
Refining the Wrapper Approach

– Smoothed Error Estimates for Feature Selection

Loo-Nin Teow

Haifeng Liu

DSO National Laboratories, 20 Science Park Drive, S 118230, Singapore

TLOONIN@DSO.ORG.SG

LHAIFENG@DSO.ORG.SG

Hwee Tou Ng

School of Computing, National University of Singapore, 3 Science Drive 2, S 117543, Singapore

NGHT@COMP.NUS.EDU.SG

Eric Yap

Defence Medical Research Institute, #20-04, Defence Technology Tower B, Depot Road, S 109678, Singapore

NMIV3@NUS.EDU.SG

Abstract

In the wrapper approach for feature selection, a popular criterion used is the leave-one-out estimate of the classification error. While being relatively unbiased, the leave-one-out error estimate is nonetheless known to exhibit a large variance, which can be detrimental especially for small samples. We propose reducing its variance (i.e. smoothing) at two levels. At the first level, we smooth the error count using estimates of posterior probabilities; while at the second level, we smooth the posterior probability estimates themselves using Bayesian estimation with conjugate priors. Furthermore, we propose using the jackknife to reduce the bias inherent in Bayesian estimators. We then show empirically that smoothing the error estimate gives improved performance in feature selection.

1. Introduction

The problem of automated feature selection has been well explored (Blum & Langley, 1997; Jain & Zonker, 1997; John, Kohavi, & Pfleger, 1994; Kohavi & John, 1997; Koller & Sahami, 1996; Kudo & Sklansky, 2000; Langley & Sage, 1994; Langley & Sage, 1997; Liu, Motoda, & Dash, 1998; Pudil, Novovicova, & Kittler, 1994; Ruck, Rogers, & Kabrisky, 1989; Yang & Pedersen, 1997). In the case of pattern recognition, the objective of feature selection is to select those features that best discriminate between the classes. In this paper, we define relevant features as those that contribute to the discrimination, while irrelevant features are those that do not.

Feature selection techniques fall broadly into two categories: filter and wrapper methods (John, Kohavi, and Pfleger, 1994). In filter methods, statistical tests, such as information entropy or class separability, are applied to various subsets of features so as to identify the optimal subset. The objective of filter methods is to predict as intelligently as possible, without actually invoking any classifiers, which subset of features would give the best classification accuracy. In wrapper methods, on the other hand, a classifier is induced for a given subset of features, and an estimate of the classification error is used to evaluate the quality of the subset. In other words, a classifier is invoked directly and the optimal feature subset is chosen based on its classification error rate.

Wrapper methods typically perform better (i.e. they find better features) although they are more computationally expensive. While there are arguments for and against either method, we believe the best way to measure the discriminative ability of a set of features is to estimate the classification error directly with a classifier.

Clearly, a key issue in the wrapper approach is estimating the error rate in a robust manner. We want an estimate that has both a small bias and a small variance. We may choose the leave-one-out estimate of the classification error rate due to the low bias of this type of estimate, and find the smallest subset of features such that the estimate is minimized. Although the leave-one-out error estimate is relatively unbiased, it has a large variance (Efron, 1983; Kohavi, 1995). In other words, it fluctuates a lot with respect to small differences in the data. This may lead to the error estimate becoming inconsistent and unreliable, especially when the number of examples is small.

We found that we can reduce the estimation variance considerably by smoothing the error rate at two levels – smoothing the error count using estimates of posterior probabilities (Glick, 1978; Pawlak, 1988; Tutz, 1985), and smoothing the posterior probability estimates themselves using Bayesian estimation with conjugate priors (Box & Tiao, 1973; Jeffreys, 1946; Laplace, 1951; Lidstone, 1920; Perks, 1947). Furthermore, we propose using the jackknife estimator (Miller, 1974; Quenouille, 1956) to reduce the bias inherent in Bayesian estimates (Bickel & Mallows, 1988; Noorbaloochi & Meeden, 1983). While these techniques by themselves are not new, no one to our knowledge has ever applied them in the context of feature selection. All these techniques combine to give a more robust error estimate, which in turn leads to better performance in feature selection, as we shall show empirically.

