The Knowledge Engineering Review, Vol. 25:1, 27–47. © Cambridge University Press, 2010
doi:10.1017/S0269888909990154

Automatic selection of reliability estimates for individual regression predictions

ZORAN BOSNIĆ and IGOR KONONENKO

University of Ljubljana, Faculty of Computer and Information Science, Tržaška 25, Ljubljana, Slovenia;
e-mail: zoran.bosnic@fri.uni-lj.si, igor.kononenko@fri.uni-lj.si

Abstract

In machine learning and its risk-sensitive applications (e.g. medicine, engineering, business), the reliability estimates for individual predictions provide more information about the individual prediction error (the difference between the true label and regression prediction) than the average accuracy of predictive model (e.g. relative mean squared error). Furthermore, they enable the users to distinguish between more and less reliable predictions. The empirical evaluations of the existing individual reliability estimates revealed that the successful estimates' performance depends on the used regression model and on the particular problem domain. In the current paper, we focus on that problem as such and propose and empirically evaluate two approaches for automatic selection of the most appropriate estimate for a given domain and regression model: the internal cross-validation approach and the meta-learning approach. The testing results of both approaches demonstrated an advantage in the performance of dynamically chosen reliability estimates to the performance of the individual reliability estimates. The best results were achieved using the internal cross-validation procedure, where reliability estimates significantly positively correlated with the prediction error in 73% of experiments. In addition, the preliminary testing of the proposed methodology on a medical domain demonstrated the potential for its usage in practice.

1 Introduction

When modeling data in supervised learning, we most commonly evaluate induced predictive models by computing the accuracy measures that are averaged across all testing examples. Such averaged accuracy measures are, for example, the mean squared error (MSE) and the relative mean squared error (RMSE), which summarize the error contributions of test examples but do not provide any information about the expected error of a particular unseen example. In modeling efforts, which include achieving the best possible prediction accuracy for the unseen examples that were not included in the learning process (Kononenko & Kukar, 2007), we might find such information about single prediction reliability (Crowder *et al.*, 1991) important and beneficial.

Important potential application areas of the individual prediction reliability estimates are the risk-sensitive applications of machine learning (e.g. medicine, financial, control applications, and so on). Namely, the availability of additional information about prediction reliability can enable users of the decision-making applications to decide to what degree they can trust the prediction. The physicians, managers, operators, and so on can therefore use the reliability estimates to decide whether they will accept the system's prediction and perform a corresponding action in the real-world (e.g. prescribe a medicine, make a business decision, change navigation direction, and so on) or not. In contrast to the averaged accuracy measures, which allow the user to evaluate the model's accuracy (induced on a given set of training examples) on the whole, the individual prediction reliability estimates, therefore, allow users to perform accuracy evaluation on an example basis.

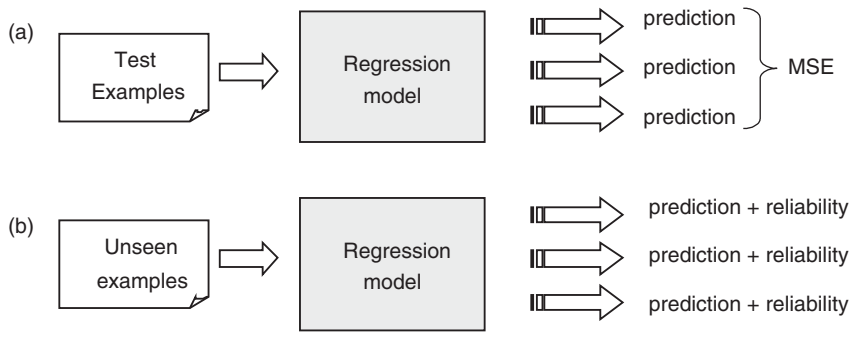

Figure 1 Reliability estimate for the whole regression model (a) in contrast to reliability estimates for individual predictions (b)

Besides providing reliability information, the concept of estimating individual predictions reliabilities has also an advantage to the averaged accuracy measures. Since this approach, in contrast to the averaged accuracy measures, does not require test examples and knowledge of their true labels, the reliability estimates can be computed for an arbitrary unseen example. This advantage is illustrated in Figure 1, which shows a contrast between the average reliability estimate (e.g. MSE) and the reliability estimates of individual predictions.

At a glance, in contrast to the averaged accuracy measures, which accurately express prediction accuracy for some test set, the task of predicting reliability for yet unseen examples poses as harder. The success of individual prediction reliability estimation (Figure 1b) indeed mostly depends on the design of such methods. To tackle this challenge, the related work in this area included the development of various methods. Basically, we can divide these methods into two families, with respect to how they are bound to the underlying prediction model. These two families are: (i) model-dependent approaches and (ii) model-independent approaches. The first group of approaches is focused on extending formalizations of specific classification and regression models, hence upgrading them to output reliability estimates as a supplemental information to the predictions. Due to exact model-related formalizations, these approaches are able to exploit the model-specific properties and can, therefore, be defined with the probabilistic interpretation (meaning that the estimate's values belong to the interval [0,1], where 0 represents the confidence of the most inaccurate prediction and 1 the confidence of the most accurate one).

In contrast, the model-independent approaches are more general and utilize the predictive models as their parameters (black box principle). The approaches from this group are mostly based on estimating the reliability of the individual examples by observing the local influence of a particular learning example to a model, either by influencing the parameters, which are available in any supervised learning framework (e.g. the learning set, attributes, and so on), local modeling or by exploiting general properties of the input space. Since the formalizations of these approaches do not depend on models' formalizations, they are harder to evaluate analytically with the individual models. The reliability estimates that are based on these approaches are, therefore, usually not probabilistically interpretable, meaning that they can take values from an arbitrary interval of numbers.

Our previous work (Bosnić & Kononenko, 2007, 2008b) aimed at developing general, model-independent approaches, which can be used to estimate prediction reliability of the arbitrary regression model. We proposed and compared nine model-independent reliability estimates (their basic versions and their variants) that were based on various approaches: sensitivity analysis, measuring variance of bagged predictors, local cross-validation, density-based estimation and local error estimation. The testing results of individual reliability estimates, achieved by measuring the correlation of the reliability estimates to the prediction error, showed that the estimates have a potential for estimation of the prediction reliability. However, although using model-independent reliability estimates, the results showed that different estimates achieve different performance on

different domains and with different regression models. This issue left an open challenge to design an approach for the automatic selection of the reliability estimate that would perform best (i.e. to achieve best correlation with the prediction error among all available estimates) for the given domain and the given model.

In this paper, we propose and test two approaches to automatic selection of the best performing reliability estimate for a given problem domain and regression model: internal cross-validation and meta-learning. The paper is organized as follows. Section 2 summarizes previous work from related areas of individual prediction reliability estimation and Section 3 summarizes the reliability estimates that we use in the proposed approaches for automatic reliability estimate selection. We describe our testing environment and the testing protocol in Section 4. Sections 5 and 6 define and evaluate the proposed two approaches for automatic reliability estimate selection. In Section 7, we describe an application of the developed methodology on a real problem from medical prognostics. The final Section 8 provides the comparison of both approaches, conclusions and ideas for further work.

2 Related work

The work presented here is based on the ideas from a number of related fields. In Section 2.1, we start by presenting the related work in *estimation of the individual prediction reliability*. Afterwards, in Sections 2.2 and 2.3, we focus on the *sensitivity analysis*-based reliability estimates and other *traditional approaches* to reliability estimation, for which we evaluate the automatic selection procedures in this paper. Sections 2.4 and 2.5 present the related work in *internal cross-validation* and *meta-learning*, which we adapt and evaluate in our paper for the purpose of the selection of the most appropriate reliability estimate.

2.1 Estimation of individual prediction reliability

The idea of reliability estimation for individual predictions originated in statistics, where confidence values and intervals are used to express the reliability of estimates. On the same basis, the reliability estimation was implemented in machine learning methods, where the statistical properties of predictive models were utilized to expand their predictions with adjoined reliability estimates. Although these approaches are specific for a particular predictive model and cannot be generalized, they provide favorable results to the general approaches. Such reliability estimates were developed for the Support Vector Machines (Gammerman *et al*., 1998; Saunders *et al*., 1999), the ridge regression model (Nouretdinov *et al*., 2001), the multilayer perceptron (Weigend & Nix, 1994), the ensembles of neural networks (Heskes, 1997; Carney & Cunningham, 1999) and others.

