Machine Learning, 57, 177–195, 2004

A Bias-Variance Analysis of a Real World Learning Problem: The CoIL Challenge 2000

PETER VAN DER PUTTEN putten@liacs.nl
Leiden Institute of Advanced Computer Science, Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands

MAARTEN VAN SOMEREN maarten@swi.psy.uva.nl
Department of Social Science Informatics, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands

Editors: Nada Lavrač, Hiroshi Motoda and Tom Fawcett

Abstract. The CoIL Challenge 2000 data mining competition attracted a wide variety of solutions, both in terms of approaches and performance. The goal of the competition was to predict who would be interested in buying a specific insurance product and to explain why people would buy. Unlike in most other competitions, the majority of participants provided a report describing the path to their solution. In this article we use the framework of bias-variance decomposition of error to analyze what caused the wide range of prediction performance. We characterize the challenge problem to make it comparable to other problems and evaluate why certain methods work or not. We also include an evaluation of the submitted explanations by a marketing expert. We find that variance is the key component of error for this problem. Participants use various strategies in data preparation and model development that reduce variance error, such as feature selection and the use of simple, robust and low variance learners like Naive Bayes. Adding constructed features, modeling with complex, weak bias learners and extensive fine tuning by the participants often increase the variance error.

Keywords: bias-variance decomposition, real world applications, overfitting

1. Introduction

Over the past few years, data mining competitions have become increasingly popular. In one of these, the CoIL Challenge 2000, a prediction problem was used with properties that appear often in real world problems: noisy, skewed, correlated and high dimensional data with a weak relation between input and target. Competitions are organized for various reasons, such as unifying the research community and promoting the field to the outside world. In addition, competitions provide data about how machine learning problems are solved in practice, when the goal is to solve the problem rather than to analyze the performance of a new method.

Extensive published results are available for the CoIL Challenge 2000. 29 out of 43 participants provided a public report on how they solved the problem. We give the results of 2 groups of in total 43 students in this paper. A previous report by van der Putten & van Someren (2000) on an earlier version of this competition covers 6 entries. With the notable exception of the PKDD Discovery Challenge (Berka, 1999), most competitions such as the KDD Cup report only the top entries.

The objective of the competition was to predict who will be interested in a particular insurance product, a caravan policy, and to provide an explanation of why people would be interested. In this paper we focus on the prediction task. Prediction models were used to select potential policy owners from a test set. The performance of the submitted solutions varied over a wide range, from one to two and a half times the number of policy owners that would have been found by a random selection and up to half of the maximum number of policy owners possible.

The main question we address here is what caused this wide range of performance. To explain the results we will evaluate the various approaches using bias-variance decomposition (Geman, Bienenstock, & Doursat, 1992; Friedman, 1997; Kohavi & Wolpert, 1996; Breiman, 1996; James, 2003). This separates the error component resulting from the inability of a learner to represent or find the appropriate model for the behavior from the error component resulting from variance in predictions due to differences in models caused by sampling. Usually, bias-variance analysis is limited to the core modeling step, but here we also apply it to other steps in the knowledge discovery process, such as feature construction and selection.

The paper is structured as follows. First the CoIL Challenge competition, problem tasks and data set are introduced (Section 2). Then we present a general overview of the results for the prediction task (Section 3). Section 4 provides more details on the method we used for analyzing the challenge problem and solutions, including bias-variance decomposition. Sections 5 and 6 focus on steps in the data mining process, data preparation and model development. Section 7 summarizes the expert evaluation of the description task of the competition. Finally we discuss the lessons learned (Section 8).

2. Competition, problem and data description

The CoIL Challenge 2000 was organized by the Computational Intelligence and Learning (CoIL) cluster, a cooperation between four EU funded research networks. The goals of the challenge were to promote the application of computational intelligence and learning technology to real world problems, to clarify the relation between different approaches and to stimulate the search for solutions that combine different methods. The competition ran from March 17 to May 8, 2000 and was organized by the authors of this paper. Only just after the challenge deadline it was decided to publish the submitted solutions (van der Putten & van Someren, 2000).

The objective of the competition was to predict who would be interested in buying a caravan insurance policy, and to give an explanation why people would buy. The problem was selected because it is representative of an important class of real world learning problems: domains with noisy, correlated, redundant and high dimensional data with a weak relation between input and target. Back then this kind of problem was not very well represented in benchmark collections like the UCI Machine Learning Archive. The UCI data sets tended to be more cleaned up and geared towards illustrating the strengths of particular machine learning algorithms rather then representing real world problems. The challenge data is now part of the KDD section of the UCI Archive (Blake & Merz, 1998). The problem was split into a prediction and a description task.

2.1. Prediction task

From a business perspective the goal of the prediction task is to rank current customers of the insurance company according to the probability that they will buy a caravan policy, so that the highest ranking customers can be contacted through a mailing.

Only data about policy ownership is available, so it is assumed that owning this policy from the company is a reasonable approximation to buying the policy in response to a mailing. Given that only 6% actually owns the policy, regular zero-one loss or classification accuracy is not an appropriate evaluation metric. A model that predicts that no one will buy has a high classification accuracy of 94% but is useless for ranking and selecting customers. This illustrates that some methods that are standard in machine learning research are not always directly applicable to real world prediction problems.

From a modeling perspective, the objective of the prediction task is to construct a model that assigns to each customer of the insurance company a probability (or at least a probability rank) that he will buy a caravan policy. If the costs of a mail piece and the profit of a mailing response are known the marketing analyst can then determine the optimal volume of customers to be mailed. However, in practice costs and benefits are not always known and in addition the behavior that is being modeled (ownership) differs from the actual behavior of interest (mail response). So we simplified the business case to the situation where there is a predetermined budget for a mail selection of 20%. The participants had to find the 800 clients in a test set of 4000 instances who were most likely to have a caravan policy. The test set was given to the participants (without the target variable, caravan ownership). The performance metric was the number of correctly identified caravan policy holders among the 800 selected cases. The test set contained 238 policy holders.

