INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 666-2900 • FAX (510) 666-2956

Semi-Supervised Model Selection Based on Cross-Validation

Matti Kääriäinen
International Computer Science Institute
mtkaaria@icsi.berkeley.edu

TR-05-010

September 2005

Abstract

We propose a new semi-supervised model selection method that is derived by applying the structural risk minimization principle to a recent semi-supervised generalization error bound. This bound that we build on is based on the cross-validation estimate underlying the popular cross-validation model selection heuristic. Thus, the proposed semi-supervised method is closely connected to cross-validation which makes studying these methods side by side very natural.

We evaluate the performance of the proposed method and the cross-validation heuristic empirically on the task of selecting the parameters of support vector machines. The experiments indicate that the models selected by the two methods have roughly the same accuracy. However, whereas the cross-validation heuristic only proposes which classifier to choose, the semi-supervised method provides also a reliable and reasonably tight generalization error guarantee for the chosen classifier. Thus, when unlabeled data is available, the proposed semi-supervised method seems to have an advantage when reliable error guarantees are called for. In addition to the empirical evaluation, we also analyze the theoretical properties of the proposed method and prove that under suitable conditions it converges to the optimal model.

1 Introduction

No learning algorithm is universally best on all learning problems and sample sizes. On the contrary, there exists a large variety of learning algorithms (support vector machines (SVMs), decision trees,...) and parametrizations thereof (the kernel and its parameters for SVMs, tree growing and pruning heuristics for decision trees,...) that each have their own advantages and weaknesses. Model selection is the task of deciding, given a sample of examples from a learning problem, which parametrized learning algorithm to use on the problem.[1]

We study model selection in a semi-supervised classifier learning setting. In this setting, the learning algorithm is given a labeled sample $S \sim D^n$ and an unlabeled sample $U \sim D_{\mathcal{X}}^m$, where D is an unknown distribution over the set $\mathcal{X} \times \mathcal{Y}$ of (example, label) pairs, and $D_{\mathcal{X}}$ is its marginal on $\mathcal{X}$. Based on these samples, the algorithm is supposed to output a classifier $f\colon \mathcal{X} \to \mathcal{Y}$ with small *generalization error* $\epsilon(f) = \mathbb{P}_{(x,y)\sim D}(f(x) \neq y)$. The goal of a learning algorithm is thus to output a classifier with small generalization error, while the goal of model selection is to find a model that succeeds in this. Thus, what differentiates this semi-supervised setting from the standard supervised model of statistical learning theory is the unlabeled sample that the learning or model selection algorithm can try to exploit. In what follows, we assume that the learning algorithms use only the labeled data, so the unlabeled data is used only in model selection. However, the results can be extended to cover the case of semi-supervised learning algorithms, too.

In practice, models are often selected using the cross-validation (CV) heuristic (or some of its variants). This is because the heuristic is easy to use, implement, and understand (at least on an intuitive level), requires no extra assumptions, and gives results that have proved to be hard to improve upon in general. However, learning theory has not yet been able to explain the success of the CV heuristic. In particular, there are no general guarantees on the quality of the models chosen by the CV heuristic, and it is indeed easy to come up with examples on which the heuristic fails. Characterizing the conditions under which the CV heuristic works is a major open problem in learning theory.

Instead of trying to analyze the CV heuristic directly, we propose an alternative semi-supervised model selection method that builds on a combination of the CV estimate and the structural risk minimization (SRM) principle. The proposed method requires a sample of unlabeled data, but is otherwise applicable whenever the CV heuristic is. The hope is that the proposed method should inherit the good observed model selection results of the CV heuristic by building on the same estimate, while employing the SRM principle should ensure we get reliable performance guarantees for the chosen model. As this indeed seems to be the case, the proposed method is potentially useful in itself. More importantly, it gives new insight into the workings of the CV heuristic by connecting it to the seemingly unrelated SRM framework.