2. The Classifier

For this work, we chose to employ the nearest neighbor classifier, which is also a popular choice for wrapper methods (Blum & Langley, 1997; Kudo & Sklansky, 2000; Langley & Sage, 1994; Langley & Sage, 1997). The nearest neighbor classifier is very simple to induce (zero training time), very amenable to analysis, has no learning parameters to adjust, can be modified easily to handle conflicting examples, and gives reasonably good performance. More importantly, it has been shown (Langley & Iba, 1993) that the error rate for the nearest neighbor classifier increases with the number of irrelevant features for a fixed number of examples. In other words, its performance is sensitive to the presence of irrelevant features. It has also been proven (Cover & Hart, 1967) that given enough examples, the nearest neighbor error rate is at most twice the Bayesian error rate. Finally, the nearest neighbor classifier is not constrained by any non-linearity in the interclass boundary, i.e. it is able to learn even when the class is determined by a complex nonlinear combination of features (unlike say, the naïve Bayes or linear perceptron classifiers).

3. Mathematical Framework

3.1 Basic Model

We have a set of n examples, $S = \{(\mathbf{x}_i, y_i) \mid i = 1 \dots n\}$, where \mathbf{x}_i is a vector containing d nominal features, and $y_i \in \{1 \dots m\}$ is a class label given m classes. Each feature $\mathbf{x}_i^{(k)}$, where $k = 1 \dots d$, can have one of several different values, possibly missing. If the class conditional (posterior) probability distributions are known, then the Bayesian *maximum a posteriori* (MAP) classifier is

$$g^*(\mathbf{x}) = \operatorname{argmax}_{h \in \{1 \dots m\}} \{P(h \mid \mathbf{x})\} \quad [1]$$

The corresponding Bayesian error probability is then

$$L^* = P(g^*(\mathbf{x}) \neq y) = P\left(\operatorname{argmax}_{h \in \{1 \dots m\}} \{P(h \mid \mathbf{x})\} \neq y\right) \quad [2]$$

The Bayesian classifier is optimal, i.e. its error probability is smaller than that of any other classifier (a simple proof of this for 2 classes can be found in Devroye, Györfi, and Lugosi (1996); it is easily extensible to multiple classes). However, the computation of the Bayesian error probability is a very complex and difficult problem, except in special cases when all the probability distributions are already known. Hence, we estimate the error probability from the empirical error rate:

$$L_{emp} = E(g(\mathbf{x}) \neq y) = \frac{1}{n} \sum_i \delta(g(\mathbf{x}_i) \neq y_i) \quad [3]$$

where $g(\mathbf{x})$ is the class decision output of a chosen classifier given \mathbf{x} as the input, and

$$\delta(Q) = \begin{cases} 1 & \text{if proposition } Q \text{ is true;} \\ 0 & \text{otherwise.} \end{cases} \quad [4]$$

Unfortunately, the above estimate is equivalent to the resubstitution rate (or apparent error rate), which is known to be very optimistically biased, particularly for the nearest neighbor classifier. To obtain a relatively unbiased estimate of the error rate, we use the leave-one-out method:

$$L_{emp}^{leave-one-out} = \frac{1}{n} \sum_i \delta(g_{-i}(\mathbf{x}_i) \neq y_i) \quad [5]$$

where g_{-i} is the classifier's decision function based on the data set with (\mathbf{x}_i, y_i) removed.

In the case of the nearest neighbor classifier,

$$g_{-i}^{<NN>} = \operatorname{argmax}_h \left| \{(\mathbf{x}_j, y_j) \in \Lambda_{-i}^{<NN>} \mid y_j = h\} \right| \quad [6]$$

$\Lambda_{-i}^{<NN>}$ is the set of examples nearest to \mathbf{x}_i , and can include examples with the same feature values as \mathbf{x}_i , i.e.

$$\Lambda_{-i}^{<NN>} = \{(\mathbf{x}_j, y_j) \in S \mid j \neq i, D(\mathbf{x}_i, \mathbf{x}_j) \leq D(\mathbf{x}_i, \mathbf{x}_k) \forall k \in \{1 \dots n\}\} \quad [7]$$

where $D(\mathbf{x}_i, \mathbf{x}_j)$ is the distance between \mathbf{x}_i and \mathbf{x}_j .