Learning methods that construct models which only predict a class but not the probability may not be optimal for this problem, even if other loss functions than zero one loss are used. A classification model may for example classify less than 800 cases as a caravan policy owner. Adding random cases to fill up the selection will not give an optimal solution. Furthermore, if a model selects more than 800 cases, without a score or a probability there is no way to prioritize these cases to extract the best 800.

2.2. Description task

The purpose of the description task is to provide insight into why customers have a caravan insurance policy. This not necessarily the same as explaining the model underlying the predictions. Participants can use different approaches and algorithms for the description and prediction task. Descriptions can be based on prediction models but also on simple tables or parts of models. Given that the value of a description is inherently subjective, a domain expert from insurance marketing evaluated the submitted descriptions. The descriptions and accompanying interpretation were scored on comprehensibility, usefulness and actionability, for a marketing professional with no prior knowledge of machine learning.

2.3. Data characterization

The effect of how steps in the analysis process are performed depends on properties of the data. As pointed out by Wolpert and Macready (1995), heuristic methods or algorithms can only be optimal on a subset of all problems. In this section we provide a characterization of the problem and data.

The data that was available for both tasks consists of 5822 training instances and 4000 test instances.[1] The data contain 83 numeric and 2 symbolic input features and the target, caravan policy yes/no, was only made available to the participants for the train set. The relation between input and target is very weak. The key features to explain policy ownership are not present in the data and measurement of input is noisy for reasons explained later in this Section. The input can be divided in sociodemographic (43 features) and product ownership data (42 features). There are no missing values and all continuous variables have been discretized into at most twelve ranges.

The sociodemographic information is linked to the postal code of the customer rather than to the individual customer. For instance a value of 5 for the feature 'Home Owner' means that presumably 50–62% of people living in the same postal code area as this client own a house. Given that these features are linked to a single hidden variable, geography, these features may be highly correlated.

Measurement noise is also high. The marketing information provider collects the sociodemographics by fusing information from various, possibly conflicting sources. Furthermore, it is certainly possible that customers living in the same area or in similar areas differ with respect to policy ownership. In spite of these obvious limitations, this kind of data is still very common for consumer marketing applications, especially if the responsible department can only access some very general internal information about its customers.

The product ownership data give an overview of the product portfolio of the customer with the insurance company. For 21 policy products, the number of policies owned and amount of revenue is given. It is also very skewed: for 36 out of 42 features over 90% of the instances falls into the majority interval; only 6% actually owns a caravan policy.

The TIC data differ from typical data sets in UCI ML archive. Table 1 shows that the TIC data set is relatively high dimensional. Furthermore we measured predictive power of input features by measuring information gain with respect to the class.

Table 1 shows that the average information gain of all features and also of the five most predictive features is very low compared to UCI sets. The ratio between average predictive power for the top five features versus all features is relatively large, so only a small proportion of the features seem to matter. These differences have consequences for the effect of different methods.

3. Overview of the prediction results

In this Section we give an overview of the results for the prediction task. In total 147 participants registered, 43 sent in a solution to both tasks and 29 supplied a public report or permission to publish the supplied solutions. The sample of published reports is still skewed: only 3 out of the top 50% best performing entries did not supply a report compared to 12 in

Table 1. Some general features of the CoIL Challenge data (unbalanced train set) and selected UCI sets.

	Instances	Input dim.	Avg. info gain	Avg. top 5 info gain	Ratio top 5 to avg.
TIC	5822	85	0.002	0.017	6.91
German credit	1000	20	0.017	0.045	2.62
Hypothyroid	3772	29	0.024	0.132	5.45
Breast cancer	286	9	0.034	0.056	1.66
Pima	768	8	0.064	0.088	1.37
Anneal	898	38	0.097	0.373	3.85
Mushroom	8124	22	0.195	0.481	2.47
Glass	214	9	0.369	0.511	1.39
Soybean	683	35	0.455	0.974	2.14

the bottom half. Still this collection of reports provides a more accurate representation of successes and failures than regular research papers or most other competitions, for which generally only the best results are published. We encouraged participants to report on the entire solution path, including approaches that didn't seem to work. Entries came from both industry (31%) and academia (59%; remainder unknown) and included participants at various skill levels.

A wide variety of methods were used including instance selection, feature selection, construction and transformation, hold out testing, cross validation, bootstrapping and ensemble learning, and cost sensitive classification; core prediction algorithms used included logistic regression, discriminant analysis, Naive Bayes, neural networks, support vector machines, evolutionary algorithms, genetic programming, fuzzy classifiers, RBF networks, self-organizing maps, decision trees, decision tables, rule based systems, ILP based methods and others.

The frequency distribution of scores for the prediction task is displayed in figure 1. To repeat, the participants had to select the 800 most probable caravan policy owners from a test set of 4000 instances (see Section 2.1). The maximum number of policy owners that can be found is 238 (all owners in the test set), the winning model selected 121 policy owners. Random selection results in 48 policy owners (6% of 800).

The performance of the submissions varies over a wide range, from one (!) to two and a half times the number of policy owners that would have been found by random selection and up to half of the maximum number of policy owners possible. These results may seem surprising, given the relatively small differences or improvements that are usually reported on UCI data for instance.