We explore the properties of the proposed method empirically and theoretically. Our experiments show that the models (parametrizations of SVMs) chosen by the semi-supervised method achieve as low error rates as the models chosen by the CV heuristic. Thus, as far as the generalization ability of the chosen model is concerned, it seems that the methods do equally well. The error guarantees given by the proposed method are always non-trivial and potentially tight enough to be useful in practice. Thus, it seems that the proposed way of using unlabeled data in model selection at least does not hurt, and is beneficial if reliable error guarantees are called for. Our theoretical analysis shows that the proposed method is consistent in the sense that if one of the

[1] In this paper, the term model is used as a synonym for parametrized learning algorithm.

candidate models is consistent with the learning problem and a mild noise condition is satisfied, then our proposed method finds an optimal model in the limit.

The other semi-supervised model selection heuristics proposed in the literature [1, 2] are similar to ours in that they, too, use unlabeled data to estimate distances between hypotheses. The methods have been observed to do well on regression tasks, but to have problems in the classifier learning setting. Furthermore, they are directly connected to neither generalization error bounds nor cross-validation, so we will not consider them further here.

2 Model selection based on cross-validation estimates

Given a parametrized learning algorithm, the CV estimate is computed as follows:

- The labeled sample S is split randomly into k equisized folds S_i, $1 \leq i \leq k$. We assume for notational simplicity that the sample size $|S|$ is divisible by k.
- For each i, $1 \leq i \leq k$, the algorithm is run on the training data $\bigcup\{S_j \mid j \neq i\}$ to get a classifier f_i. The generalization error of f_i is estimated by its error rate on S_i given by $\hat{e}(f_i, S_i) = \frac{1}{|S_i|} \sum_{(x,y) \in S_i} I(f(x) \neq y)$.
- The algorithm is run one more time using the whole of S as training data to get a final classifier f. The CV estimate of $\epsilon(f)$ is the average of the estimates $\hat{e}(f_i, S_i)$.

Even though the CV estimate seems to be often close to $\epsilon(f)$ in practice, there is no guarantee that this should be the case in general. The first reason is that even though the random variables $\hat{e}(f_i, S_i)$ are unbiased estimates for $\epsilon(f_i)$, their variance may be large as both f_i and S_i are random. These random fluctuations of are not taken into account in the CV estimate. The second reason is that the random variables $\epsilon(f_i)$ may be far from $\epsilon(f)$ as nothing guarantees that the f_is are close to f. It is easy to construct learning algorithms whose behavior changes completely when the training set size increases from $(k-1)/k|S|$ to $|S|$. Such changes—known as phase transitions of the learning algorithm—have been observed in connection with real learning algorithms, too.

The CV estimate is used for model selection by evaluating it for each of a finite set of candidate models and choosing the model corresponding to the smallest estimate. Despite its shortcomings, this strategy—the CV model selection heuristic—is probably the most widely used model selection method in practice and has been observed to work very well on a large variety of problems.

3 Model selection based on generalization error bounds

3.1 Structural risk minimization

Simply choosing the model with minimal generalization error is impossible as the distribution D is unknown. To circumvent the problems arising from this, the goals of learning and model selection can be restated in terms of *generalization error bounds*. A generalization error bound is a random variables that depends only on the observed data and that provably upper bound the generalization error of the learned classifier with probability at least $1-\delta$. Here, $\delta > 0$ is a confidence parameter for the bound that upper bounds its probability of failure. If a generalization error bound is tight in the sense that it doesn't overshoot the true generalization error by much, then a classifier minimizing the generalization error bound has close to minimal generalization error (at least with probability

$1-\delta$). This may of course be true also if the bound is not tight, but it is hard to justify why this should be the case.

The principle of choosing the model with the best generalization error bound is called Structural Risk Minimization (SRM) [3]. Assuming for simplicity that we have a finite number N of candidate models, the SRM principle works as follows:

- Split the confidence parameter δ into shares $\delta_j > 0$ so that $\sum_{j=1}^{N} \delta_j = \delta$. It is possible to encode prior beliefs in the performance of the models by giving a larger share of δ to those models that are expected to perform well. In lack of such beliefs, we split δ uniformly to all candidate models.