In contrast to the latter group of methods, the general (i.e. model-independent) methods utilize approaches, such as local modeling of prediction error based on input space properties and local learning (Birattari *et al*., 1998; Giacinto & Roli, 2001), meta-predicting the leave-one-out error of a single example (Tsuda *et al*., 2001), transductive reasoning (Vapnik, 1995; Kukar & Kononenko, 2002) and sensitivity analysis (Breierova & Choudhari, 1996; Kearns & Ron, 1997; Kleijnen, 2001; Bousquet & Elisseeff, 2002). In our previous work, we focused mostly on applying the sensitivity analysis to the context of reliability estimation for regression predictions (Bosnić *et al*., 2003; Bosnić & Kononenko, 2007, 2008a). We proposed the standard framework for the use of sensitivity analysis in machine learning, supported by the Minimum Description Length principle (Li & Vitányi, 1993), and evaluated the sensitivity reliability estimates with many different regression models.

2.2 Sensitivity analysis in the context of reliability estimation

An approach which allows to analyze the *local* particularities of learning algorithms is the *sensitivity analysis* (Breierova & Choudhari, 1996; Kearns & Ron, 1997; Kleijnen, 2001; Bousquet & Elisseeff, 2002), which is used in statistics and mathematical programming. Sensitivity analysis aims at determining how much the variation of input can influence the output of a system.

The idea for putting the reliability estimation in the context of the sensitivity analysis framework is, therefore, in observing the changes in model outputs by modifying its inputs. Treating the predictive model as a black box, the sensitivity analysis approach, therefore, indirectly analyzes qualitatively describable aspects of the model, such as generalization ability, bias, resistance to noise, avoidance of overfitting, and so on.

In our previous work (Bosnić & Kononenko, 2007), we based the sensitivity reliability estimates on observing the change in prediction when expanding the initial learning set with an additional example. The motivation came from the related fields, which have implied the dependencies between learning set composition and model accuracy. The related fields were data perturbation (Wolpert, 1992; Breiman, 1996; Freund & Schapire, 1997; Tibshirani & Knight, 1999; Elidan *et al.*, 2002), usage of unlabeled examples in supervised learning (de Sa, 1993; Blum & Mitchell, 1998; Mitchell, 1999; Goldman & Zhou, 2000; Seeger, 2000), active learning (Cohn *et al.*, 1990; Linden & Weber, 1992; Cohn *et al.*, 1995), transductive reasoning, meta-learning and reinforcement learning (Whitehead, 1991; Schmidhuber & Storck, 1993). The testing results on numerous benchmark and real-world domains showed the potential for the usage of the sensitivity estimates in practice.

2.3 Traditional approaches to reliability estimation for individual examples

In the later work (Bosnić & Kononenko, 2008b), the performance of sensitivity estimates was compared to four other approaches to reliability estimation for individual examples. The novel reliability estimates were either adapted from the traditional approaches (generalized for the usage with the arbitrary regression model) or proposed as a novelty. The summarized definition of these estimates is given in Section 3.

The performance (correlation of the estimate to the prediction error) of all implemented reliability estimates was evaluated using eight regression models (regression trees, linear regression, neural networks, bagging with regression trees, support vector machines, locally weighted regression, random forests and generalized additive model) on 28 testing domains. The testing results of individual reliability estimates showed that the estimates have a potential for estimation of the prediction reliability. However, the results also demonstrated that different estimates achieved different performance on different domains and with different regression models. Finding this out, we tried to define a new estimate as a combination of individual estimates, which would achieve such performance using a particular regression model/domain as the best estimate constituting the combination. Although the results showed the improvement in the performance, we are still motivated to explore the other approaches that might improve the results.

2.4 Internal cross-validation

If lacking relevant problem-specific knowledge, cross-validation methods may be used to empirically select a learner (Schaffer, 1993). When faced with a number of possible learning strategies and having no prior knowledge about the data, a natural idea is to allow the data itself to indicate which method will work best. Using the cross-validation approach, we divide the data into two parts, use one part as an input to a number of learning algorithms and then choose the algorithm that produces the most accurate model on the second part. This idea can also be conducted as a cross-validation study, partitioning the data into a number of groups, using each in turn as a test set for models produced on the basis of the remaining data. The finally chosen method is the one that achieves the highest average accuracy.

Besides selecting the most appropriate learning algorithm, internal cross-validation can be used also for related purposes, for example, for examining the stability of data in clustering (Krieger & Green, 1999). Our work utilizes the adapted general cross-validation approach for selection of the most appropriate reliability estimate for a given domain and model. We define our approach in Section 5.

2.5 Meta-level reasoning

Similarly to internal cross-validation, a way to relate the performance of the algorithms to the characteristics of the data sets is using the meta-level learning (Michie *et al*., 1994). The purpose of the meta-learning process is in generating a set of rules capable of relating these two concepts based on our past empirical knowledge concerning the algorithms. In order to achieve this aim, one needs to determine which features for describing this problem are relevant, thus defining a set of meta-level attributes for this problem. The meta-level rules can be constructed manually or with the help of machine learning methods on the basis of past cases.

In related work, other approaches to meta-learning have also been implemented. Schmidhuber *et al*. (1996) interpret meta-learning as generating useful shifts of inductive bias by adapting the learning strategy. Based on the ability to choose the bias dynamically, meta-learning differs from the base-learning in which the bias is fixed *a priori* or user-parameterized (Vilalta & Drissi, 2002). Based on this definition, Gordon and desJardins (1995) develop a framework for the study of dynamic bias as a search over three tiers: hypothesis space, hypothesis parameters (strength and size) and a tier for defining the meta-spaces. Among the other such approaches, the most well-known include: stacked generalization (Wolpert, 1992), which is considered a form of meta-learning because the transformation of the training set conveys information about the predictions of the base-learners; selecting a learning algorithm for each individual test example based on the algorithm's performance exhibited in the example's neighborhood (Merz, 1996); and inductive transfer of learnt knowledge across domains or tasks (Caruana, 1997; Pratt & Jennings, 1998).

This paper focuses on adapting a general approach of meta-learning, which consists of defining a set of domain characteristics or meta-features that are relevant to the performance of the learning algorithm and to inducing a predictive model (Aha, 1992; Gama & Brazdil, 1995). Instead of selecting the most appropriate learning algorithm, we adapt the general meta-learning approach to select the most appropriate reliability estimate for a given domain and model, based on the domain and model characteristics. Section 6 provides the description of the proposed approach.

3 Overview of reliability estimates

In this paper, we propose two approaches to automatic selection of the reliability estimates for individual regression predictions, which were proposed and evaluated in our previous work (Bosnić & Kononenko, 2008b). These nine reliability estimates (basic versions and their variants) were based on the following five approaches: sensitivity analysis, variance of bagged models, local cross-validation, density-based estimation and local error estimation.

Since the estimates were designed to be independent of a used regression model, they belong to the family of estimates that are generally not probabilistically interpretable (see Section 1). The presented reliability estimates are therefore designed as metrics, of which values estimate the true prediction error (which is unknown for the unlabeled examples). Values of each estimate belong to estimate-dependent interval, which is not related to the prediction error interval. This therefore does not enable the users to estimate the exact prediction error, but to relatively distinguish between more and less reliable predictions (greater estimate values denote greater estimated prediction error). In the previous work, we measured the performance of such reliability estimates by evaluating their correlation coefficients with the prediction errors of test cases (in the leave-one-out cross-validation setting). We denoted the experiment (using a particular model and domain) as successful, if the estimate statistically significantly positively correlated with the prediction error, and unsuccessful if the correlation was significantly negative (undesired result) or not significant (estimate does not correlate to the prediction error). For statistical evaluation, t-test for correlation coefficients was used with the significance level $\alpha \leqslant 0.05$.

In the following, we summarize the core ideas behind the definition of each estimate and briefly present the most relevant results. All testing results are expressed as the percentage of successful experiments with respect to the number of all experiments performed (for each combination of 28

testing domains and 8 regression models, $28 \times 8 = 224$ experiments). For more details, see Bosnić and Kononenko (2007, 2008b).

Savar: Reliability estimate, based on the sensitivity analysis (see Section 2.2), which measures the prediction variance, achieved with different *sensitivity models* (i.e. models built on the modified versions of the learning set). The testing results showed the prevailing percent of experiments with the significant positive correlations to the prediction error (54%) with linear regression and generalized additive model.

SAbias-s: Estimate, based on the sensitivity analysis (see Section 2.2), which measures the local prediction bias. The bias is computed using the differences between the prediction of the original regression model and the predictions of the *sensitivity models*. Since values of this reliability estimate can also be negative (suffix -s denotes *signed*), the estimate also provides the additional information about the error direction (whether the value of prediction was too high or too low), which holds a potential for the further work in correcting the initial predictions. The estimate achieved outstanding results using the regression trees (82% of experiments with the significant positive correlations to the prediction error).