Figure 1 also displays the performance of two reference groups of students who worked on this problem after the competition as an assignment for a data mining course. The first group of students did not receive the test set targets nor were they informed of the CoIL Challenge or any of the results. In contrast, the second group of students read a paper written by the winner of the prediction task (Elkan, 2001). Both groups compete very well with the

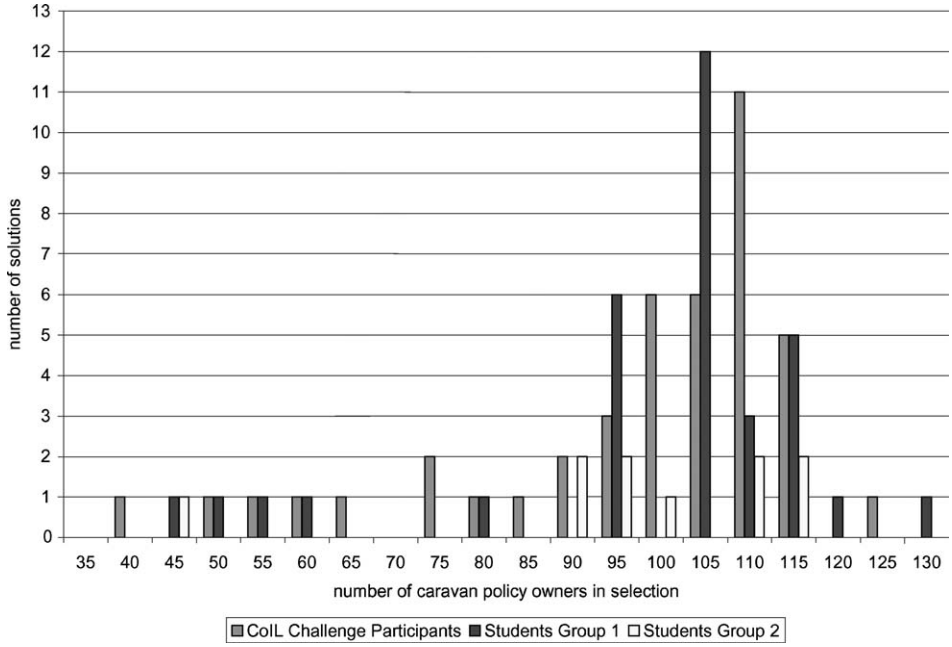

Figure 1. Histogram of prediction task performance for CoIL Challenge participants and two reference groups of students (bucket size is 5).

CoIL participants on this problem. This an interesting result, given that these students were new to data mining. We will suggest an explanation for this in the discussion.

4. Meta analysis approach: Bias variance decomposition

The purpose of this study is to explain the results of the different solutions of the CoIL Competition to better understand the factors that determine the success of real world data mining projects. We organize the analysis by the main steps in the data mining process. According to the CRISP process model (Chapman et al., 1999) the top-level knowledge discovery process consists of business understanding, data understanding, data preparation, modeling, evaluation and deployment. Neither the business and data understanding step nor the evaluation and deployment steps were part of the prediction task of the competition. Therefore we focus on the data preparation (feature construction and selection) and modelling steps.

For each solution to the competition task we collected the following data: prediction accuracy (number of caravan policy owners in test selection), if feature construction was used (and if so, which method was used), features selection (and method), learning method and representation of the result, performance on the description task (comprehensibility, usefulness and actionability).

One characteristic property of this problem is the large noise component in the data. This means that overfitting is likely to be an important source of prediction errors. To analyze this effect we use the concept of bias-variance decomposition. When a model is constructed by a learning method from a sample taken from a given domain, and the model is used to make predictions then some predictions are false. Bias-variance decomposition distinguishes between (1) the *bias error*, a systematic component in the error associated with the learning method and the domain, (2) the *variance error*, a component associated with differences in models between samples and (3) an *intrinsic error* component associated with the inherent uncertainty in the domain. High variance error indicates varying, unstable predictions and the intrinsic error is the variance within each point in the instance space, the error for the Bayes optimal classifier. The variance component is associated with overfitting: if a method overfits the data the predictions for a single instance will vary between samples.

The concept of bias-variance decomposition was introduced to machine learning for mean squared error (Geman, Bienenstock, & Doursat, 1992) and later versions for "zero-one-loss" (predictions are correct or false) were given by Friedman (1997), Kohavi and Wolpert (1996), Breiman (1996), Domingos (2000) and James (2003). Here we use the definition of Kohavi and Wolpert (1996). For a particular target function f and a size of the training set m, the expected misclassification rate $E(C)$ (an error has cost 1 and a correct prediction cost 0) is defined as:

$$E(C) = \sum_x P(x)\left(\sigma_x^2 + bias_x^2 + variance_x\right) \tag{1}$$

with

$$bias_x^2 \equiv \frac{1}{2}\sum_{y\in Y}[P(Y_F = y \mid x) - P(Y_H = y \mid x)]^2$$

$$variance_x \equiv \frac{1}{2}\left(1 - \sum_{y\in Y} P(Y_H = y \mid x)^2\right)$$

$$\sigma_x^2 \equiv \frac{1}{2}\left(1 - \sum_{y\in Y} P(Y_F = y \mid x)^2\right)$$

where x ranges over the instance space X and Y is the predicted variable, with elements $y \in \{0, 1\}$. The actual target function f is a conditional probability distribution $P(Y_F = y_F \mid x)$ and the hypothesis or model h generated by the learner is also a distribution $P(Y_H = y_H \mid x)$. Although this is not included in the equations, the conditioning events in the conditional probabilities are actually parametrised over f and m, the training set. For example, the expression $P(Y_H = y \mid x)$ means $P(Y_H = y \mid x, f, m)$. In terms of the trainingset d, f and m this corresponds to

$$\sum_d P(d \mid f, m, x)P(Y_H = y \mid d, f, m, x) \tag{2}$$

which is equal to

$$\sum_{d} P(d \mid f, m) P(Y_H = y \mid d, x) \tag{3}$$

Here $P(d \mid f, m)$ is the probability of generating data d from target function f and $P(Y_H = y \mid d, x)$ is the probability that the learning algorithm makes guess y for point x after learning from d. Therefore $P(Y_H = y \mid x)$ is the average Y value over all training sets that is guessed for x by the learner. Details can be found in Kohavi and Wolpert (1996).