- For each j, $1 \leq j \leq N$, run the learning algorithm and the generalization error bounding scheme corresponding to model j with confidence parameter δ_j. The result is a classifier f_j and a bound α_j satisfying $\epsilon(f_j) \leq \alpha_j$ with probability at least $1 - \delta_j$.

- Choose the model indexed by $j^* = \arg\min \alpha_j$, and output the corresponding classifier f_{j^*} as the final hypothesis.

By the union bound, α_{j^*} upper bounds the generalization error of the classifier corresponding to the chosen model with probability at least $1 - \delta$. Thus, if all the models are put into a black box containing the SRM machinery, the modularity introduced by the model selection phase disappears and the box will behave exactly as if it was a single learning algorithm equipped with a generalization error bounding scheme.

3.2 Model selection using the semi-supervised cross-validation bound

We apply the SRM principle to a recent semi-supervised generalization error bound [4]. The bound is obtained by first transforming the CV estimate to a generalization error bound for a related randomized classifier, and then applying a recent semi-supervised derandomization method [5] to this bound to obtain a bound for the classifier learned based on all labeled data.

The CV bound uses the ensemble of classifiers f_i learned during the CV procedure to build a *randomized classifier* f_{rand}. This f_{rand} works by selecting the f_i to use independently and uniformly at random for each example to be classified. The unlabeled data is then used in estimating the disagreement probability $d(f_{\text{rand}}, f) = P(f_{\text{rand}}(X) \neq f(X))$ of f_{rand} and f by

$$\hat{d}(f_{\text{rand}}, f) = \frac{1}{|U|} \sum_{x \in U} I(f(x) \neq f_{\text{rand}}(x)).$$

With this notation, the semi-supervised CV bound reads as follows [4].

Theorem 1 *Let f and f_{rand} be obtained by the k-fold CV procedure as explained above. Then with probability at least $1 - \delta$ over the choice of $S \sim D^n$, $U \sim D_{\mathcal{X}}^m$, and the randomness in f_{rand}, we have*

$$\epsilon(f) \leq \frac{1}{k} \sum_{i=1}^{k} \overline{\text{Bin}}\left(\hat{e}(f_i, S_i), |S_i|, \delta/(2k)\right) + \overline{\text{Bin}}\left(\hat{d}(f_{rand}, f), m, \delta/2\right).$$

The inverse binomial tail $\overline{\text{Bin}}(\hat{p}, m, \delta)$ is the q for which $\sum_{i=0}^{\lceil \hat{p}m \rceil} \binom{m}{i} q^i (1-q)^{m-i} = \delta$.

The first of the two terms in the bound is the CV estimate turned into an upper bound for the generalization error of the randomized ensemble f_{rand}. Its expectation is thus close to the test set bound one would obtain by using a $(k-1)/k$ fraction of labeled examples for training and the remaining $1/k$ fraction for testing (the only difference being the change $\delta \to \delta/(2k)$ in the confidence parameter). The second term uses unlabeled data to bound the potential change in generalization error that results from derandomizing f_{rand} by replacing it with the classifier f learned from all the data. It measures explicitly how far the classifiers f_i learned during the CV procedure are from the final classifier f learned on all labeled data. Thus, the second term will detect phase transitions and other changes in the behavior of the learning algorithm that may occur when the sample size is increased from $(k-1)/k|S|$ to $|S|$.

In summary, the semi-supervised CV bound addresses both problems of the CV estimate: the variance of the random variables $\hat{e}(f_i, S_i)$ is explicitly taken into account by using the CV bound for f_{rand} instead of the CV estimate, and the fact that we are interested in the performance of f and not in the average performance of the f_is is taken care of by the second term. The unavoidable draw-back is that the upper bound of Theorem 1 is no longer an unbiased estimate of anything interesting and contains some slack (due to neglecting the dependences between the f_is and the use of triangle inequality in the derandomization). Thus, whether to use the CV estimate or the bound of Theorem 1 depends on whether one prefers unreliable but unbiased estimates or reliable but therefore biased guarantees.

