on $\mathcal{X} \times \{0,1\}$. If the number of examples m in each sample S_i satisfies*

$$m \geq \frac{88}{\epsilon^2}\left[2d_{\mathbf{H}}(n)\log\frac{22}{\epsilon} + \frac{1}{n}\log\frac{4}{\delta}\right],$$

then with probability at least $1-\delta$ (over the choice of $S_1, \ldots, S_n$), for any $H \in \mathbf{H}$, and $h_1, \ldots, h_n \in H$,

$$\left| \frac{1}{n}\sum_{i=1}^{n} Er^{P_i}(h_i) - \frac{1}{n}\sum_{i=1}^{n} \hat{Er}^{S_i}(h_i) \right| \leq \epsilon.$$

[1] Note that although [2] only states that $\frac{1}{n}\sum_{i=1}^{n} Er^{P_i}(h_i) \leq \frac{1}{n}\sum_{i=1}^{n} \hat{Er}^{S_i}(h_i) + \epsilon$, it is clear from the proofs in [2] that this stronger form holds.

[2] Permissibility, discussed in [4] is a "weak measure-theoretic condition satisfied by almost all 'real-world' hypothesis space families" ([2], Appendix D). Throughout this paper we shall assume that all our classes are permissible.

4 Multiple Task Learning Versus the Single Task Approach

We have provided upper bounds on the information complexity of multiple task learning where the tasks are related via some set of transformations, $\mathcal{F}$. We now address the question of how this compares to the information complexity of single task learning.

In the following, let D =VC-dim($\mathbb{H}$), $d_{max} = \max_{h\in\mathbb{H}}$VC-dim($[h]_{\sim_{\mathcal{F}}}$).

The standard single task learning result [7] guarantees that for a single sample S_0 sampled from distribution P_0, $|S_0| \geq$

$$(64/\epsilon^2)[\log(4/\delta) + 2D\log(12/\epsilon)] \tag{2}$$

is sufficient to ensure that with probability at least $1-\delta$,

$$\left| Er^{P_0}(h) - \hat{Er}^{S_0}(h) \right| \leq \epsilon.$$

Analogously, the results in this paper guarantee that for n $\mathcal{F}$-similar tasks, the total number of examples needed is at most

$$\begin{aligned} n \times \max\left(\frac{352}{\epsilon^2}\left[2d_{\mathbf{H}}(n)\log\frac{44}{\epsilon} + \frac{1}{n}\log\frac{8}{\delta}\right], \left(\frac{256}{\epsilon^2}\right)\left[2d_{max}\log\frac{24}{\epsilon} + \log\frac{8}{\delta}\right]\right) \\ < \quad n \times \frac{352}{\epsilon^2}\left[2d_{\mathbf{H}}(n)\log\frac{44}{\epsilon} + \frac{8}{11}\log\frac{8}{\delta}\right]. \end{aligned} \tag{3}$$

It is clear that if n is relatively large, then this bound for the total number examples required for the multitask approach is not an improvement over the single task approach. This is not surprising. However, the number of examples needed *per task* is at most $\frac{1}{n}th$ of this quantity. Thus, if $d_{\mathbf{H}}(n) \ll D$, then the information complexity *per task* is less for learning n tasks than for learning a single data task. This means that if the $\sim_{\mathcal{F}}$-equivalence classes of $\mathbb{H}$ are sufficiently less rich than $\mathbb{H}$ itself, then one can compensate for insufficient training data for a task by using additional tasks.

Now, in order to compare the *per task* information complexity advantage of multiple task learning, we must first compare $d_{\mathbf{H}}(n)$ and $VC\text{-}dim(\mathbb{H})$.

4.1 Analysis of $d_{\mathbf{H}}(n)$

It is easy to see that $d_{max} \leq d_{\mathbf{H}}(n) \leq D$. Thus, the best we can hope for is $d_{\mathbf{H}}(n) = d_{max}$. The following theorem gives a general scenario in which this bound is attained.

Theorem 4.1 *If there exists M such that $|h| \leq M$ for all $h \in \mathbb{H}$, then there exists n_0 such that for all $n \geq n_0$,*

$$d_{\mathbf{H}}(n) = \max_{h\in\mathbb{H}} VC\text{-}dim([h]_{\sim_{\mathcal{F}}}).$$

Proof: Assume $d_{\mathbf{H}}(n) \geq m$, and let $\overline{x_1}, \ldots, \overline{x_n}$ be such that

$$\left| \left\{ \begin{bmatrix} \overline{h\circ f_1}(\overline{x_1}) \\ \vdots \\ \overline{h\circ f_n}(\overline{x_n}) \end{bmatrix} : f_1 \ldots f_n \in \mathcal{F}, h \in \mathbb{H} \right\} \right| = 2^{nm}$$

Consider $h_0 \in \mathbb{H}$ and $f_1, \ldots, f_n$ such that

$$\begin{bmatrix} \overline{h_0 \circ f_1}(\overline{x_1}) \\ \vdots \\ \overline{h_0 \circ f_n}(\overline{x_n}) \end{bmatrix} = \begin{bmatrix} 1 & \ldots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \ldots & 1 \end{bmatrix}$$

Note that for each i, there exists $S_i \subseteq h_0$ such that $\overline{x_i}$ is some permutation of $\{f_i^{-1}(z) \ : \ z \in S_i\}$.