SAbias-a: The absolute version of SAbias-s (suffix -a denotes *absolute*), which is tested for correlation with the absolute prediction error only.

BAGV: The variance of predictions in the bagged (Breiman, 1996) aggregate. In the previous work, the estimate was tested using aggregates of 50 model instances. Since bagging can be used with an arbitrary regression model, this estimate was adapted from the usage with the neural networks (Heskes, 1997; Carney & Cunningham, 1999). Besides achieving the best average performance, the evaluation showed that this estimate is the most appropriate for the usage with the locally weighted regression (46% of the successful tests).

LCV: The estimate locally models the prediction error by applying the cross-validation procedure locally (Woods *et al*., 1997; Birattari *et al*., 1998; Schaal & Atkeson, 1998, 1994; Giacinto & Roli, 2001). The estimate is computed as the weighted average of leave-one-out prediction errors, obtained by applying the leave-one-out cross-validation procedure only to the subspace defined by the nearest neighbors of the particular example (for which we are estimating the prediction reliability). The evaluation revealed that *LCV* is the most appropriate estimate for the usage with the support vector regression (61% of the successful tests), locally weighted regression (46% of the successful tests) and random forests (61% of the successful tests).

DENS: The reliability estimate, based on the distribution of learning examples in the input space (Wand & Jones, 1995). It is defined as an inverted value of the estimated probability density function (Silverman, 1986; Jeon & Landgrebe, 1994) for a given unlabeled example. The estimate did not achieve remarkable results with any of the testing regression models.

CNK-s: Reliability estimate that models the prediction error locally as the difference between averaged nearest neighbors' label and the prediction of the example in question. Similarly to *SAbias-s*, the estimate is signed and therefore provides the potential for the further work in correcting the initial predictions. It achieved the best performance using the regression trees (86% of the successful tests).

CNK-a: The absolute version of CNK-s, which is tested for correlation with the absolute prediction error only. The estimate achieved the best results using the linear regression and the generalized additive model (57% of the successful tests with both models).

BVCK: Linear combination of estimates *BAGV* and *CNK-a*, which outperformed (achieved higher percent of the significant positive correlations with the prediction error) all of the above individual estimates. The estimate achieved the best results with neural networks (54% of the successful tests) and with bagging (61% of the successful tests).

The best average testing results were achieved using the estimates *BVCK*, *BAGV*, *CNK-a*, *LCV* and *SAvar* in the decreasing order with respect to the percent of the significant positive correlations. The estimate *SAbias-a* achieved the worst average results. The comparison of the results, averaged across all regression models, is for all reliability estimates displayed in Figure 2.

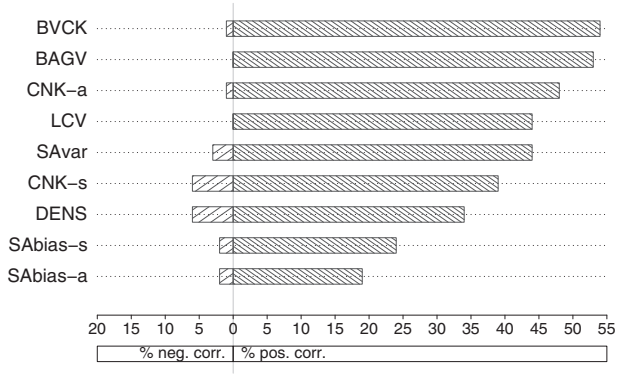

Figure 2 Ranking of reliability estimates by the average percent of significant positive correlations with the prediction error (average percent of significant negative correlations is also displayed)

The results indicate that the estimates *BVCK*, *SAbias*, *CNK*, *BAGV* and *LCV* have a good potential for estimation of the prediction reliability. However, the results also showed that these estimates performed differently with different regression models and in different domains. This can be seen from Table 1, which shows a contingency of successful reliability estimates (i.e. estimates that significantly positively correlated to the prediction error) among testing domains and regression models (testing domains and model abbreviations are further described in Section 4). The dependence of the reliability estimates on the particular models and domains motivated us to explore approaches to automatic selection of the best-performing reliability estimate for a given domain and regression model. These are discussed in Sections 5 and 6.

4 Experimental environment

We tested and compared the performance of nine individual estimates, presented in previous section (*SAvar*, *SAbias-s*, *SAbias-a*, *BAGV*, *LCV*, *DENS*, *CNK-s*, *CNK-a* and *BVCK*) to the performance of the two proposed approaches, presented in Sections 5 (internal cross-validation) and 6 (meta-learning). For each given domain and model, the most appropriate reliability estimate was first selected. Afterwards, the testing was performed using the leave-one-out cross-validation procedure. For each learning example that was left out in the current iteration, the prediction and all the tested reliability estimates were computed. Having completed all iterations, the leave-one-out errors and reliability estimates were therefore computed for all available examples. The performance of the reliability estimates was measured by computing the Pearson correlation coefficient between the reliability estimate and the prediction error. The significance of the correlation coefficient was afterwards statistically evaluated using *t*-test.

The testing was performed using eight regression models, implemented in statistical package R (R Development Core Team, 2006). In the following, we provide a brief description of some key properties of used models.

Regression trees (RT): Trees (Breiman *et al.*, 1984) with the mean squared error used as the splitting criterion; the values in leaves represent the average label of corresponding training examples;

Linear regression (LR): Linear regression with no explicit parameters;

Neural networks (NN): Three-layered perceptron (Rumelhart *et al.*, 1986) with five hidden neurons, *tanh* activation function, the backpropagation learning algorithm using adaptive gradient descent;

Bagging (BAG): Bagging (Breiman, 1996) with 50 regression trees;

Support vector machines (SVM): The version of regression SVM (Vapnik, 1995; Smola & Schölkopf, 1998), implemented in the LIBSVM library (Christiannini & Shawe-Taylor, 2000; Chang & Lin, 2001); When we used the third-degree RBF kernel, the precision parameter was $\varepsilon = 0.1$;

Table 1 The numbers of reliability estimates (out of nine) that were successful (achieved significant positive correlation to the prediction error) in different domains and using different regression models

	RT	LR	NN	BAG	SVM	LWR	RF	GAM	Success (%)
cpu	9	7	8	6	9	6	6	7	81
fishcatch	8	6	7	8	8	7	6	6	78
auto_price	9	6	9	7	6	6	6	6	76
transplant	6	8	8	7	9	2	7	8	76
hungarian	8	5	5	8	7	6	8	5	72
cloud	9	5	5	6	9	5	6	5	69
autohorse	8	6	6	5	7	2	5	6	63
triazines	7	6	4	7	2	3	5	6	56
auto93	8	3	6	5	3	6	5	3	54
bodyfat	6	4	6	4	9	2	4	4	54
pharynx	9	6	0	7	0	5	5	6	53
servo	8	4	4	5	2	5	5	4	51
pwlinear	8	5	3	1	4	6	0	5	44
elusage	5	1	3	5	6	6	4	1	43
pyrim	4	3	7	2	4	5	1	5	43
grv	8	4	2	6	1	0	1	3	35
sleep	2	5	4	2	2	0	1	5	29
echomonths	6	0	0	2	4	3	5	0	28
pollution	5	6	0	0	1	0	0	6	25
diabetes	3	3	2	1	1	3	0	2	21
tumor	3	0	1	1	4	0	1	0	14
wpbc	3	2	0	1	1	0	1	2	14
baskball	4	1	0	0	0	0	1	1	10
lowbwt	1	0	1	3	1	1	0	0	10
breasttumor	2	0	0	0	2	0	0	0	6
mbagrade	1	0	0	0	0	1	2	0	6
fruitfly	1	0	0	0	0	0	1	0	3
brainsize	0	0	0	0	0	1	0	0	1

RT = Regression trees; LR = Linear regression; NN = Neural networks; BAG = Bagging; SVM = Support vector machines; LWR = Locally weighted regression; RF = Random forests; GAM = Generalized additive model. The rows are sorted in decreasing order of values in column *Success* that present a percentage of the successful estimates, averaged across all regression models. The ordering therefore ranks the domains by their difficulty of the reliability estimation task with the nine estimates at hand.

Locally weighted regression (LWR): Local regression with Gaussian kernel for weighting examples according to their distance;

Random forests (RF): Random forests (Breiman, 2001) with 100 trees;

Generalized additive model (GAM): Linear model (Hastie & Tibshirani, 1990; Wood, 2006) with no special parameters.