In this paper, we focus on the nominal domain as a preliminary study. For nominal feature vectors, we have found the following distance measure to perform well:

$$D(\mathbf{x}_i, \mathbf{x}_j) = \sum_k^d \left\| \mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)} \right\|^2 \quad [8]$$

where $\left\| \mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)} \right\|$

$$= \begin{cases} 0 & \text{if } \mathbf{x}_i^{(k)} = \mathbf{x}_j^{(k)} \text{ or if both values are missing.} \\ \frac{1}{2} & \text{if only one of the values is present.} \\ 1 & \text{if both values are present and } \mathbf{x}_i^{(k)} \neq \mathbf{x}_j^{(k)}. \end{cases} \quad [9]$$

3.2 Smoothing the Error Rate

As mentioned earlier, the main objection against using the leave-one-out estimate is its large variance. To overcome this problem, we can smooth $L_{emp}^{leave-one-out}$ as follows:

$$L_{emp}^{smooth} = \frac{1}{n} \sum_i^n (1 - P_{-i}(y_i | \mathbf{x}_i)) \quad [10]$$

where $P_{-i}(y_i | \mathbf{x}_i)$ is the posterior probability estimated from the data with (\mathbf{x}_i, y_i) removed. Smoothing the classification error estimate was first proposed by Glick (1978) as a means of reducing the estimate's variance. This approach has been further explored in Devroye, Györfi, and Lugosi (1996), Pawlak (1988), Tutz (1985).

$P_{-i}(y_i | \mathbf{x}_i)$ can be estimated as a parameter of the multinomial distribution. First, we define a symmetric Dirichlet distribution as a Dirichlet distribution with all its parameters $\alpha_1, \dots, \alpha_m$ set equal to α . Then, taking a symmetric Dirichlet distribution as the conjugate prior to the multinomial distribution, a Bayesian estimate of $P_{-i}(y_i | \mathbf{x}_i)$ is

$$\hat{P}_{-i}(y_i | \mathbf{x}_i) = \frac{l_i + \alpha}{L_i + m\alpha} \quad [11]$$

where α is the hyperparameter of the conjugate prior,

$$l_i = \left| \left\{ (\mathbf{x}_j, y_j) \in \Lambda_{-i}^{<NN>} \mid y_j = y_i \right\} \right| \quad [12]$$

$$L_i = \left| \Lambda_{-i}^{<NN>} \right| \quad [13]$$

The above formulation neatly handles the case of conflicting examples as well as avoids the need for breaking distance ties, which can be rather problematic especially for nominal-valued domains.

α can also be regarded as the smoothing parameter. Using arbitrary α corresponds to Lidstone's law (Lidstone, 1920). Setting $\alpha = 0$ gives the maximum likelihood estimator, while $\alpha = 1$ corresponds to Laplace's law of succession (Laplace, 1951). $\alpha = 0.5$ is a popular choice that corresponds to Jeffreys-Perks' law or

Expected Likelihood Estimation (Box & Tiao, 1973; Jeffreys, 1946; Perks, 1947). More recently, Kohavi, Becker, and Sommerfield (1997) showed that setting $\alpha = 1/L_i$ is a reasonably good heuristic for reducing the bias in Laplace's law of succession. Other approaches to smoothing the probability estimator, tailored for sparsely represented domains such as natural language processing, can be found in Chen and Goodman (1996), Friedman and Singer (1999), Ristad (1995).

Simple algebraic manipulation shows that Equation 11 can be rewritten as a linear combination:

$$\begin{aligned} \hat{P}_{-i}(y_i | \mathbf{x}_i) &= \frac{L_i}{L_i + m\alpha} \cdot \frac{l_i}{L_i} + \frac{m\alpha}{L_i + m\alpha} \cdot \frac{1}{m} \\ &= \left(1 - \frac{m\alpha}{L_i + m\alpha} \right) \cdot \frac{l_i}{L_i} + \frac{m\alpha}{L_i + m\alpha} \cdot \frac{1}{m} \end{aligned} \quad [14]$$

3.3 Jackknifing the Bayesian Estimator

It is known that Bayes estimators are generally biased except in very special cases (Bickel & Mallows, 1988; Noorbaloochi & Meeden 1983). Hence, we propose using a statistical method to reduce the bias in the Bayesian estimator of the posterior probabilities. The jackknife technique introduced by Quenouille (1956) and reviewed by Miller (1974) has been proven effective in reducing the bias in many estimators.