Using the bound of Theorem 1 for model selection is straightforward: Just plug it into the SRM framework. What results is a new semi-supervised model selection criterion that can also be viewed as a semi-supervised learning algorithm that actively uses unlabeled data in choosing a classifier.

4 Experiments

We compared the proposed semi-supervised model selection method to the CV heuristic on a model selection problem borrowed from [6]. The task is to find the best combination of the cost parameter C and the kernel parameter γ for a Gaussian kernel support vector machine (`libsvm`). The parameter space is discretized to the grid defined by $C = 2^4, 2^3, \ldots, 2^{-10}$ and $\gamma = 2^{12}, 2^{11}, \ldots, 2^{-2}$. Thus, the number of candidate models is $N = 225$. As datasets we used the `dna`, `letter`, `satimage`, `segment`, `shuttle`, and `usps` benchmarks (obtained from [7]) that were transformed into semi-supervised datasets by forgetting 10% of the labels. The observed error rate of the classifiers is their error rate on the "unlabeled" data. The bounds were computed by `semibound`, a package available at `http://hunch.net/semibound`.

Figure 4 presents the results of the experiments over the whole model space, whereas Figure 4 shows the 2d restrictions of the 3d plots to the value of C on which the minimal error rate is attained. It can be seen that the CV estimate is always very close to the level of the observed error rate (being sometimes below and sometimes above it), while the bound is always clearly above the observed error rate. Thus, as should be expected, the bound is no good as an estimate of the error rate. However, the gap between the bound and the observed error rate is most of the time so small that the error guarantee provided by the bound is non-trivial and potentially useful. This is quite good a result given the complex multi-class nature of the learning problems: All bounds in our knowledge that require no labeled data to be held out for testing would have failed to drop below 1 (the level of triviality), at least on the multi-class problems on which margin based bounds are not directly applicable.