For testing we used 28 standard benchmark data sets, well known across the whole machine learning community. Each data set is a regression problem. The application domains vary from medical, ecological and technical to mathematical and physical domains. Most of the data sets are available from UCI Machine Learning Repository (Asuncion & Newman, 2007) and from StatLib DataSets Archive (Department of Statistics at Carnegie Mellon University, 2005). All data sets are available from authors upon request. A brief description of data sets is given in Table 2.

5 Internal cross-validation

We adapted the basic cross-validation approach (Schaffer, 1993) to select the most appropriate reliability estimate instead of the most appropriate learning algorithm. The proposed approach is described in the following.

Table 2 Basic characteristics of testing data sets

Data set	No. of examples	No. of discrete attributes	No. of continuous attributes
autoprice	159	1	14
auto93	93	6	16
autohorse	203	8	17
baskball	96	0	4
bodyfat	252	0	14
brainsize	20	0	8
breasttumor	286	1	8
cloud	108	2	4
cpu	209	0	6
diabetes	43	0	2
echomonths	130	3	6
elusage	55	1	1
fishcatch	158	2	5
fruitfly	125	2	2
grv	123	0	3
hungarian	294	7	6
lowbwt	189	7	2
mbagrade	61	1	1
pharynx	195	4	7
pollution	60	0	15
pwlinear	200	0	10
pyrim	74	0	27
servo	167	2	2
sleep	58	0	7
transplant	131	0	2
triazines	186	0	60
tumor	86	0	4
wpbc	198	0	32

5.1 Definitions

The internal cross-validation approach divides the learning examples to n equally sized subsets, as in the standard cross-validation approach. Each subset (selection set) is used for performance evaluation (correlation to the prediction error) of all testing reliability estimates. Based on acquired n correlation coefficients for each reliability estimate (one acquired between the estimate and the prediction error on each of the n subsets), the final (most appropriate) reliability estimate is selected as the one with the highest average correlation. This estimate is then used to estimate the reliability of all testing examples in that particular model and domain. The pseudocode of the procedure is shown in Figure 3.

5.2 Empirical evaluation

The testing was performed by correlating the reliability estimate, chosen by internal cross-validation procedure, with the prediction error of examples. The reliability estimate and the prediction errors were computed using the leave-one-out cross-validation procedure, that is, for each learning example that was left out in the current iteration. Having completed all iterations, the leave-one-out errors and reliability estimates were, therefore, computed for all available examples. The performance of the reliability estimates was measured by computing the Pearson correlation coefficient between the reliability estimate and the prediction error. The significance of the correlation coefficient was afterwards statistically evaluated using t-test for correlation coefficients with significance level $\alpha \leqslant 0.05$.

```
Input: data set Data = {(x_1, C_1), (x_2, C_2), ..., (x_n, C_n)}
Output: E (optimal estimate on Data)
PROGRAM ICV
    divide Data to subsets {S_1, ..., S_10}
    FOR outer_loop = 1 TO 10     // 10-fold CV
        FOR EACH (x_i, C_i) ∈ S_outer_loop // leave-one-out
            compute all reliability estimates for (x_i, C_i)
        END FOR EACH
        compute correl. coeff. on S_outer_loop for all estimates
    END FOR
    compute averages of estimates' correl. coeff. across 10 folds
    select the estimate with the highest average correlation E
END ICV
```

Figure 3 The pseudocode for the internal cross-validation testing

Table 3 The performance comparison of the most successful individual estimate *BVCK* and the estimate, selected with the internal cross-validation

Model	BVCK (+/−)	Internal cross-validation (+/−)
RT	71/4	87/0
LR	50/0	73/0
NN	54/0	73/0
BAG	61/0	67/0
SVM	46/0	67/0
LWR	43/0	73/0
RF	50/0	60/0
GAM	54/0	80/0
Average	54/1	**73/0**

RT = Regression trees; LR = Linear regression; NN = Neural networks;
BAG = Bagging; SVM = Support vector machines; LWR = Locally weighted regression;
RF = Random forests; GAM = Generalized additive model.
The table shows the percentage of experiments exhibiting significant positive/negative correlations between the *reliability estimates* and the *prediction error*.

The summarized testing results are shown in Table 3 and the detailed results in Table 8 in the Appendix. Due to the extensive time complexity of nested cross-validation and the computation of all reliability estimates, it was not timely feasible to evaluate this approach on all available testing domains. The approach was, therefore, tested on the subset, consisting of 15 domains (brainsize, diabetes, elusage, sleep, pollution, mbagrade, pyrim, tumor, auto93, baskball, cloud, grv, fishcatch, autoprice, servo). To achieve representative results, these domains were approximately uniformly sampled from the ordering in Table 1 to include the domains with higher as well as with lower coverage by the reliability estimates.

The results show that the automatic selection of reliability estimates using the internal cross-validation achieved better performance than any of the individual testing estimates. We can see that on average, with the internal cross-validation the 73% of estimates significantly positively correlated with the prediction error, while 0% of estimates significantly negatively correlated with the prediction error. With the aim to compare this approach to another, possibly less time-demanding, we also propose a meta-learning approach, presented in the next section.

6 Meta-learning in domain/model problem space

To develop an approach for automatic prediction of the most appropriate reliability estimate, we also adapted the basic meta-learning approach (Aha, 1992; Gama & Brazdil, 1995). The proposed approach, which performs the estimate selection on the given domain and on the model basis, is defined and empirically evaluated in the following.

6.1 Definitions

Within our meta-learning approach we propose a meta-classifier, which is intended to predict the optimal reliability estimate for a given *problem domain*/*regression model* pair. The reliability estimates therefore represent class values to be predicted, the learning set consists of optimal selections for known domain/model pairs, and the attributes describe the domain/model characteristics that are related to a particular example.

Within this meta-learning task, a challenge arises on how to define the suitable meta-attributes. Since this task differs from meta-selecting the predictive model, we cannot rely on the fact that the same meta-attributes would be suitable for our task, as well. In addition, since the reliability estimates are model-independent, we cannot use attributes that extract specific model features, but can focus only on more general model description (i.e. performance of the model on particular domain). With this objective, we defined and proposed the following seven attributes for this meta-learning task:

- **Regression model** (discrete nominal attribute), used on a problem domain (in the following referred to as *model*);
- **Number of learning examples** in a given domain (*no.examples*);
- **Number of attributes** in a given domain (*no.attr*);
- **Relative mean squared error**, achieved by the given regression model on a domain using the tenfold cross-validation (*cv.rmse*);
- **Average density** of the problem space, estimated by Parzen windows and sampled in points, given by learning examples (*avg.dens*);
- **Average distance to the five nearest neighbors**, averaged across all learning examples (*avg.DA*);
- **Average difference between the prediction for a given example and the predictions for the five nearest neighbors**, averaged across all learning examples (*avg.DK*).

We can see that the attributes *model*, *cv.rmse* and *avg.DK* describe the regression model used, while the others provide a description of the data.

To each meta-learning example we assigned a class value, representing a reliability estimate that achieved maximum positive correlation with the prediction error (irrespective of whether the correlation was statistically significant or not) for a given regression model. The correlation coefficients between estimate values and prediction errors were computed using a leave-one-out scenario on each particular data set. The possible class values therefore were: *SAvar*, *SAbias-s*, *SAbias-a*, *CNK-s*, *CNK-a*, *LCV*, *BAGV*, *DENS* and *BVCK*.

6.2 Empirical evaluation

Since we experimented with 28 domains and 8 regression models, we, therefore, formed the meta-learning set consisting of $28 \times 8 = 224$ learning examples. To assure the unbiasedness of the testing procedure, in every experiment we removed the example from the meta-learning set that represented that domain/model combination, for which we were meta-predicting the optimal reliability estimate.

Meta-classifier. Using the above set of the meta-learning examples, we constructed a decision tree meta-classifier for prediction of the most appropriate reliability estimate. We decided to use the decision tree as the meta-model due to its interpretability, which helps us understand the decision on when the use of each estimate is more appropriate. An example of such decision tree (constructed on all 224 meta-learning examples, without removing any) is shown in Figure 4 (the tree was pruned using the 1-SE rule and the cost-complexity pruning algorithm (Breiman *et al.*, 1984; Torgo, 2003)).