The intrinsic error and bias error cannot be estimated separately. Here we mostly compare the bias and variance of different methods and in that case the intrinsic error contributes as a constant factor. The combined bias / intrinsic error effect and the variance error are estimated using the implementation in the WEKA toolkit (Witten & Frank, 2000) following Kohavi and Wolpert (1996). The data are split into two parts. From one part samples are drawn, learning is applied and the prediction error on the other half is calculated. We used 50 samples of size 200. The average error is the estimate for the *bias-inherent* error component and the variance between predictions estimates the variance component in the error.

To analyze the differences between methods we reconstructed a number of solutions to estimate the bias-variance decompositions. For these experiments we used data sets with balanced distributions because random sampling leaves a too few buyers instances and does not allow reliable estimates of the error components.

In general there is a trade-off between the strength of learning bias and overfitting. Methods with a strong learning bias are less likely to overfit, because their results depend less on the data than the result of unbiased methods. However, if the learning bias of a method is not correct for a domain than this bias will be a source of prediction errors. Unlike the prediction errors caused by overfitting, errors caused by incorrect learning bias will not decrease with more data.

Table 2 summarises the relation between strength and correctness of learning bias and bias and variance error.

We can now characterize learning methods and also operations as feature selection and feature construction by the effect that they have on the bias and variance components in the error.

Table 2. The relation between strength and correctness of learning bias and variance and bias error.

Learning bias	Correct	Incorrect
Strong	Low variance error	Low variance error
	Low bias error	High bias error
Weak	High variance error	High variance error
	Low bias error	High bias error

5. Lessons learned: Data preparation

In this section we review methods for feature construction, transformation and selection and attempt to explain the influence of these steps for the CoIL problem using bias and variance.

5.1. Feature construction and transformation

Feature construction can be used to avoid bias error caused by limitations in the representation language of a model. Feature construction can reduce bias error by relaxing learning bias (e.g. when features are added that allow models that were initially excluded), or by changing the search bias of a learning method. A risk of adding constructed features is that the variance component in the error increases. However, feature construction can also reduce the variance error even without changing the representation bias. We illustrate the effect of feature construction and discuss the effect on solutions in the challenge problem.

An example of feature construction for avoiding representation bias is described in the winning entry by Elkan (2001). The algorithm he uses, Naive Bayes, is limited in the sense that it cannot model interactions between several input features. To compensate for this limitation he constructed a new feature for each of the two most important products, car and fire policies, by taking the Cartesian product of the number of policies and the turnover amount. The new combinations replace the original features. Elkan claims that this was a vital contribution to his model. We repeated these experiments by comparing the bias and the variance of a Naive Bayes model for the original data and for the data with the constructed features (see figure 2). There is a clear reduction of bias, but balanced to a large extent by an increase in variance. So the reduction in total error is small. Using the product of two features causes "data fragmentation": it reduces the number of instances that are used for estimating the conditional probabilities in each cell, especially for values that already have low marginal frequencies. For a detailed discussion of conditions for the appropriateness of construction of Cartesian products for Naive Bayes and conditions for applicability of the bias of Naive Bayes in general, see Domingos and Pazzani (1997).

Feature construction can also reduce the variance component of the error. For instance, Jorgensen and Linneberg[2] first computed aggregate features such as the total number of policies and the total contribution. These were entered as features in Linear Discriminant Analysis and they were included in the discriminant function after pruning. This approach was found to have lower variance error than the standard method. In this case feature construction does not relax representational bias but it lowers the variance error of the method, in particular the pruning step. Pruning decisions about individual features are more sensitive to sampling variance than are decisions about composite features, for example when one feature represents the sum of several others. In this way, feature construction will reduce the variance error.

If we broaden the definition of feature construction to include transformations on a single feature the entry of White & Liu is also an interesting example. They cross-tabulated all predictors against the caravan policy ownership. All predictors showing "substantial evidence of a quadratic relationship (or higher order polynomial) were recoded into constructed

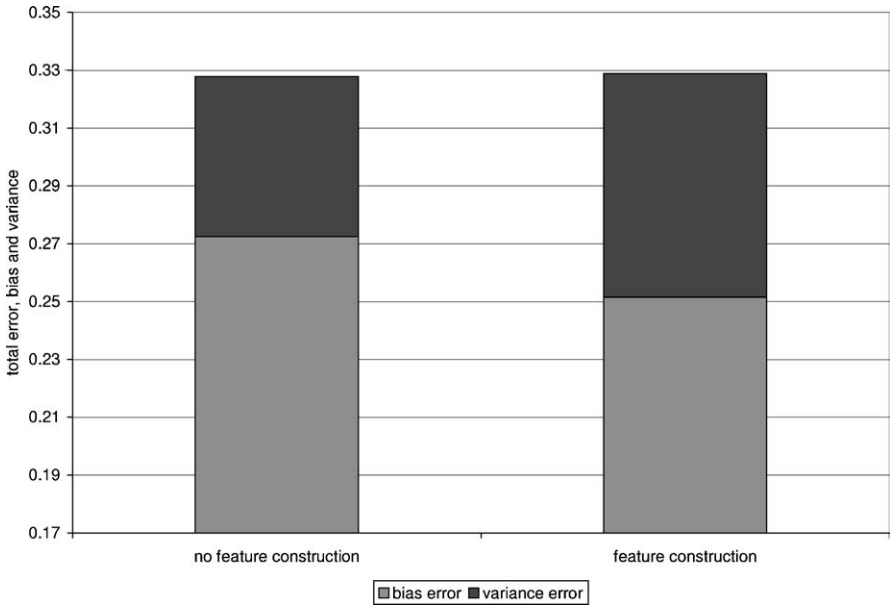

Figure 2. Bias and variance for winning model with and without constructed feature.

variables having a monotone relationship with the class" (van der Putten & van Someren, 2000). This does not relax the representation bias of the learning method that they used, decision tree learning, but is does improve the search bias. This form of linearisation reduces the number of intervals that must be constructed, making interval construction and pruning less sensitive to sampling and thereby this method reduces variance error. Various participants also use discretization to minimize the number of intervals. Apart from changing the feature type from numeric to categorical, which may be needed because of practical limitations of the learner, these methods also aim at reducing the variance by decreasing the degrees of freedom.