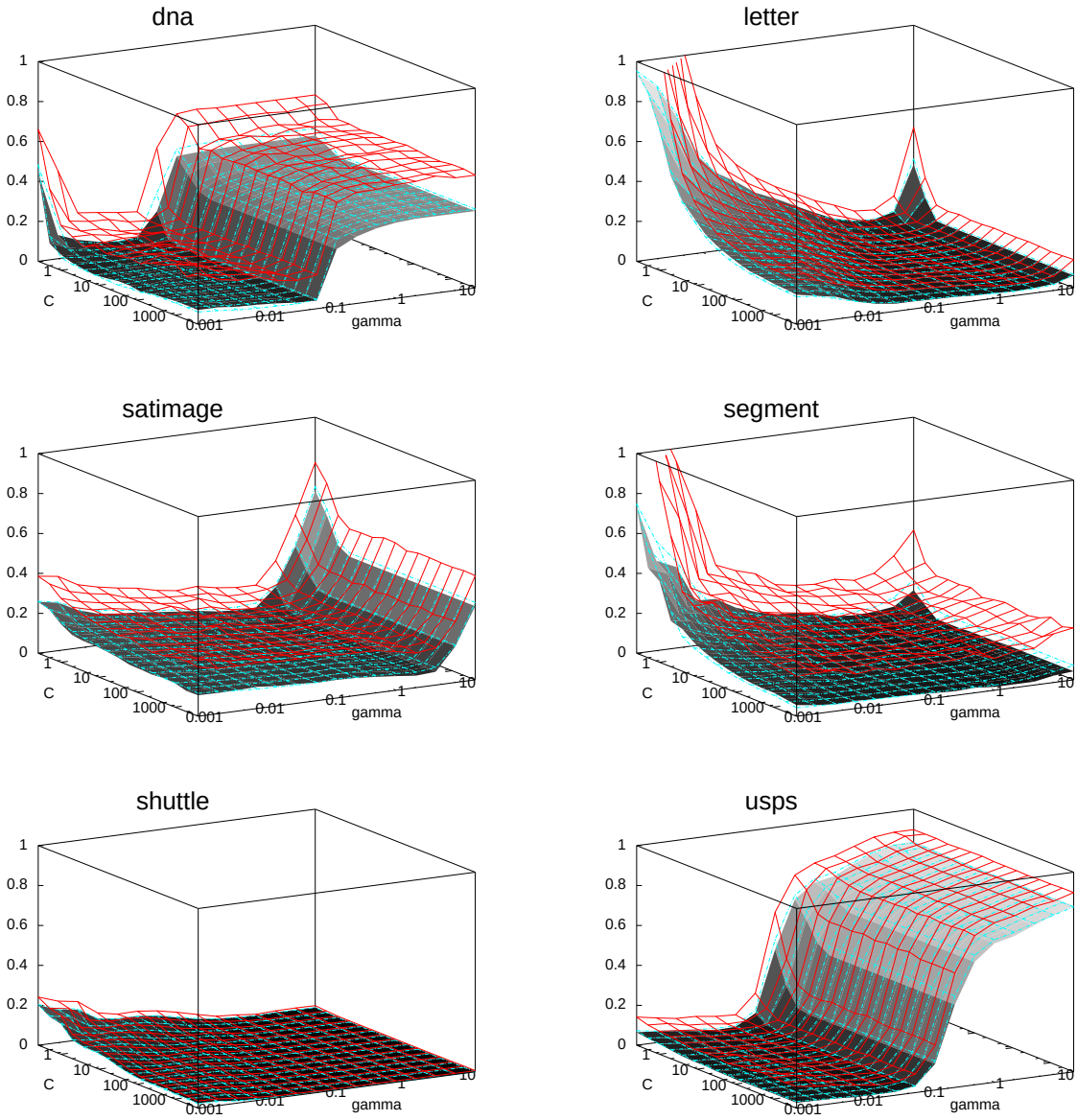

Figure 1: Plots of the observed error rate (gray surface), the CV estimate (cyan mesh), and the semi-supervised CV bound (red mesh). In all experiments, we have $\delta = 0.01$ for model selection, resulting in a share of $0.01/225 < 10^{-4}$ for each of the individual models.

For model selection, though, shape is more important than size. By visual inspection, the shape of the proposed bound seems to reflect the shape of the observed error very closely. Although the gap between the bound and the observed error rate is not constant, it seems to be monotonically increasing with the observed error rate (modulo small random fluctuations). Thus, minimizing the bound should result in about the same model as minimizing the error rate on test set. This is indeed the case: On three of the six datasets we experimented with, the semi-supervised model