Although knowing that the performance of reliability estimates varies with different regression models, the results are rather interesting. We can see that the decision tree in Figure 4 shows that the most important attribute (more important than regression model), which constitutes the root

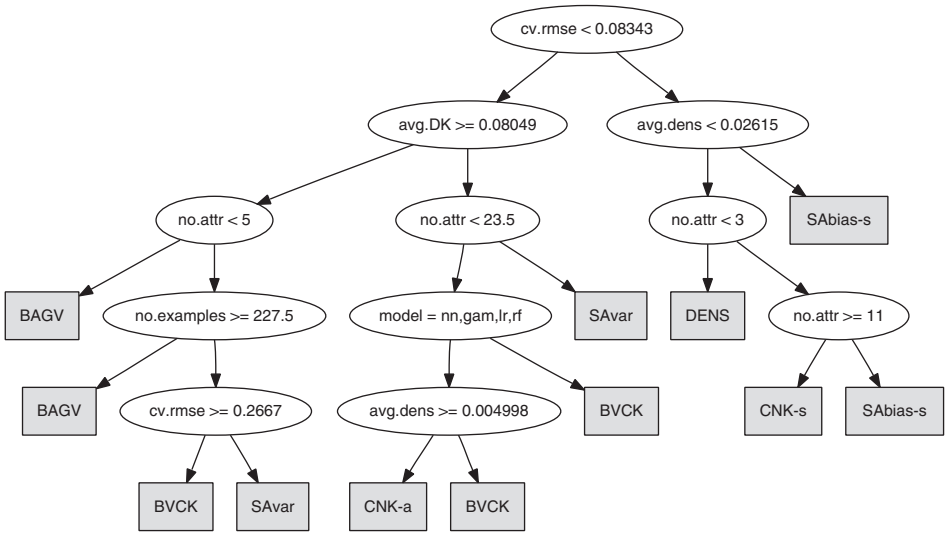

Figure 4 Meta-decision tree for prediction of the most appropriate reliability estimate

node of the tree, is the RMSE, achieved by the given regression model on a domain using the tenfold cross-validation (*cv.rmse*). The tree shows that the estimates *BAGV*, *BVCK*, *SAvar* and *CNK-a* better perform with more accurate regression models, while the estimates *DENS*, *CNK-s* and *SAbias-s* better perform with less accurate regression models.

The testing results of the automatic selection of the best performing estimate using the meta-learning approach are shown in Table 4. The detailed results are shown in Table 7 in the Appendix. From comparison with the most successful (the largest number of significant positive correlations with the prediction error) individual reliability estimate *BVCK* we can see, that we achieved better average results using the meta-learning approach. However, compared to the results of the internal cross-validation approach (Table 3) we can see that the latter approach outperforms the meta-learning approach. Nevertheless, being less time-demanding and achieving better results than the most successful individual estimate *BVCK*, the proposed meta-learning approach shows the potential for predicting the domain/model-based optimal reliability estimate.

Note that the present meta-learning approach requires further study (see Section 8). Besides evaluating stability and properties of the current meta-model, performance of other classifiers shall also be tested and evaluated. The current meta-learning problem shall also be optimized in terms of proposing/selecting more appropriate meta-attributes, using larger set of meta-examples and selecting the most appropriate set of target classes (i.e. set of reliability estimates).

7 Application on a real domain

The proposed methodology for automatic selection of the most appropriate estimate was preliminarily tested in a real domain. The data consisted of 1035 breast cancer patients, who had surgical treatment for cancer between 1983 and 1987 in the Clinical Center in Ljubljana, Slovenia. The patients were described using standard prognostic factors for breast cancer recurrence. The goal of the research was to predict the time of possible cancer recurrence after the surgical treatment. The analysis showed that this is a difficult prediction problem, because the possibility for recurrence is continuously present for almost 20 years after the treatment. Furthermore, the data present a mixture of two prediction problems, which additionally hinders the learning performance: (i) yes/no classification problem, whether the illness will recur at all, and (ii) the regression problem for the prediction of the recurrence time. In our study, the bare recurrence predictions were therefore complemented with our reliability estimates, helping the doctors with the additional validation of the predictions' accuracies.

Table 4 Performance comparison of the most successful individual estimate *BVCK* and the meta-predicted optimal estimate

Model	BVCK (+/−)	Meta-learning (+/−)
RT	71/4	79/0
LR	50/0	54/0
NN	54/0	46/0
BAG	61/0	61/0
SVM	46/0	54/0
LWR	43/0	43/4
RF	50/0	64/0
GAM	54/0	54/0
Average	54/1	**57/1**

RT = Regression trees; LR = Linear regression; NN = Neural networks;
BAG = Bagging; SVM = Support vector machines; LWR = Locally weighted regression;
RF = Random forests; GAM = Generalized additive model.
The table shows the percentage of experiments exhibiting significant positive/negative correlations between the *reliability estimates* and the *prediction error*.

After the performance comparison of various regression models, the locally weighted regression was chosen for the use with this prediction problem, due to its low RMSE. For test examples, all nine available reliability estimates were computed, achieving the correlation to the prediction error as shown in Table 5. The statistical evaluation of the correlation coefficients revealed that only the estimate *BVCK* significantly positively correlates to the prediction error. This indicates that the oncological problem is a difficult domain for estimation of prediction reliability, in terms of domain ordering as shown in Table 1.

On this particular problem, we tested both of our proposed methods for automatic selection of the most appropriate reliability estimate. Both, the internal cross-validation approach, as well as the meta-learning approach, selected the *BVCK* as the most appropriate estimate, which also turned out to be the optimal estimate for this problem, according to the results in Table 5. This result, therefore, shows the potential of the proposed methodology in practice.

8 Discussion

The testing of the individual reliability estimates, developed in our previous work, exhibited different potentials for the usage of different estimates with particular regression models. In addition, the results also showed that the success of the chosen reliability estimate may also be domain-dependent.

To deal with this problem, in the paper we proposed and tested two novel approaches for automatic selection of the most appropriate estimate for a given domain and regression model: the internal cross-validation approach and the meta-learning approach. With both approaches, the selection of estimates was performed from the set of nine estimates, based on various approaches, that is sensitivity analysis, variance of bagged models, local cross-validation, density-based estimation and local error estimation.

We implemented the internal cross-validation approach, which selects the most appropriate reliability estimate based on a subset of available examples. For the purpose of testing, we performed the partitioning as in standard cross validation, thus nesting the inner leave-one-out procedure (for computation of reliability estimates and the prediction error) in an outer tenfold cross-validation loop.

Next, we showed how reliability estimate selection can be turned into a meta-learning task by proposing meta-attributes and defining other meta-space parameters. By building a meta-decision tree classifier, we used it to predict a most appropriate reliability estimate for various domain and model combinations.

Table 5 The correlation coefficients and their significance levels between the *reliability estimates* and the *prediction error* on the oncological application domain

	Correlation coefficient	Significance level α
SAvar	−0.046	0.140
SAbias-s	0.033	0.298
SAbias-a	0.001	0.985
BAGV	0.059	0.060
LCV	0.008	0.793
DENS	−0.074	0.018
CNK-s	−0.042	0.179
CNK-a	0.034	0.282
BVCK	**0.071**	**0.024**

We compared the performance of both procedures for automatic selection of estimate, as well as to the performance of each individual estimate. The empirical results of both approaches showed the advantage of dynamically selected reliability estimate for a given domain/model pair, when compared to individual reliability estimates in terms of higher positive correlation to the prediction error. The best results were achieved using the internal cross-validation procedure. With this approach, the selected reliability estimates significantly positively correlated with the prediction error in 73% of domains and significantly negatively correlated with the prediction error in none.

Figure 5 displays a performance comparison of individual estimates and both approaches for the most appropriate estimate selection. The graphs show the percentage of experiments with the significant positive correlations (desired result) and the percentage of experiments with the significant negative correlation (undesired). We can see that internal cross-validation performed better than any of the other estimates in seven out of eight regression models. Meta-learning approach has empirically proven to perform good as well, since its performance was in top half of the ranking with all used regression models. The achieved results indicate that it is reasonable to approach the problem of automatic estimate selection using the proposed approaches.

Comparison of both approaches. By analyzing how often was each of nine reliability estimates selected (in percent), as shown in Table 6, we can compare both of the proposed approaches. The comparison shows that the internal cross-validation approach most frequently selected such estimates that also perform well in *specific* domains (this can be seen by comparing Table 8 to detailed performance results of individual estimates, which cannot be presented here due to their extent—approximately 10 pages. The results are available from the authors on request). In contrast, the meta-learning approach most frequently selected the estimates that *generally* perform well also as individual reliability estimates, that is *BVCK*, *BAGV* and *CNK-a*.