Let $\hat{\theta}$ be an estimator of the parameter θ based on L examples. Also let $\hat{\theta}_{-j}$ be the corresponding estimator based on $L-1$ examples with the j^{th} example removed. Then the jackknife estimator $\tilde{\theta}$ is defined as

$$\tilde{\theta} = L\hat{\theta} - \frac{L-1}{L} \sum_j^L \hat{\theta}_{-j} \quad [15]$$

Applying this to the posterior probability estimator, we get (proof not shown):

$$\begin{aligned} \tilde{P}_{-i}(y_i | \mathbf{x}_i) &= L_i \hat{P}_{-i}(y_i | \mathbf{x}_i) - \frac{L_i - 1}{L_i} \sum_j^{L_i} \hat{P}_{-i-j}(y_i | \mathbf{x}_i) \\ &= L_i \frac{l_i + \alpha}{L_i + m\alpha} \\ &\quad - \frac{L_i - 1}{L_i} \left(l_i \frac{l_i - 1 + \alpha}{L_i - 1 + m\alpha} + (L_i - l_i) \frac{l_i + \alpha}{L_i - 1 + m\alpha} \right) \end{aligned} \quad [16]$$

Again, the above can be expressed as a linear combination, although the combination weights here are more complex:

$$\begin{aligned} \tilde{P}_{-i}(y_i | \mathbf{x}_i) = & \left(1 - \frac{m^2 \alpha^2}{(L_i + m\alpha)(L_i - 1 + m\alpha)} \right) \cdot \frac{l_i}{L_i} \\ & + \frac{m^2 \alpha^2}{(L_i + m\alpha)(L_i - 1 + m\alpha)} \cdot \frac{1}{m} \end{aligned} \quad [17]$$

We may also employ bootstrap methods (Efron, 1979), whereby estimates are averaged over many “bootstrap samples” each drawn with replacement from the original sample. However, this would be too computationally expensive, especially in the context of feature selection whereby the search space can be very large.

4. Search Engine

When the number of features is large (say ≥ 20), an exhaustive search through all possible subsets of features for the optimal one would be impractical. We need to employ a more intelligent strategy for searching through the solution space. Comparison of a number of existing search algorithms (Jain & Zonker, 1997; Kudo & Sklansky, 2000) seem to favor the sequential forward/backward floating search methods (SFFS/SBFS) (Pudil, Novovicova, & Kittler, 1994) as well as the genetic algorithm (GA) (Holland, 1975).

The SFFS/SBFS algorithm adds or deletes one feature at a time but backtracks (i.e. delete a selected feature or add a deleted feature) whenever it can find an improvement in the criterion. The genetic algorithm, on the other hand, is a stochastic search algorithm that mimics the evolutionary process to find the solution.

As the genetic algorithm requires fine-tuning of many parameters, we chose to employ a combined SFFS/SBFS method as the search engine, whereby the better solution between those generated by SFFS and SBFS is chosen. The quality of a solution is based on its error estimate and the number of features chosen. A good solution would have low error and few features.

5. Experiments

5.1 Bias-Variance Decomposition

It is well known in statistical machine learning that the mean squared error (MSE) can be decomposed into the variance and the square of the bias:

$$\begin{aligned} MSE &= E \left[(\hat{L} - L^*)^2 \right] \\ &= E \left[(\hat{L} - E[\hat{L}])^2 \right] + (E[\hat{L}] - L^*)^2 \\ &= \text{Variance} + \text{Bias}^2 \end{aligned} \quad [18]$$

where L^* is the Bayesian error probability, \hat{L} is an estimate of the error, and the expectation is taken over all possible data sets of a given size.

In this subsection, we conduct simulation studies of the behavior of the mean squared error, variance and bias of each type of error estimate discussed in this paper.

5.1.1 EXPERIMENTAL SETUP

For a given sample of size n , we generate random vectors of 3 nominal features. Each feature is randomly assigned one of 3 values with equal probabilities. We fix the number of classes at 3, and randomly assign a class label to each feature vector according to a predefined set of *a priori* class probabilities. The class probabilities are in turn generated randomly from a uniform distribution. Hence, we can take the conditional probabilities to be equal to the corresponding *a priori* class probabilities. This implies that we can directly compute the Bayesian error as:

$$L^* = P \left(\operatorname{argmax}_{h \in \{1 \dots m\}} \{P(h)\} \neq y \right) = 1 - \max_{h \in \{1 \dots m\}} \{P(h)\} \quad [19]$$