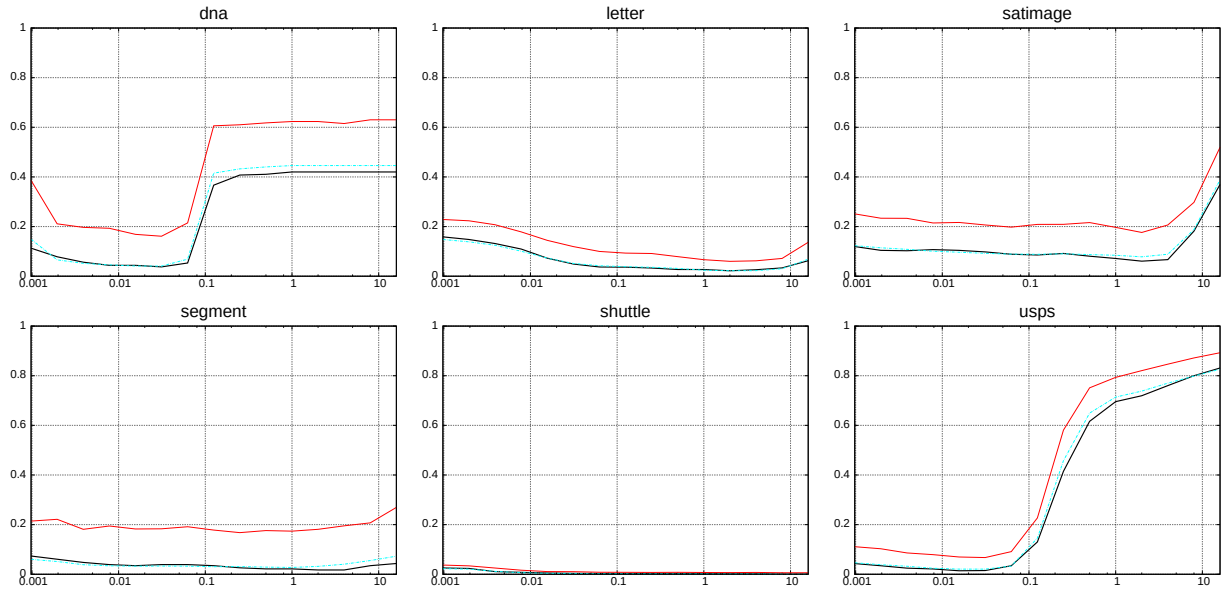

Figure 2: Cross sections of plots in Figure 1 for the value of C achieving smallest error rate.

selection method and the CV estimate choose the same model. On the datasets on which the methods choose different models, the observed error rate (again on the "unlabeled" data) of the model chosen by the semi-supervised method is once better, once worse, and once exactly equal to the observed error rate of the model chosen by the CV heuristic. Thus, the performance of the two methods seems to be roughly the same.

It is clear from the experiments that, at least on these benchmarks, the CV estimate is so close to the observed error rate and thus to the generalization error that the CV model selection heuristic is close to optimal. Thus, the result that the semi-supervised method did as well as the CV heuristic can be considered good.

5 Consistency of semi-supervised cross-validation model selection

In this section we give conditions under which the semi-supervised CV model selection method converges to an optimal solution when the number of learning examples increases without limit. The results are asymptotic and build on reasonable but hard to verify assumptions beyond iid, so they provide little to no information directly relevant in practice. Nevertheless, as asymptotic results in general, they may complement the experimental justification for the method and give some understanding on why and when it is at its best.

Before going into the exact formulations of the assumptions we make toward proving the consistency result, let us begin with the following lemma. Its proof is only a sketch, as are all the proofs in this section. In all that follows we assume that m goes monotonically to infinity with n.

Lemma 1 *For any $\varepsilon > 0$, the probability that the CV bound undershoots the generalization error by more than ε goes to zero as $n \to \infty$.*

Proof. It is always true that $\epsilon(f) \leq \epsilon(f_{\text{rand}}) + d(f, f_{\text{rand}})$. By applying the law of large numbers and the Chernoff approximation for the inverse binomial tail, one can show that the terms of the

semi-supervised CV bound converge to the terms on the right hand side of the above inequality. Thus, the claim follows. □

The above lemma shows that the CV bound is never over-optimistic in the limit. To prove the converse, we have to make additional assumptions, as the CV bound will fail to converge to the generalization error unless the learning algorithm it is applied to is stable in a suitable sense. Instead of assuming the stability directly, we show that it is implied by the following two easier to interpret conditions. In the conditions, the Bayes classifier f_{bayes} is given by $f_{\text{bayes}}(x) = \arg\max_{y \in \mathcal{Y}} P(Y = y|X = x)$ (ties broken arbitrarily). It is the classifier with minimal generalization error.

Consistency [8] There exists a candidate model that is consistent with the learning problem at hand: as $n \to \infty$, the generalization error of the classifiers related to at least one of the candidate models converges to $\epsilon(f_{\text{bayes}})$ in probability.

Noise condition There exists some $c > 0$ such that for all $x \in \mathcal{X}$ and all $y \neq f_{\text{bayes}}(x)$, $P(Y = f_{\text{bayes}}(x)|X = x) \geq P(Y = y|X = x) + c$.