We explain this phenomenon based on the main characteristics of both the approaches, as follows:

1. The proposed meta-learning approach is based on modeling of the dependence between the domain/model properties and the reliability estimate. By using the available meta-attributes, the meta-classifier's predictive power is, therefore, limited by their descriptiveness of the particular domain and model. In contrast to this, the internal cross-validation operates directly with the available data and tests the candidate estimates on the subset of examples. If we assume that learning and test examples originate from the same distribution, it is likely that the estimate that correlates best with the prediction error on a subset of examples will also correlate well on the rest of the examples. This therefore enables the internal cross-validation approach to select estimates that are more *tailored* to the domain itself.
2. The purpose of the meta-learning approach is to induce a generalized rule for selection of the most appropriate estimate. This fact implies that, for preserving the generality of the rule (avoiding overfitting), the predicted estimates would most often be the ones that also perform well on average.

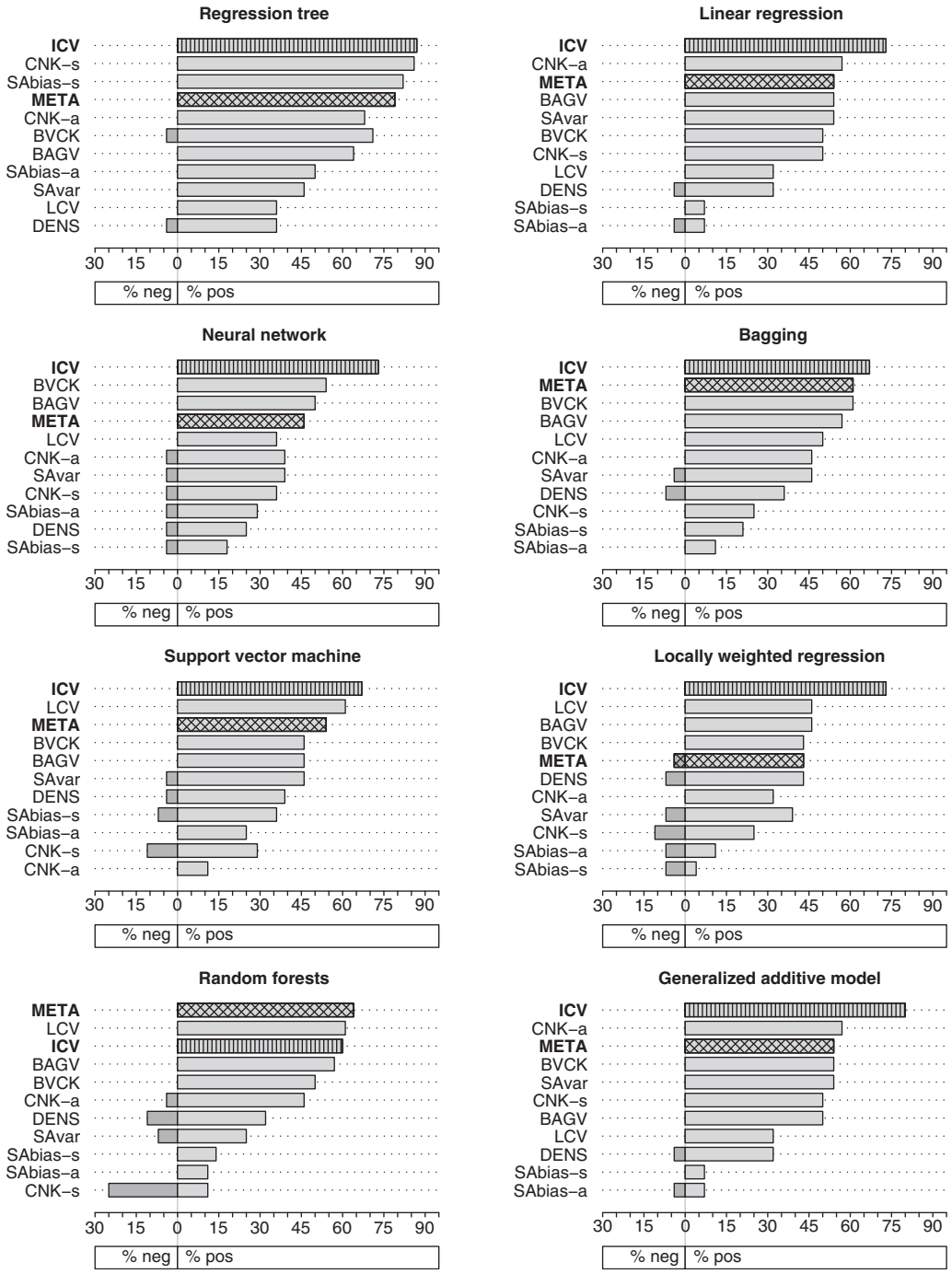

Figure 5 Comparison (ranking) of meta-learning (META) and internal cross-validation (ICV) performance to the performance of individual reliability estimates. Graphs show the percent of experiments with the significant positive and negative correlations with the prediction error

Time complexity of both approaches. The experiments showed that the internal cross-validation approach turned out to be more demanding than the meta-learning approach. Since reliability estimates are independent of a particular regression model, we discuss the time complexity of both approaches in terms of the number of regression models they require to be constructed.

The design of the internal cross-validation approach shows that nine reliability estimates have to be computed in the two nested loops (10 times leave-one-out procedure), out of which the most demanding estimate *BAGV* requires constructing 50 regression models for each example.

Table 6 The percent of each automatically selected reliability estimates using the meta-learning and internal cross-validation approach

	Meta-learning	Internal cross-validation
SAvar	11	12
SAbias-s	14	5
SAbias-a	0	3
BAGV	15	10
LCV	2	17
DENS	8	8
CNK-s	11	22
CNK-a	19	9
BVCK	20	14

Considering other reliability estimates (SAvar and SAbias require ten models, LCV five and CNK one), this therefore means that for testing domain with 100 examples, at least 6600 models would need to be constructed.

We can discuss the complexity of the meta-learning approach from the following two perspectives: the complexity of generating the meta-learning set (which depends on the proposed meta-attributes) and the complexity of modeling and predicting the final estimate. The former requires computing six meta-attributes, which include computing distances, local models and prediction errors. Among these attributes, the computation of the RMSE is the most time-consuming (tenfold cross-validation requires constructing ten models). Not considering computing distances and auxiliary costs, this means that for computation of the meta-learning set of size 100, we would require construction of 1000 models. In contrast, the other complexity perspective of the meta-learning approach (generating the final model and predicting the most appropriate estimate) requires inducing of only one model.

The case of our study shows that the greater time complexity of the internal cross-validation approach clearly pays off with its better performance.

Further work. The achieved results in the field of estimating reliability for individual predictions offer the challenges for further work, which includes:

- A good correlation of signed reliability estimates (*SAbias-s* and *CNK-s*) with the signed prediction error implies the potential for the usage of reliability estimates for the correction of regression predictions. We shall, therefore, explore whether these two reliability estimates can be utilized to reduce the error of regression predictions.
- Different performances of reliability estimates in different testing domains (see Table 1) indicate that the potential for estimation of prediction reliability is in some domains more feasible than in the others. The domain characteristics, which lead to a good performance of reliability estimates, shall be analyzed in more detail.
- The meta-learning approach shall be further analyzed by employing other predictive models, proposing new attributes and evaluating their quality.
- The preliminary experiments with meta-learning have shown that the selection of a regression model with low RMSE and the selection of a model on which reliability estimates will perform well, is a trade-off criteria. This phenomenon shall be analyzed in further theoretical and empirical work.

Appendix A: Detailed results

Table 7 Automatically selected *reliability estimates using the meta-predictor* and their correlation coefficients with the *prediction error*