We compared the five solutions with highest accuracy (over 110) with the five with the lowest accuracy (below 97). Of these only the best solution, by Elkan, constructed new features and used this to replace the original features. This suggests that for the TIC data, feature construction is not critical. In our analysis of the winning entry, a reduction in bias is countered by an increase in variance (and vice versa). This risk should be taken into account when constructing or transforming features.

5.2. Feature selection

In theory, feature selection can have various effects. Removing features that are irrelevant for the learner will not change the bias error. In other words, it will not change the loss in accuracy due to bias mismatch. As far as the learner is concerned, irrelevant features are noise. Features are irrelevant for a learner either because there is no real relation with the

target, or the learning bias prevents capturing this relation. In the latter case intrinsic error may increase because information is lost.

Feature selection is generally aimed at variance reduction: fewer parameters need to be estimated, whereas the amount of relevant information that is removed is minimized. Removing relevant features may lead to an increase of intrinsic, bias and variance error.

The question is whether feature selection plays a major role in the challenge prediction task. In Section 2.3 and Table 1 we have already shown that the TIC data set has very many features and that the predictive power of individual features is low measured in information gain (it is actually zero for more than half of the features). So the risk of overfitting the relation between individual features and the target is high and eliminating irrelevant features is likely to reduce the variance error.

Another indication is the predictive power of features relative to the best predictor. Figure 3 shows the information gain of features ordered from high to low, with the information gain scaled as percentage of the information gain of the most predictive feature. For the TIC set, only a small proportion of the features has relatively high predictive power.

The effect of feature selection on the final result will depend on the learning method that is used. If the learner itself selects features, a separate feature selection step will have less effect. If we compare the submissions with highest and lowest accuracy, we see that four of the five highest scoring solutions used feature selection and the fifth used a learner that eliminates irrelevant features. Of the five solutions with lowest performance

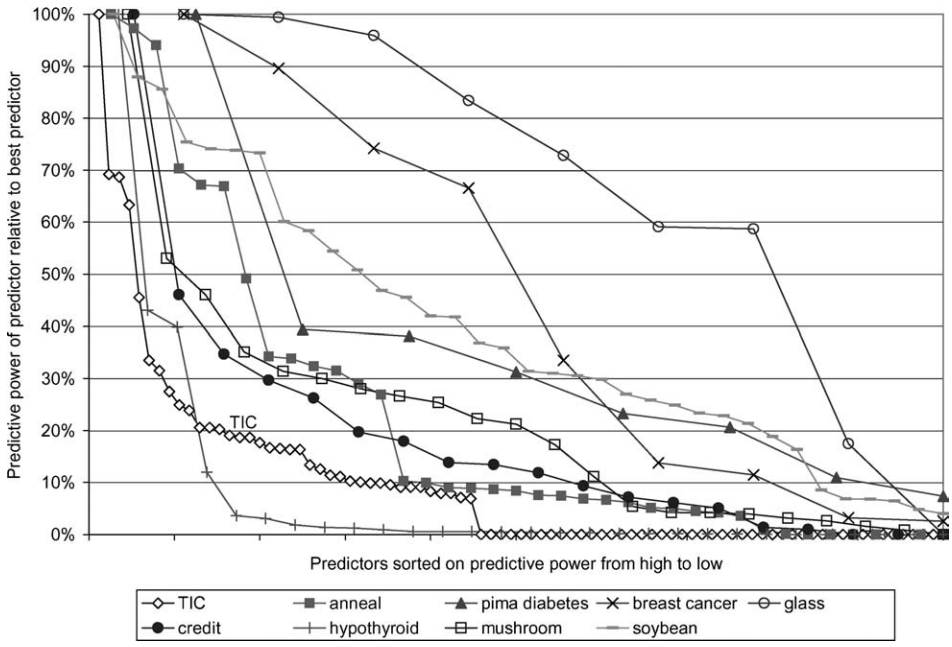

Figure 3. Distribution of predictive power for predictors, measured in information gain in proportion to the best predictor, for TIC (unbalanced training set) and a selection of UCI sets.

only one includes separate feature selection and two use a learning method that eliminates features.

To evaluate the relative importance of feature selection, we performed bias-variance analysis on eight different learners, given the full set and a data set that was reduced to seven variables using the best first version of the CFS feature subset selection algorithm (Hall, 1998; Witten & Frank, 2000). This algorithm takes both the predictive power of a predictor and its correlation to features that are already selected into account. To see the effects of extreme feature selection, we included decision stumps (decision trees of depth 1, Holte, 1993), which were not used by any participant.

As can be seen from figure 4, feature selection improves classification results for seven out of eight learners. When all 16 models are compared, six out of the top eight results are achieved using feature selection. So for TIC using feature selection or not seems to be more important than the choice of learning algorithm. Feature selection reduces variance error for all eight learners.