We conduct 1000 trials of 1000 simulations each. In each trial, a set of class probabilities is generated and 1000 simulations are conducted using these class probabilities. In each simulation, we generate data sets of various sizes. We perform two analyses: small-sample analysis with data set sizes ranging from 5 to 50 at increments of 5, and large-sample analysis with data set sizes ranging from 50 to 500 at increments of 50. The analyses are performed on 6 types of error estimates: the unsmoothed leave-one-out error count, the smoothed error estimates with $\alpha = 0$ (maximum likelihood estimator), $\alpha = 0.5$ (Jeffreys-Perks’ law), $\alpha = 1$ (Laplace’s law of succession), and the jackknifed versions of Jeffreys-Perks’ law and Laplace’s law of succession. (The jackknife estimator for the maximum likelihood estimate in this case happens to be the maximum likelihood estimator itself.) For each data set size and each type of error estimate, we compute the mean squared error, variance and bias over 1000 simulations in a trial. These statistics are then averaged over the 1000 trials. The results for the small-sample analysis are illustrated in Figure 1, while the results for the large-sample analysis are illustrated in Figure 2.

5.1.2 DETAILED ANALYSIS

As these figures show, the unsmoothed error estimate exhibits the largest variance and the smallest bias, while the smoothed error estimate using Laplace’s law of succession for the posterior probabilities has the smallest variance and the largest bias. There is clearly a trade-off between bias and variance as α varies. The jackknife lowers the bias for both Jeffreys-Perks’ law and Laplace’s law of succession, as expected.

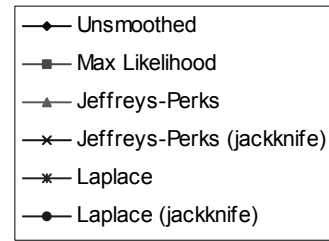
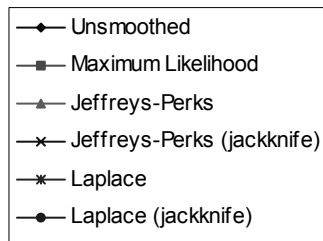
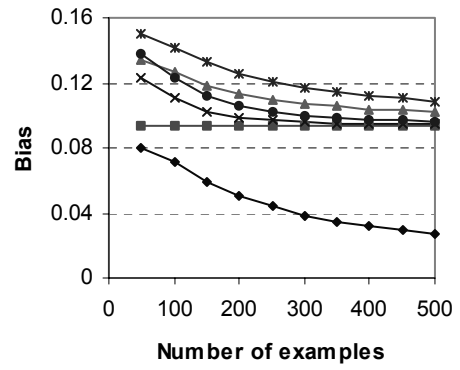
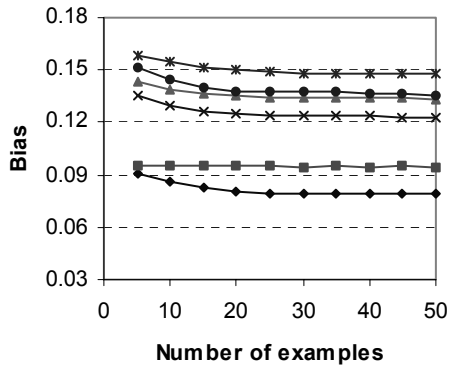
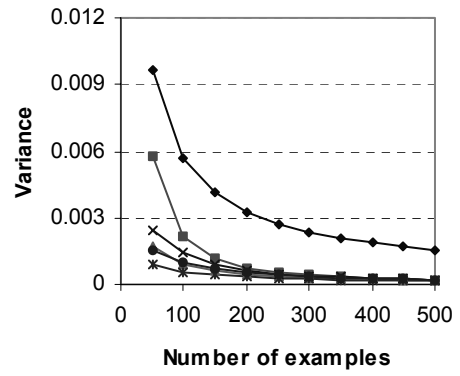
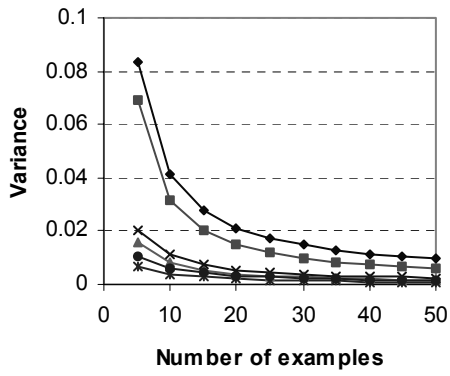
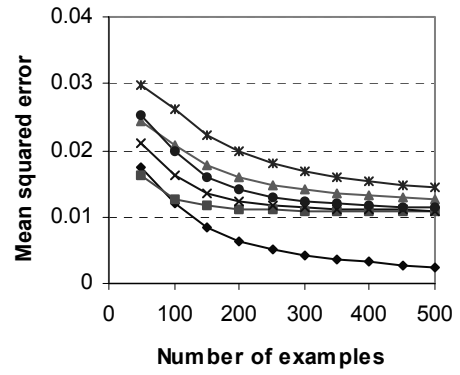
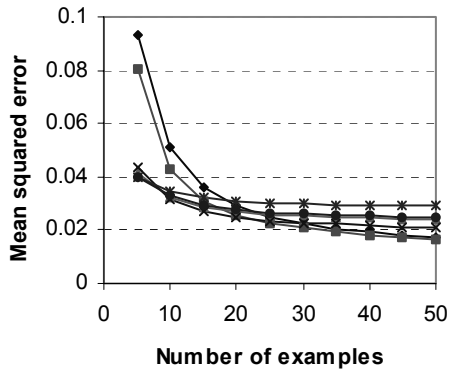


