Concept Learning using Complexity Regularization[1]

Gábor Lugosi[2] and Kenneth Zeger[3]

[2] Dept. of Mathematics, Faculty of Elect. Engineering, Technical University of Budapest, Hungary.
email: lugosi@vma.bme.hu.

[3] Coordinated Science Lab., Dept. of Elect. and Comp. Engineering, University of Illinois, Urbana-Champaign, IL 61801
email: zeger@uiuc.edu.

Abstract — **We apply the method of complexity regularization to learn concepts from large concept classes. The method is shown to automatically find the best balance between the approximation error and the estimation error. In particular, the error probability of the obtained classifier is shown to decrease as $O(\sqrt{\log n/n})$ to the achievable optimum, for large nonparametric classes of distributions, as the sample size n grows.**

In pattern recognition—or concept learning—the value of a $\{0,1\}$-valued random variable Y is to be predicted based upon observing an $\mathcal{R}^d$-valued random variable X. A *prediction rule* (or *decision*) is a function $\phi : \mathcal{R}^d \to \{0,1\}$, whose performance is measured by its error probability $\mathbf{P}\{\phi(X) \neq Y\}$. The error probability $L^* = \mathbf{P}\{g^*(X) \neq Y\}$ of the optimal decision g^* is called the Bayes risk. Assume that a training sequence

$$D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$$

of independent, identically distributed random variables is available, where the (X_i, Y_i) have the same distribution as (X, Y), and D_n is independent of (X, Y). A *classifier* is a function $\phi_n : \mathcal{R}^d \times (\mathcal{R}^d \times \{0,1\})^n \to \{0,1\}$, whose error probability is the random variable $L(\phi_n) = \mathbf{P}\{\phi_n(X, D_n) \neq Y | D_n\}$.

The method of empirical risk minimization picks a classifier from a class $\mathcal{C}$ of functions $\mathcal{R}^d \to \{0,1\}$ that minimizes the empirical error probability over $\mathcal{C}$. More precisely, define the empirical error probability of a decision ϕ by $\hat{L}_n(\phi) = (1/n)\sum_{i=1}^n I_{\{\phi(X_i) \neq Y_i\}}$, where I denotes the indicator function. Let $\tilde{\phi}_n$ denote a classifier chosen from $\mathcal{C}$ by minimizing $\hat{L}_n(\phi)$, i.e., $\hat{L}_n(\tilde{\phi}_n) \leq \hat{L}_n(\phi)$, $\phi \in \mathcal{C}$. Vapnik and Chervonenkis [4], [5] proved distribution-free exponential inequalities for empirical error minimization. One of the implications is that $\mathbf{E}L(\tilde{\phi}_n) - \inf_{\phi \in \mathcal{C}} L(\phi) \leq c\sqrt{(V \log n)/n}$, where V is the VC dimension of the class $\mathcal{C}$ and c is a universal constant (independent of the distribution). Thus, the error probability of the empirically chosen decision is always within $O(\sqrt{\log n/n})$ of that of the best in $\mathcal{C}$. Unfortunately, if $V < \infty$, then for some distributions, $\inf_{\phi \in \mathcal{C}} L(\phi)$ may be arbitrarily far from L^* On the other hand, if $V = \infty$, then $L(\tilde{\phi}_n) - \inf_{\phi \in \mathcal{C}} L(\phi)$ will be large for some distributions [3], [5].

A possible solution to this problem may be derived from the idea of *structural risk minimization* (Vapnik and Chervonenkis [5]), also known as *complexity regularization* (see Barron [1], Barron and Cover [2]). The basic idea is to minimize the sum of the empirical error and a term corresponding to the "complexity" of the candidate classifier. In our application, this complexity is a simple function of the VC dimension of the class from which the candidate classifier is taken.

Theorem 1 *Let $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \ldots$ be a sequence of classes of classifiers whose VC dimensions $V_1, V_2, \ldots$ are finite. Let ϕ_n^* be the classification rule based on structural risk minimization. Then for all n,*

$$\mathbf{E}\{L(\phi_n^*)\} - L^* \leq \inf_{k \geq 1} \left(\sqrt{\frac{16 V_k \log n + 8(k+11)}{n}} + \left(\inf_{\phi \in \mathcal{C}^{(k)}} L(\phi) - L^* \right) \right).$$

This result is close on spirit of those obtained by Barron [1], and Barron and Cover [2], who select a classifier from a countable list of candidates by minimizing the sum of the empirical error and a properly chosen penalty. A significant difference is that the method we study here does not restrict the search to a countable set of candidates, allowing thus better approximation ability.

Corollary 1 *Let $\mathcal{C}^{(1)}, \mathcal{C}^{(2)}, \ldots$ be a sequence of classes of classifiers such that the VC dimensions $V_1, V_2, \ldots$ are all finite. Assume further that the Bayes rule is contained in the union of these classes, i.e., $g^* \in \mathcal{C}^* \stackrel{\text{def}}{=} \cup_{j=1}^{\infty} \mathcal{C}^{(j)}$. Let K be the smallest integer such that $g^* \in \mathcal{C}^{(K)}$. Then for every n, the error probability of the classification rule based on structural risk minimization, ϕ_n^*, satisfies*

$$\mathbf{E}L(\phi_n^*) - L^* \leq 4\sqrt{\frac{V_K \log n + K/2 + 6}{n}}.$$

Corollary 1 shows that the rate of convergence is always of the order of $\sqrt{\log n/n}$, and the constant factor V_K depends on the distribution. The number V_K may be viewed as the inherent complexity of the Bayes rule for the distribution. One great advantage of structural risk minimization is that it finds automatically where to look for the optimal classifier.

References

[1] A. R. Barron. Complexity regularization with application to artificial neural networks. In G. Roussas, editor, *Nonparametric Functional Estimation and Related Topics*, pages 561–576, Dordrecht, 1991. NATO ASI Series, Kluwer Academic Publishers.

[2] A. R. Barron and T. M. Cover. Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37:1034 – 1054, 1991.

[3] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36:929–965, 1989.

[4] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[5] V. N. Vapnik and A. Ya. Chervonenkis. *Theory of Pattern Recognition.* Nauka, Moscow, 1974. (in Russian); German translation: *Theorie der Zeichenerkennung*, Akademie Verlag, Berlin, 1979.

[1] The research was supported in part by the National Science Foundation under Grants No. NCR-92-96231 and INT-93-15271.