The consistency condition ensures that there is a "correct" model to be found. The noise condition can be seen as a (multi-class) version of a limiting case of Tsybakov's noise condition [9]. We need it to guarantee that the Bayes classifier is identifiable in the sense that convergence of generalization error to the minimum implies convergence to f_{bayes} (the latter with respect to d). The condition says essentially that the optimal classifier f_{bayes} has a uniform advantage of c over randomly guessing a label, whence the name noise condition. The actual value of c is immaterial and does not need to be known.

Lemma 2 *Suppose the above two assumptions are true. Let f_n be the classifier chosen by (an arbitrary one of) the consistent candidate models given n examples. Then $d(f_n, f_{bayes}) \to 0$ in probability as $n \to \infty$.*

Proof. By the noise condition, $\epsilon(f_n) \geq \epsilon(f_{\text{bayes}}) + cd(f_n, f_{\text{bayes}})$, where $c > 0$ is the constant from the noise condition. Thus, as $\epsilon(f_n) \to \epsilon(f_{\text{bayes}})$ in probability, $d(f_n, f_{\text{bayes}}) \to 0$ in probability as well. □

Note that if we only assume consistency, no analog of Lemma 2 can be true as the sequence (f_n) does not need to converge to anything.

Lemma 3 *Assuming the above conditions, the semi-supervised CV bounds for consistent models converge to $\epsilon(f_{bayes})$ in probability.*

Proof. The idea is to show that the first term of the bound converges to $\epsilon(f_{\text{bayes}})$ and the second term to 0. The first follows easily and the second by Lemma 2. □

Combining all of the above, we get the consistency theorem.

Theorem 2 *Assuming the above conditions, the generalization error of the classifier chosen by the semi-supervised model selection method converges to $\epsilon(f_{bayes})$ in probability.*

Proof. By Lemmas 1 and 3 we know that the minimum of the bounds for the candidate models converges to $\epsilon(f_{\text{bayes}})$, and again by Lemma 1 so does the generalization error of the selected classifier. □

We note that it is easy to see that the above consistency result is quite robust to small violations of the above conditions: if the conditions are almost true, then so are the conclusions.

6 Conclusions

The goal of model selection is to find the model with best generalization performance. Both the CV heuristic and the methods based on SRM (including the method proposed in this paper) do something different: The CV heuristic makes its choice based on unbiased but potentially unreliable estimates of the error of a wrong classifier, while the SRM alternatives suffer from the implications of optimizing generalization error bounds instead of the error itself. In case of the proposed method we hope the latter implications are not that harmful as the underlying CV bound is rather tight.

According to our initial experiments, it seems that by using the semi-supervised method, one gets about the same model selection result as one would have obtained using the CV heuristic. In addition, one gets an error guarantee for the chosen classifier. The guarantee does not come entirely for free, since the semi-supervised method can be used only if unlabeled data is available. Whether the value of the obtained guarantee over-weighs the cost of the unlabeled sample depends on the application, but we suspect that this is not unlikely. In such situations where error guarantees are desired, the proposed semi-supervised method may have great potential.

References

[1] Dale Schuurmans and Finnegan Southey. Metric-based methods for adaptive model selection and regularization. *Machine Learning*, 42(1–3):51–84, 2002.

[2] Yoshua Bengio and Nicolas Chapados. Extensions to metric-based model selection. *Journal of Machine Learning Research*, 3(7):1209–1227, 2003.

[3] Vladimir N. Vapnik. *Statistical Learning Theory*. John Wiley and Sons, 1998.

[4] Matti Kääriäinen and John Langford. A comparison of tight generalization error bounds. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

[5] Matti Kääriäinen. Generalization error bounds using unlabeled data. In *Learning Theory: 18th Annual Conference on Learning Theory, COLT 2005*, pages 127–142. Springer, 2005.

[6] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13:415–425, 2002.

[7] Chih-Chung Chang and Chih-Jen Lin. The libsvm page, 2005. `http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html`.

[8] Luc Devroye, L'azl'o Györfi, and G'abor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31 of *Applications of Mathematics*. Springer, Berlin Heidelberg New York, 1996.

[9] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.