Figure 1. Small-sample analysis of the various types of classification error estimates with respect to their mean squared errors, variances and biases.

Figure 2. Large-sample analysis of the various types of classification error estimates with respect to their mean squared errors, variances and biases.

In every case, the mean squared error and the variance drop quickly as the number of examples increases. As the number of examples grows, the mean squared errors for low-bias/high-variance estimates drop more rapidly than for those high-bias/low-variance estimates. For example, the unsmoothed error estimate has the largest mean squared error when the number of examples is 5, but it has the smallest mean squared error when the number of examples is 100 or more. Conversely, the smoothed error estimate using Laplace’s law of succession has the smallest mean squared error when the number of examples is 5, but it has the largest mean squared error when the number of examples is 100 or more.

Interestingly, the bias of the error estimate using the maximum likelihood estimator remains relatively constant regardless of the number of examples, while its mean squared error appears to converge towards a constant (larger than that of the unsmoothed estimate). As the number of examples increases, both the bias and the mean squared error of the other smoothed estimates converges to the maximum likelihood case, while those of the unsmoothed estimate drop considerably lower. At the same time, the bias and the mean squared error for the jackknifed estimators drop as a faster rate than those of the non-jackknifed ones.

On the other hand, the variance of the unsmoothed estimate remains relatively higher than the smoothed estimates even when the number of examples is large, while the variances of all the smoothed estimates drop to almost zero as the number of examples increases. We can thus infer that the mean squared error in the smoothed estimate is due mainly to its inherent bias when the number of examples is large.

In the context of feature selection, the variance is a more important factor than the bias, especially when the number of examples is small, since we are more concerned with *relative* error estimates rather their absolute values when comparing different feature subsets. The above studies suggest that smoothed error estimates are better suited for small samples.

5.2 Feature Selection

We now show empirically that in the wrapper approach for feature selection, using a smoothed error estimate can give better performance than using the unsmoothed error count. We conduct experiments on both artificial and real-world data. The artificial data sets are Corral (John, Kohavi, & Pfleger, 1994) and LED-24 from the UCI repository (Blake & Merz, 1998). The real-world data sets are Voting and SPECT-Heart, both from the UCI repository. The features (i.e. attributes) in these data sets are all nominal, and their properties are given in Table 1.

Only the training and testing examples for the SPECT-Heart data has been given explicitly. For the rest of the data sets (i.e. Corral, LED-24 and Voting), we arbitrarily fix the number of training and testing examples, and conduct 10 simulations whereby we randomly partition the data set into the training and testing examples; the results are then averaged over the 10 simulations.

We perform feature selection on the training set and test the selected features on the testing set to give the percentage of classification errors. The results are given in Table 2 and Table 3. Using the smoothed error estimate with the jackknifed Jeffreys-Perks estimator gives the best overall performance on the test data. Note that the test error rates using all the original features are typically very high because the nearest neighbor classifier is particularly sensitive to irrelevant features.

Table 1. The UCI data sets used and their properties.

Data Set	# Classes	# Features	# Train	# Test
Corral	2	6	32	128
LED-24	10	24	200	3000
Voting	2	16	300	135
SPECT-Heart	2	22	80	167

Table 2. Error rates on the test data for the features selected using various criteria. For each data set, the lowest test error rate obtained is boldfaced.