	Meta-predicted reliability estimate							
	RT	LR	NN	BAG	SVM	LWR	RF	GAM
autoprice	**COMB**	**CNK-a**	**COMB**	**COMB**	**COMB**	**COMB**	**COMB**	**CNK-a**
	0.534	**0.415**	**0.413**	**0.642**	**0.267**	**0.383**	**0.460**	**0.415**
auto93	**COMB**	**SAvar**	**SAvar**	**COMB**	**SAvar**	**SAvar**	**COMB**	**CNK-a**
	0.378	**0.251**	**0.236**	**0.276**	**0.379**	**0.290**	**0.454**	**0.505**
autohorse	**SAvar**	**COMB**	**COMB**	**SAvar**	**SAvar**	COMB	**CNK-a**	**COMB**
	0.275	**0.436**	**0.591**	**0.266**	**0.549**	0.089	**0.182**	**0.451**
baskball	**CNK-s**	CNK-s	CNK-s	CNK-s	CNK-s	CNK-s	**CNK-s**	CNK-s
	0.432	0.085	0.074	0.105	0.098	0.058	**0.207**	0.085
bodyfat	**BAGV**	**COMB**	**COMB**	**BAGV**	**BAGV**	COMB	**BAGV**	**COMB**
	0.275	**0.513**	**0.522**	**0.327**	**0.524**	0.090	**0.415**	**0.515**
brainsize	CNK-s	SAbias-s	BAGV	SAbias-s	SAbias-s	BAGV	SAbias-s	BAGV
	−0.175	−0.171	−0.336	0.186	−0.328	−0.078	0.110	0.170
breasttumor	**SAbias-s**	LCV	SAbias-s	SAbias-s	LCV	CNK-a	SAbias-s	BAGV
	0.151	0.082	−0.014	0.114	−0.013	−0.040	0.101	−0.027
cloud	**BAGV**	**BAGV**	**BAGV**	**BAGV**	**SAvar**	**COMB**	**BAGV**	**BAGV**
	0.455	**0.319**	**0.269**	**0.451**	**0.664**	**0.302**	**0.551**	**0.343**
cpu	**CNK-a**	**CNK-a**	**CNK-a**	**COMB**	**COMB**	**COMB**	**CNK-a**	**CNK-a**
	0.679	**0.753**	**0.754**	**0.782**	**0.707**	**0.375**	**0.656**	**0.753**
diabetes	**DENS**	**DENS**	**DENS**	**DENS**	**DENS**	**DENS**	DENS	**DENS**
	0.375	**0.419**	**0.425**	**0.414**	**0.451**	**0.439**	0.225	**0.419**
echomonths	**SAbias-s**	CNK-a	COMB	**COMB**	SAbias-s	**COMB**	**BAGV**	CNK-a
	0.355	0.134	0.084	**0.197**	0.056	**0.214**	**0.184**	0.134
elusage	**BAGV**	BAGV	SAvar	**COMB**	DENS	**SAbias-s**	**CNK-a**	BAGV
	0.390	0.050	−0.070	**0.328**	0.261	**−0.368**	**0.318**	−0.104
fishcatch	**CNK-a**	**CNK-a**	**CNK-a**	**COMB**	**COMB**	**COMB**	**CNK-a**	**CNK-a**
	0.668	**0.498**	**0.502**	**0.821**	**0.715**	**0.701**	**0.324**	**0.498**
fruitfly	**SAbias-s**	SAbias-s	SAbias-s	SAbias-s	LCV	SAbias-s	**SAbias-s**	BAGV
	0.219	0.045	−0.039	0.078	−0.079	−0.056	**0.246**	−0.136
grv	**CNK-a**	CNK-a	CNK-a	**COMB**	COMB	COMB	**BAGV**	CNK-a
	0.303	0.108	0.158	**0.242**	0.020	0.023	**0.189**	0.108
hungarian	**COMB**	**COMB**	**CNK-s**	**BAGV**	**SAvar**	**COMB**	**BAGV**	**COMB**
	0.470	**0.331**	**0.469**	**0.629**	**0.319**	**0.401**	**0.588**	**0.330**
lowbwt	BAGV	SAvar	SAvar	**BAGV**	**SAvar**	COMB	CNK-a	SAvar
	0.090	−0.021	0.000	**0.276**	**0.205**	0.076	0.107	−0.021
mbagrade	**SAbias-s**	DENS	DENS	SAbias-s	DENS	SAbias-s	**SAbias-s**	DENS
	0.263	−0.078	−0.077	0.102	−0.035	−0.049	**0.283**	−0.078
pharynx	**COMB**	**SAvar**	CNK-a	**COMB**	SAvar	**BAGV**	**COMB**	**SAvar**
	0.457	**0.171**	−0.086	**0.311**	0.125	**0.203**	**0.228**	**0.171**
pollution	**CNK-s**	**CNK-s**	CNK-s	BAGV	CNK-s	BAGV	CNK-s	**CNK-s**
	0.483	**0.488**	0.189	−0.111	0.071	−0.130	0.097	**0.488**
pwlinear	**COMB**	**SAvar**	**BAGV**	**COMB**	**SAvar**	**COMB**	COMB	**SAvar**
	0.286	**0.229**	**0.287**	**0.140**	**0.287**	**0.306**	0.126	**0.229**
pyrim	**CNK-a**	**CNK-a**	**CNK-a**	DENS	**SAvar**	**CNK-a**	CNK-a	**CNK-a**
	0.307	**0.849**	**0.636**	0.024	**0.369**	**0.797**	0.037	**0.849**
servo	DENS	SAvar	SAvar	**BAGV**	SAvar	**BAGV**	**BAGV**	SAbias-s
	−0.100	0.006	0.125	**0.608**	0.103	**0.760**	**0.765**	0.095
sleep	LCV	SAbias-s	**SAbias-s**	CNK-s	**SAbias-s**	SAbias-s	CNK-s	SAbias-s
	0.238	0.031	**0.358**	0.186	**0.281**	0.202	0.037	0.058
transplant	**DENS**	**DENS**	**DENS**	**DENS**	**COMB**	COMB	**CNK-a**	**CNK-a**
	0.509	**0.482**	**0.473**	**0.474**	**0.755**	0.075	**0.507**	**0.387**
triazines	**CNK-s**	**CNK-s**	CNK-a	CNK-s	CNK-s	CNK-a	**BAGV**	**CNK-s**
	0.362	**0.513**	0.135	−0.043	0.064	0.071	**0.469**	**0.513**
tumor	SAbias-s	SAbias-s	BAGV	SAbias-s	**CNK-s**	SAbias-s	SAbias-s	BAGV
	0.097	0.040	−0.042	0.108	**0.238**	0.033	0.150	0.016

Table 7 *(Continued)*

	Meta-predicted reliability estimate							
	RT	LR	NN	BAG	SVM	LWR	RF	GAM
wpbc	CNK-a 0.039	CNK-a 0.053	CNK-a 0.019	CNK-a 0.008	CNK-a 0.094	CNK-a 0.111	CNK-a 0.003	CNK-a 0.053
+:	**79%**	**54%**	**46%**	**61%**	**54%**	**43%**	**64%**	**54%**
−:	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**

RT = Regression trees; LR = Linear regression; NN = Neural networks; BAG = Bagging; SVM = Support vector machines; LWR = Locally weighted regression; RF = Random forests; GAM = Generalized additive model. Statistically significant values ($\alpha \leqslant 0.05$) are denoted with boldface. Significant negative values (undesired negative correlations) are additionally underlined.

Table 8 Automatically selected *reliability estimates using internal cross-validation* and their correlation coefficients with the *prediction error*

	Reliability estimate selected with the internal cross-validation							
	RT	LR	NN	BAG	SVM	LWR	RF	GAM
autoprice	**CNK-a** **0.624**	**COMB** **0.520**	**COMB** **0.482**	**COMB** **0.519**	**SAvar** **0.517**	**SAvar** **0.537**	**CNK-a** **0.559**	**COMB** **0.438**
auto93	**CNK-s** **0.525**	**CNK-a** **0.465**	**COMB** **0.522**	**CNK-s** **0.335**	**SAvar** **0.287**	**LCV** **0.436**	**COMB** **0.442**	**CNK-a** **0.537**
autohorse	**CNK-s** **0.381**	SAvar 0.120	SAvar 0.087	CNK-s 0.059	CNK-s 0.137	CNK-s −0.011	CNK-s 0.176	**LCV** **0.236**
baskball	LCV 0.017	CNK-s 0.347	SAbias-s 0.378	SAbias-a −0.193	DENS −0.152	**CNK-s** **0.472**	SAbias-a 0.254	CNK-s 0.349
bodyfat	**BAGV** **0.461**	**SAvar** **0.400**	**SAvar** **0.507**	**DENS** **0.538**	**SAvar** **0.639**	**SAvar** **0.654**	**DENS** **0.518**	**SAvar** **0.364**
brainsize	**DENS** **0.415**	**DENS** **0.409**	**DENS** **0.316**	**DENS** **0.420**	**DENS** **0.354**	**CNK-s** **0.375**	DENS 0.174	**DENS** **0.382**
breasttumor	**COMB** **0.426**	**LCV** **0.279**	**BAGV** **0.397**	**COMB** **0.403**	**SAvar** **0.537**	**CNK-s** **0.430**	**COMB** **0.429**	**LCV** **0.417**
cloud	**COMB** **0.658**	**CNK-a** **0.515**	**CNK-a** **0.482**	**COMB** **0.760**	**COMB** **0.749**	**COMB** **0.635**	**COMB** **0.560**	**CNK-a** **0.518**
cpu	**SAbias-s** **0.346**	**CNK-s** **0.235**	**CNK-s** **0.285**	**LCV** **0.270**	**LCV** **0.202**	**BAGV** **0.241**	**SAvar** **0.196**	**CNK-s** **0.333**
diabetes	SAbias-s 0.165	LCV 0.018	SAbias-s −0.137	LCV 0.105	SAvar 0.047	**BAGV** **0.365**	**BAGV** **0.256**	SAbias-a 0.139
echomonths	**CNK-s** **0.457**	**CNK-s** **0.650**	**CNK-s** **0.270**	CNK-s 0.137	SAbias-s 0.212	SAvar 0.043	CNK-a 0.117	**CNK-s** **0.505**
elusage	**CNK-s** **0.370**	**CNK-a** **1.000**	**COMB** **0.566**	**CNK-s** **0.273**	**COMB** **0.536**	**CNK-a** **0.786**	LCV −0.102	**CNK-a** **0.834**
fishcatch	**BAGV** **0.600**	**LCV** **0.514**	**LCV** **0.427**	**BAGV** **0.391**	**LCV** **0.748**	**BAGV** **0.700**	**BAGV** **0.833**	**LCV** **0.418**
fruitfly	**SAbias-s** **0.512**	**BAGV** **0.749**	**COMB** **0.535**	**LCV** **0.319**	**LCV** **0.345**	LCV 0.213	**LCV** **0.283**	**BAGV** **0.767**
grv	**CNK-s** **0.330**	CNK-s 00.022	CNK-s 0.129	LCV 0.195	BAGV 0.202	SAbias-a 0.157	LCV 0.203	CNK-s 0.200
+:	**87%**	**73%**	**73%**	**67%**	**67%**	**73%**	**60%**	**80%**
−:	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**	**0%**