The feature selection methods used by the participants can be divided into three main categories (see Guyon & Elissee, 2003 for a recent overview of the state of the art in feature selection). The first category consists of approaches where candidate features are evaluated independently of other features. Simple evaluation measures like correlation with the target feature are used, and feature selection is sometimes confirmed or guided by prior domain knowledge as well. For instance the author of the winning entry simply writes: "As is common in commercial data mining, only data about the wealth and personal behavior of

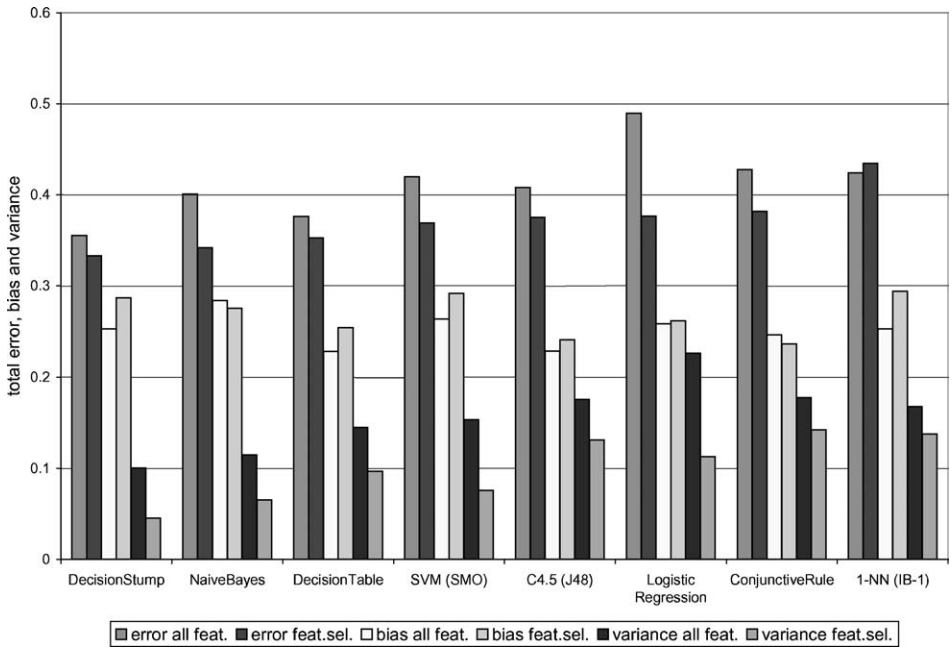

Figure 4. Bias-variance decomposition for eight learners with and without feature selection.

individuals is useful here". He discarded all sociodemographic features but one and the other five selected features were related to product ownership. The winner of the small scale Benelearn competition (van der Putten & van Someren, 2000) that preceded the CoIL Challenge just used some simple cross tabulation to select the best features. The average performance of entries that were restricted to this kind of feature selection (as far as known) was 99 caravan policies selected from the test set.

The second category contains methods that select subsets of features rather than just evaluate features individually and independently. There may be several reasons to look at subsets instead of single features (Kohavi & John, 1997). A feature with high individual predictive power but also high correlation to variables that are already selected does not add much information to the model, so it should not be included. A feature with low individual predictive power may have some complex joint relationship with a selected variable that is highly predictive, in which case it may be advisable to include it. Several specialized subset feature selection algorithms exist (e.g. Hall, 1998), however most participants in this category use a regular learner such as decision trees, decision tables or Naive Bayes to select features. We constrain this category to methods that use a different learner for feature selection than for model development, so these are all so-called filter methods. The average performance in this category was 110 policy owners.

The third category are the so-called wrapper methods (see also Kohavi & John, 1997). These methods select subsets of variables using the same learner both for feature selection and for model development. We use a broad definition for this category and include some participants that experimented extensively with manually selecting and deselecting features and retraining the learner. The average performance in this category was also 110 policy owners.

A difficulty in feature selection is the comparison between models with different sets of features. This means that they involve estimates of different sets of parameters. Comparing only the accuracies hides the uncertainty associated with the predictions. Participants use different approaches for this. Popular approaches are extensive cross-validation experiments and testing on a holdout set.

In conclusion, the majority of the participants use some form of feature selection. The median number of features selected is 10 out of 85. Our experiments suggest that feature selection is of key importance and that is likely to be a greater factor in determining the success than the choice of the learning method. The reason is that feature selection is a powerful tool for variance reduction. Although the number of observations is small, feature subset selection methods seem to outperform methods that evaluate candidate features individually, but for this problem we see no significant difference between subset filter and wrapper methods. This has been confirmed for other domains by a recent study on wrappers and filters (Tsamardinos & Aliferis, 2003).

6. Lessons learned: Model representation and the learning method

The selection of a method for constructing the model of the data is generally considered an important decision in the data mining process, in addition to feature selection and feature construction. The most important property of methods is the model representation. As

discussed briefly in Section 4 an inadequate learning bias of a method can cause bias error. An important characteristic of a method is therefore the strength and the content of the representation (or language) bias. Strong bias will in general reduce the variance error because it forces the learner into a small class of models. If the learning bias is incorrect then this will cause bias error. The advantage of methods with stronger bias is lower variance.

Another way in which the representation affects the error variance is by its exploitation of redundancy. Models that involve (weighted) addition of features or of constructed predictors can exploit the fact that the noise in these predictors will average out and therefore the total prediction error will be smaller than that of the individual predictions.

What is the effect of differences in learning bias between methods? The precise underlying pattern of the TIC domain is not known. The most predictive features are "level of car insurance", "number of car insurance policies", "purchasing power class", "level of fire insurance". These features are correlated and their relation with the target class is approximately monotonic, although not completely, see Section 5.1. This means that Naive Bayes, rule-based, Additive Feature Combinations and Ensembles are all adequate representations. Because of the uncertainty in the relation, additive models are most attractive. Non-linear relations make Naive Bayes and Rule-based representations competitive. Because of the correlation between features, subsets of two or three of these (possibly discretized) variables give comparable optimal models within most model representations.

Table 3 shows the mean and maximum accuracies of the solutions of methods that are based on different model representations.

Although we must interpret these data with care because of the small numbers, this suggests that differences in accuracy between model representations are relatively small. It is remarkable that Naive Bayes performs quite well although it has a relatively strong representational learning bias compared with the other methods. The best solution using Naive Bayes included feature construction and feature selection and this may have corrected the representational learning bias.