Error Estimate	Corral	LED-24	Voting	SPECT-Heart
Unsmoothed	5.55 ± 9.31	34.18 ± 6.26	7.19 ± 0.67	33.16
Max Likelihood	0.00 ± 0.00	37.88 ± 0.26	6.67 ± 0.00	29.41
Jeffreys-Perks	0.00 ± 0.00	32.10 ± 0.03	6.74 ± 0.22	29.41
Jeffreys-Perks (jackknife)	0.00 ± 0.00	32.10 ± 0.03	5.78 ± 0.86	29.41
Laplace	11.33 ± 9.25	32.10 ± 0.03	6.74 ± 0.22	29.41
Laplace (jackknife)	5.70 ± 8.71	32.10 ± 0.03	6.30 ± 0.76	31.55
[All features]	10.16 ± 1.10	60.76 ± 0.28	9.78 ± 0.44	41.71

Table 3. Number of features selected for the various criteria.

Error Estimate	Corral	LED-24	Voting	SPECT-Heart
Unsmoothed	3.80 ± 0.75	7.30 ± 0.46	3.80 ± 0.60	9
Max Likelihood	4.00 ± 0.00	10.00 ± 0.00	5.00 ± 0.00	5
Jeffreys-Perks	4.00 ± 0.00	5.00 ± 0.00	3.00 ± 0.00	5
Jeffreys-Perks (jackknife)	4.00 ± 0.00	5.00 ± 0.00	4.00 ± 0.45	6
Laplace	2.80 ± 0.98	5.00 ± 0.00	2.10 ± 0.30	6
Laplace (jackknife)	3.40 ± 0.92	5.00 ± 0.00	3.00 ± 0.00	6

6. Further Discussion

Instead of using a symmetric Dirichlet distribution, we can set each hyperparameter of the Dirichlet distribution to be proportional to the class prior probabilities, i.e. $\alpha_h = \lambda P(h) \forall h = 1 \dots m$, where λ is a scaling constant. Each class probability can in turn be estimated from the entire data set using a symmetric Dirichlet distribution as the conjugate prior. This is similar to the hierarchical model proposed by Mackay and Peto (1995). We then get the following Bayesian estimates:

$$\hat{P}_{-i}(y_i | \mathbf{x}_i) = \frac{I_i + \lambda \hat{P}(y_i)}{L_i + \lambda} \quad [20]$$

$$\hat{P}(y_i) = \frac{\eta(y_i) + \beta}{n + m\beta} \quad [21]$$

where β is the hyperparameter of the symmetric Dirichlet distribution for the class prior, and

$$\eta(y_i) = \left| \{(\mathbf{x}_j, y_j) \in S \mid y_j = y_i\} \right| \quad [22]$$

We can again express the above as linear combinations:

$$\hat{P}_{-i}(y_i | \mathbf{x}_i) = \left(1 - \frac{\lambda}{L_i + \lambda}\right) \cdot \frac{I_i}{L_i} + \frac{\lambda}{L_i + \lambda} \cdot \hat{P}(y_i) \quad [23]$$

$$\hat{P}(y_i) = \left(1 - \frac{m\beta}{n + m\beta}\right) \cdot \frac{\eta(y_i)}{n} + \frac{m\beta}{n + m\beta} \cdot \frac{1}{m} \quad [24]$$

Like α , we can set $\beta = 0, 0.5$ or 1 . We would also suggest setting $\lambda = m\beta$, so that the combination weights of the two

estimates in Equation 23 and Equation 24 become equivalent in form. Of course, there is no real strong justification for setting the parameters in this way, but it can be argued that the rate of change of the combination weights with respect to the number of examples should be the same regardless of the probability being estimated.

The jackknife can also be applied to the above estimates in a similar fashion. The hierarchical model should be more accurate since it uses additional information from the data set to estimate the class prior probabilities. Further analysis of this model remains for future work.

7. Conclusion

We have proposed a refinement to the wrapper method for feature selection by smoothing the error estimate. We have also showed empirically that using a version of the smoothed error estimate can give improved performance (i.e. better able to find the optimal features) compared with using the unsmoothed error count.

We would need to extend our study to k -nearest neighbors and other distance measures. Generalization to numeric features may be possible either by discretization or by using continuous distributions. Smoothing the error rate is a general approach that should be applicable to other cross-validation methods and other types of classifiers.