Say $|h_0| = K$. Then if $n > \left(\binom{K}{m} - 1\right) 2^m$, then there exists $S \subseteq h_0$ and $i_1, \ldots i_{2^m}$ such that $S_{i_j} = S$ for $1 \leq j \leq 2^m$. Let $\sigma_1, \ldots \sigma_{2^m}$ be the corresponding permutations.

Finally, letting $v_1, \ldots, v_{2^m}$ be an enumeration of all vectors of length m over $\{0,1\}$, letting N be any $m \times n$ matrix over $\{0,1\}$ whose i_j^{th} row is $\sigma_j(v_{i_j})$, and letting h_* and $f'_1, \ldots, f'_n$ be such that

$$\begin{bmatrix} \overline{h_* \circ f'_1}(\overline{x_1}) \\ \vdots \\ \overline{h_* \circ f'_n}(\overline{x_n}) \end{bmatrix} = N,$$

we see that $[h_*]_{\sim_{\mathcal{F}}}$ shatters S, so $m \leq \text{VC-dim}([h_*]_{\sim_{\mathcal{F}}})$.

To eliminate the dependence on $|h_0| = K$, we set $n_0 = \left(\binom{M}{M/2} - 1\right) 2^M$, noting that $n_0 \geq \left(\binom{K}{m} - 1\right) 2^m$ for all $K, m \leq M$. □

Ben-David, et. al. [3] provide the following further results on $d_{\mathbf{H}_{\sim_{\mathcal{F}}}}(n)$.

Theorem 4.2 *If $\mathcal{F}$ is finite and $\frac{n}{\log(n)} \geq$ VC-dim$(\mathbb{H})$, then $d_{\mathbf{H}_{\sim_{\mathcal{F}}}}(n) \leq 2\log(|\mathcal{F}|)$*

Theorem 4.3 *If $\sim_{\mathcal{F}}$ is of finite index k, and $n \geq \frac{\log k}{4d \log d}$, then $d_{\mathbf{H}_{\sim_{\mathcal{F}}}}(n) \leq \frac{\log k}{n} + 4d \log d$, where*

$$d = \max\left(\max_{H \in \mathbb{H}/\sim_{\mathcal{F}}} \text{VC-dim}(H), 3\right).$$

4.2 When Multiple Task Learning is Provably Advantageous

As we observed earlier, if $d_{\mathbf{H}}(n) \ll D$, then the multiple task approach provides a provable information gain. Observe that $d_{\mathbf{H}}(n) < D/8$ is sufficient to ensure that $1/n$ times the quantity in equation 3 is less than the quantity in equation 2. The analysis in section 4.1 provides some cases where this holds. In particular, theorem 4.3 gives us this guarantee for sufficiently large n, provided that $(d_{max}) \log(d_{max}) < D/8$. Additionally, theorem 4.2 gives us the following corollary.

Corollary 4.4 *If $\mathcal{F}$ is finite, then given a sequence of $\mathcal{F}$-similar tasks, the multiple task approach offers an information complexity advantage over the single task approach for any hypothesis space $\mathbb{H}$ with VC-dim$(\mathbb{H}) \leq 16\log(|\mathcal{F}|)$.* [3]

[3] Assuming $\mathcal{F}$ is closed under function composition and inverse, and $\mathbb{H}$ is closed under the action of $\mathcal{F}$.

5 Conclusions and Future Work

We have presented a useful notion of relatedness between tasks for multiple task learning. Our notion of relatedness provides a natural model for learning situations in which data for the same classification task is collected by a set of similar yet different recording devices. We derived generalization error bounds for learning of multiple tasks related in this manner, and provided general conditions under which these bounds guarantee better generalization than the known bounds for the single task approach. This is a small but significant step towards the goal of a full theory of multiple task learning. Due to the restriction to a special type of relatedness of tasks, we have been able to obtain sample size bounds which are significantly better than previously proven bounds for the learning to learn scenario.

Hopefully, this work will stimulate future work in several directions. There is room for a more thorough understanding of the conditions under which multi-task learning is advantageous over the single task approach in our scenario. It would also be fruitful to relax the requirements on the set of transformations through which the tasks are related, allowing these transformations to be arbitrary rather than bijections, and perhaps even allowing the actual transformations between the tasks to be merely approximated by the set of known transformations. Finally, the quest for more general notions of similarity between tasks remains the key to a thorough understanding of multiple task learning.

We believe that this work provides convincing evidence that a theoretical understanding of multiple task learning is a promising research endeavor worth pursuing.

References

[1] Jonathan Baxter. Learning internal representations. In *COLT: Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, 1995.

[2] Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

[3] Shai Ben-David, Johannes Gehrke, and Reba Schuller. A theoretical framework for learning from a pool of disparate data sources. In *Proceedings of the The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002.

[4] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, 1989.

[5] N. Intrator and S. Edelman. How to make a low-dimensional representation suitable for diverse tasks. *Connection Science*, 8, 1996.

[6] S. Thrun. Is learning the n-th thing any easier than learning the first? In D. Touretzky and M Mozer, editors, *Advances in Neural Information Processing Systems (NIPS)*, pages 640–646, 1996.

[7] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theoret. Probl. And Its Appl*, 16(2):264–280, 1971.