Many authors mentioned that they experimented with a number of learning tools, and parameters of tools. This experimentation causes "procedural bias" (Quinlan & Cameron-Jones, 1995; Domingos, 1997): a new method is tried, or a variation of an earlier method and if the accuracy increases then the it is assumed that the new method is better. This may not be true because the new method may have a weaker bias or more degrees of freedom, allowing a closer fit to the data but with weaker support for the learned model.

Table 3. Representation bias and accuracy of CoIL solution methods.

	Bias strength	Additive	Bias correctness	Mean	Max.	Number
Naive Bayes	Strong	Yes	Good	118	121	2
Rule-based	Weak	No	Good	100,5	112	13
Additive linear	Strong	Yes	Medium	111	111	2
Non-linear	Weak	Good	Good	106	115	8
Ensembles	Medium	Yes	Medium	112	115	1

For example, consider first learning "decision stumps", trees of depth one, and then decision trees of depth two. If we would simply compare the accuracies of these two methods, we would probably prefer the trees of depth two because these achieve higher accuracy but the support for the predictions of trees of depth two, for the path from root to leaf, is weaker for the deeper trees and the variance error and risk of overfitting will increase. Therefore only comparing the accuracies is not enough. Using cross validation can resolve this problem but only when the test set is very large. Cross validation itself relies on a sample and is therefore subject to error. As an aside we note that Domingos (1997) claims that overfitting is caused by testing too many hypotheses on a single data set but this is not the main problem. For example, trying more candidate decision stumps will only improve the quality of the decision stump that is found. The problem is in the comparison of hypotheses (or classes of hypotheses) with different complexities or degrees of freedom. Usually a learning process that involves more hypothesis testing involves more comparisons between hypotheses with different degrees of complexity (or degrees of freedom) and therefore the risk of incorrect decisions is increased.

Because we do not have enough data about the number and nature of the experiments performed by participants, we focus on the methods that were used. Table 4 compares the five methods with highest scores with the five with lowest scores for which the method is known with enough detail.

These results show that methods with weak learning bias tend to have low accuracies. A possibility is that participants have been seduced by the high accuracies that can be obtained by learners with weak learning bias and did not realize that the models found in this way cannot be directly compared with models found by methods with stronger learning bias.

We explore the effect of model representation shape by reconstructing some of the solutions. Specifically, we can look at estimates of bias error and variance error in the CoIL solutions. Figure 4 gives the result for bias-variance decomposition. Even after feature selection, the bias component is relatively large because its estimate here includes the (constant) inherent error. The spread in bias error is relatively small. If we exclude decision stumps

Table 4. Accuracy and method selection.

Accuracy	Method
121	Naive Bayes
115	Ensemble of 20 pruned Naive Bayes
112	GA with numerical and Boolean operators
111	SVM regression
111	SVM regression
96	Subgroup discovery
96	Decision tree
80	Fuzzy rules
74	CART
72	Neural nets

then the range of the bias component in the errors ranges from 0.22 to 0.26. The variance component shows more variability. The rule-based methods have a variance component in the error of around 0.20. The most stable results are produced by Naive Bayes which has a variance error of only 0.09. The results show that for the TIC domain the bias error is more stable between methods than the variance error. This suggests that selection or even improvements of methods should be found in reducing the variance component associated with overfitting, rather than the bias component.

7. Lessons learned: Description task

The goal in the description task was to explain why people own a caravan policy, given the data, modeling methods and subjective, domain-based interpretation. The descriptions and accompanying interpretation had to be comprehensible, useful and actionable for a marketing professional with no prior knowledge of computational learning technology.

Submitted descriptions were evaluated by a marketing expert (Stephan van Heusden from MSP Associates in Amsterdam). The expert commented: "Almost all entries lack a good description in words: participants seem to forget that most marketeers find it difficult to read statistics (and understand it)". The expert stressed the importance of actionability. "A good entry combines a description of the results with a tool to use the results." Participants from industry had a better score than academic participants (4.3 versus 3.5 out of a maximum of 6), although these differences were not significant given the standard deviations (1.6 resp. 2.1). Similar to the prediction solutions, a wide variety of approaches was chosen, although there was a tendency to use simple statistics, cross tabulations and rule based solutions.

The winners, Kim and Street, used a variety of techniques for the description task, including proprietary evolutionary algorithms, chi square tests and association rules. The marketing expert remarked: "This participant clearly explained which steps preceded his conclusions. He may have discovered all conclusions the others also found, but additionally he really tries to interpret them." The expert also appreciated the association rule results, the discussion of the complex nonlinear relation between purchasing power and policy ownership and the explanations why people were not interested in a caravan policy. Their prediction model scored 107 policy owners (65th percentile).

The success of the winning entry demonstrates that to explain behavior the prediction models used do not necessarily need to be of top quality. This was confirmed by analyzing the correlation between prediction and description scores for all entries. Rather than giving a complete overview of a single model, good results are achieved by applying multiple methods and choosing the most comprehensible, useful and actionable patterns from these models.

8. Discussion and conclusion

The CoIL competition resulted in analyses of a real-world problem by a number of experts. We used the bias-variance decomposition of errors to identify causes of success and failure in solving the competition problems. The TIC problem is characterized by a relatively large

number of intercorrelated features, much uncertainty and skewed distributions of inputs and target. This is common for many real world problems. In this case variance error was a bigger problem than bias error.

Attempts to discover complex models using methods with weak learning bias and weak methods to avoid overfitting lead to complex models that were unstable and that overfit. The best approach is to simplify the data set through data preparation first and then use simple, robust, strong bias methods for modeling. This suggests a potential reason why the CoIL experts didn't outperform the students. Apparently a simple model and experimental setup suffice to solve this noisy prediction problem.