References

- Bickel, P.J., & Mallows, C.L. (1988). A note on unbiased Bayes estimates. *American Statistician*, 42, 132-134.
- Blake, C.L., & Merz, C.J. (1998). UCI repository of machine learning databases [http://www.ics.uci.edu/~mllearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.
- Blum, A.L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence* 97, 245-271.
- Box, G.E.P., & Tiao, G.C. (1973). *Bayesian inference in statistical analysis*, Addison-Wesley.
- Chen, S.F., & Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. *Proceedings of the Thirty-fourth Annual Meeting of the Association of Computational Linguistics* (pp.310-318).
- Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13, 21-27.
- Devroye, L., Györfi, L., and Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. New York: Springer-Verlag.

- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7, 1-26. University of Michigan Press.
- Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382), 316-330.
- Friedman, N., & Singer, Y. (1999). Efficient Bayesian parameter estimation in large discrete domains. *Advances in Neural Information Processing Systems 11* (pp.417-423).
- Glick, N. (1978). Additive estimators for probabilities of correct classification. *Pattern Recognition*, 10, 211-222.
- Holland, J. (1975). *Adaptation in natural and artificial systems*. University of Michigan Press.
- Jain, A., & Zonker, D. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2), 153-158.
- Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society A*, 186, 453-461.
- John, G., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. *Proceedings of the Eleventh International Conference on Machine Learning* (pp.121-129).
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (pp.1137-1143).
- Kohavi, R., Becker, B., & Sommerfield, D. (1997). Improving simple Bayes. *Proceedings of the Ninth European Conference on Machine Learning* (pp.78-87).
- Kohavi, R., & John, G.H. (1997). Wrappers for feature subset selection. *Artificial Intelligence (Special Issue on Relevance)*, 97(1-2), 273-324.
- Koller, D., & Sahami, M. (1996). Toward optimal feature selection. *Proceedings of the Thirteenth International Conference on Machine Learning* (pp.284-292).
- Kudo, M., & Sklansky, J. (2000). Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 33, 25-41.
- Langley, P., & Iba, W. (1993). Average-case analysis of a nearest neighbor algorithm. *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence* (pp.889-894).
- Langley, P., & Sage, S. (1994). Oblivious decision trees and abstract cases. *Working Notes of the AAAI'94 Workshop on Case-Based Reasoning* (pp.399-406).
- Langley, P., & Sage, S. (1997). Scaling to domains with irrelevant features. In R. Greiner (Ed.), *Computational Learning Theory and Natural Learning Systems*, 4, 17-30. Cambridge, MA: MIT Press.
- Laplace, P.-S. (1951). *Philosophical essay on probabilities*. New York: Dover Publications. (Original work published in 1814)
- Lidstone, G. (1920). Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities. *Transactions of the Faculty of Actuaries*, 8, 182-192.
- Liu, H., Motoda, H., & Dash, M. (1998). A monotonic measure for optimal feature selection. *Proceedings of the Tenth European Conference on Machine Learning* (pp.101-106).
- Mackay, D.J.C., & Peto, L.C.B. (1995). A hierarchical Dirichlet language model. *Natural Language Engineering*, 1(3), 1-19.
- Miller, R.G. (1974). The jackknife – a review. *Biometrika*, 61(1), 1-15.
- Noorbaloochi, S., & Meeden, G. (1983). Unbiasedness as the dual of being Bayes. *Journal of the American Statistical Association*, 78, 619-623.
- Pawlak, M. (1988). On the asymptotic properties of smoothed estimators of the classification error rate. *Pattern Recognition*, 21(5), 515-524.
- Perks, W. (1947). Some observations on inverse probability, including a new indifference rule. *Transactions of the Faculty of Actuaries*, 73, 285-312.
- Pudil, P., Novovicova, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, 15, 1119-1125.
- Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353-360.
- Ristad, E.S. (1995). *A natural law of succession* (Technical Report CS-TR-895-95). Department of Computer Science, Princeton University.
- Ruck, D.W., Rogers, S.K., & Kabrisky, M. (1989). Feature selection using a multilayer perceptron. *Journal of Neural Network Computing*, 2(2), 40-48.
- Tutz, G.E. (1985). Smoothed additive estimators for non-error rates in multiple discriminant analysis. *Pattern Recognition*, 18(2), 151-159.
- Yang, Y., & Pedersen, J.O. (1997). A comparative study on feature selection in text categorization. *Proceedings of the Fourteenth International Conference on Machine Learning* (pp.412-420).