8.1. Lessons

To summarize we would like to single out the following lessons learned for problems similar to the challenge:

1. In the case of noisy prediction problems, choices in the analysis process should be aimed at reducing the variance component rather than at finding an appropriate bias.
2. The potential impact of data preparation (e.g. selecting the right features) on variance reduction is larger than for model development, e.g. by selecting the right features.
3. Attempting to improve the fit between the bias of the problem and method by using complex learners is only useful when parameters can be estimated reliably. For problems like the one of the CoIL Challenge 2000, simple models with few degrees of freedom are most robust so these should be tried first.
4. All steps of the data mining process risk increasing variance error. This suggests that model stability should be tested: if a method, applied to different samples, produces different models with different predictions, this suggests that variance error may be a problem. The method may be overfitting the data or may be otherwise unstable.
5. Measures against overfitting all have their weaknesses. Cross validation, being an empirical method, is not guaranteed to result in the best model, especially if it is used to limit the complexity of a model such as a decision tree. If intrinsic error is high and the amount of data is low, estimates of validation set performance are uncertain. Extensive experimentation while only keeping the 'best' models is also risky. Quite a few participants reported that they were seduced to spoil their original models (cf. Seewald; Abonyi & Roubos; Sathiya Keerthi & Jin Ong; Kaymak & Setnes). They increased the fit on the data by additional heuristics or fine-tuning, ending up with a model that is worse rather than better than the original.

8.2. Further research

We identify a number of directions for further research:

1. Bias-variance decomposition is an elegant framework for the analysis of learning problems because it provides a diagnosis of error into various components. Each component

requires a different strategy of error reduction, so the data miner has more insight into what action to take to improve the model. To become a standard tool, a more universally accepted definition of bias and variance is needed, both for zero one loss and for other loss functions and evaluation metrics, such as for asymmetric cost functions and the area under the ROC curve.

2. This study confirms the importance of steps before and after the core modeling step such as feature construction, feature selection and model evaluation. Bias-variance decomposition could be used more to analyze and improve methods used in these steps. Evaluation of bias and variance components should be integrated more tightly in methodology, methods and systems for Machine Learning. In addition to using the concepts of bias and variance for analysis, methods could be developed that explicitly minimize bias and variance error.

Acknowledgments

This work builds on the remarkable group effort carried out by the participants in the Benelearn 1999 and CoIL Challenge 2000 competitions. The CoIL competition was organized by the network of Computational Intelligence and Learning, supported by the European Commission. The authors would like to thank the participants for both for their solutions and for the many discussions during and after the competitions. We also want to thank Maarten Keijzer at the Free University in Amsterdam and Boris Kovalerchuk of Central Washington University for the student results, Edwin de Jong for his comments, Sentient Machine Research in Amsterdam for their support in preparing the competition, the other members of the CoIL competition committee and the reviewers who helped us to significantly improve the paper.

Notes

1. The CoIL Challenge 2000 problem is also referred to as The Insurance Company (TIC) benchmark. Data sets, problem descriptions, background info and reports can be found on the benchmark homepage at http://www.liacs.nl/~putten/library/cc2000.
2. All public challenge submissions can be found in a tech report by van der Putten and van Someren (2000).

References

Berka, P. (1999). Workshop notes on discovery challenge PKDD-99. Technical report, Laboratory of Intelligent Systems, University of Economics, Prague.

Blake, C., & Merz, C. (1998). UCI Repository of machine learning databases.

Breiman, L. (1996). Bias, variance, and arcing classifiers. Technical Report, Statistics Department, University of California.

Chapman, P., Clinton, J., Khabaza, T., Reinartz, T., & Wirth, R. (1999). The CRISP-DM process model. Technical Report, Crisp Consortium. http://www.crisp-dm.org/.

Domingos, P. (1997). The role of Occam's Razor in knowledge discovery. *Data Mining and Knowledge Discovery, 3*, 409–425.

Domingos, P. (2000). A unified bias-variance decomposition and its applications. In *Proceedings of the Seventeenth International Conference on Machine Learning* (pp. 231–238). CA: Morgan Kaufmann.

Domingos, P., & Pazzani, M. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning, 29*, 103–130.

Elkan, C., (2001). Magical thinking in data mining: Lessons from CoIL challenge 2000. In *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining (KDD'01)* (pp. 426–431).

Friedman, J. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery, 1*, 55–77.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation, 4*, 1–58.

Guyon, I., & Elissee, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*, 1157–1182.

Hall, M. A. (1998). Correlation-based feature subset selection for machine learning. Ph.D. thesis, University of Waikato.

Holte, R. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning, 11*, 63–91.

James, G. M. (2003). Variance and bias for general loss functions. *Machine Learning, 51*, 115–135.

Kohavi, R. & John, G., (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*, 273–324.

Kohavi, R., & Wolpert, D. H. (1996). Bias plus variance decomposition for zero-one loss functions. In L. Saitta (Ed.), *Machine learning: Proceedings of the thirteenth international conference* (pp. 275–283). Morgan Kaufmann.

Quinlan, J. R., & Cameron-Jones, R. M. (1995). Oversearching and layered search in empirical learning. In *IJCAI* (pp. 1019–1024).

Tsamardinos, I., & Aliferis, C. (2003). Towards principled feature selection: Relevancy, filters and wrappers. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*.

van der Putten, P., & van Someren, M. (2000). CoiL challenge 2000: The insurance company case. Technical Report 2000-09, Leiden Institute of Advanced Computer Science, Universiteit van Leiden. http://www.liacs.nl/∼putten/library/cc2000.

Witten, I., & Frank, E. (2000). *Data mining: Practical machine learning tools and techniques with java implementations*. San Francisco: Morgan Kaufmann Publishers.

Wolpert, D. H., & Macready, W. G. (1995). No free lunch theorems for search. Technical Report SFI-TR-95-02-010, The Santa Fe Institute.

Received April 28, 2003
Accepted April 8, 2004
Final manuscript April 13, 2004