RT = Regression trees; LR = Linear regression; NN = Neural networks; BAG = Bagging; SVM = Support vector machines; LWR = Locally weighted regression; RF = Random forests; GAM = Generalized additive model. Statistically significant values ($\alpha \leqslant 0.05$) are denoted with the boldface. There were no significant negative values (undesired negative correlations).

References

Aha, D. W. 1992. Generalizing from case studies: A case study. In *Proceedings of the Ninth International Workshop on Machine Learning (ML 1992)*, Aberdeen, Scotland, UK, 1–10.

Asuncion, A. & Newman, D. J. 2007. UCI machine learning repository, `http://www.ics.uci.edu/~mlearn/MLRepository.html`, Irvine, CA: University of California, School of Information and Computer Science.

Birattari, M., Bontempi, H. & Bersini, H. 1998. Local learning for data analysis. In *Proceedings of the 8th Belgian-Dutch Conference on Machine Learning*, Wageningen, The Netherlands, 55–61.

Blum, A. & Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, Madison, Wisconsin, 92–100.

Bosnić, Z. & Kononenko, I. 2007. Estimation of individual prediction reliability using the local sensitivity analysis. *Applied Intelligence* **29**(3), 187–203.

Bosnić, Z. & Kononenko, I. 2008a. Estimation of regressor reliability. *Journal of Intelligent Systems* **17**(1/3), 297–311.

Bosnić, Z. & Kononenko, I. 2008b. Comparison of approaches for estimating reliability of individual regression predictions. *Data & Knowledge Engineering* **67**(3), 504–516.

Bosnić, Z., Kononenko, I., Robnik-Šikonja, M. & Kukar, M. 2003. Evaluation of prediction reliability in regression using the transduction principle. In *Proceedings of Eurocon 2003*, Zajc, B. & Tkalčič, M. (eds), 99–103. IEEE (Institute of Electrical and Electronics Engineering, Inc.)

Bousquet, O. & Elisseeff, A. 2002. Stability and generalization. *Journal of Machine Learning Research* **2**, 499–526.

Breierova, L. & Choudhari, M. 1996. An introduction to sensitivity analysis. MIT System Dynamics in Education Project.

Breiman, L. 1996. Bagging predictors. *Machine Learning* **24**(2), 123–140.

Breiman, L. 2001. Random forests. *Machine Learning* **45**(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. 1984. *Classification and Regression Trees*. Wadsworth International Group.

Carney, J. & Cunningham, P. 1999. Confidence and prediction intervals for neural network ensembles. In *Proceedings of IJCNN'99, The International Joint Conference on Neural Networks*, Washington, USA, 1215–1218.

Caruana, R. 1997. Multitask learning. *Machine Learning* **28**(1), 41–75.

Chang, C. & Lin, C. 2001. *LIBSVM: A Library for Support Vector Machines*. Software available at `http://www.csie.ntu.edu.tw/cjlin/libsvm`.

Christiannini, N. & Shawe-Taylor, J. 2000. *Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Cohn, D. A., Atlas, L. & Ladner, R. 1990. Training connectionist networks with queries and selective sampling. In *Advances in Neural Information Processing Systems*, Touretzky, D. (ed.) **2**, 566–573. Morgan Kaufman.

Cohn, D. A., Ghahramani, Z. & Jordan, M. I. 1995. Active learning with statistical models. In *Advances in Neural Information Processing Systems*, Tesauro, G., Touretzky, D. & Leen, T. (eds) **7**, 705–712. The MIT Press.

Crowder, M. J., Kimber, A. C., Smith, R. L. & Sweeting, T. J. 1991. *Statistical Concepts in Reliability. Statistical Analysis of Reliability Data*. Chapman & Hall.

de Sa, V. 1993. Learning classification with unlabeled data. In *Proc. NIPS'93, Neural Information Processing Systems*, Cowan, J. D., Tesauro, G. & Alspector, J. (eds), 112–119. Morgan Kaufmann Publishers.

DesJardins, M. & Gordon Diana, F. 1995. Evaluation and Selection of Biases in Machine Learning. *Machine Learning* **20**, 5–22.

Department of Statistics at Carnegie Mellon University 2005. *Statlib – Data, Software and News from the Statistics Community*. `http://lib.stat.cmu.edu/`.

Elidan, G., Ninio, M., Friedman, N. & Schuurmans, D. 2002. Data perturbation for escaping local maxima in learning. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence*, July 28 - August 1, 2002, Edmonton, Alberta, Canada, 132–139. AAAI Press.

Freund, Y. & Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **55**(1), 119–139.

Gama, J. & Brazdil, P. 1995. Characterization of classification algorithms. In *Progress in Artificial Intelligence, 7th Portuguese Conference on Artificial Intelligence, EPIA-95*, Pinto-Ferreira, C. & Mamede, N. (eds), 189–200. Springer-Verlag.

Gammerman, A., Vovk, V. & Vapnik, V. 1998. Learning by transduction. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, Madison, Wisconsin, 148–155.

Giacinto, G. & Roli, F. 2001. Dynamic classifier selection based on multiple classifier behaviour. *Pattern Recognition* **34**(9), 1879–1881.

Goldman, S. & Zhou, Y. 2000. Enhancing supervised learning with unlabeled data. In *Proc. 17th International Conf. on Machine Learning*, Morgan Kaufmann, San Francisco, CA, 327–334.

Hastie, T. & Tibshirani, R. 1990. *Generalized Additive Models*. Chapman and Hall.

Heskes, T. 1997. Practical confidence and prediction intervals. In *Advances in Neural Information Processing Systems*, Mozer, M. C., Jordan, M. I. & Petsche, T. (eds), **9**, 176–182. The MIT Press.

Jeon, B. & Landgrebe, D. A. 1994. Parzen density estimation using clustering-based branch and bound. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 950–954.

Kearns, M. J. & Ron, D. 1997. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Computational Learning Theory*, Freund Y. & Shapire R. (eds), 152–162, Morgan Kaufmann.

Kleijnen, J. 2001. Experimental designs for sensitivity analysis of simulation models. *Tutorial at the Eurosim 2001 Conference*.

Kononenko, I. & Kukar, M. 2007. *Machine Learning and Data Mining: Introduction to Principles and Algorithms*. Horwood Publishing Limited.

Krieger, A. M. & Green, P. E. 1999. A cautionary note on using internal cross validation to select the number of clusters. *Psychometrika* **64**, 341–353.

Kukar, M. & Kononenko, I. 2002. Reliable classifications with machine learning. In *Proc. Machine Learning: ECML-2002*, Elomaa, T., Manilla, H. & Toivonen, H. (eds), 219–231. Springer-Verlag.

Li, M. & Vitányi, P. 1993. *An Introduction to Kolmogorov Complexity and its Applications*. Springer-Verlag.

Linden, A. & Weber, F. 1992. Implementing inner drive by competence reflection. In *Proceedings of the 2nd International Conference on Simulation of Adaptive Behavior*, Hawaii, 321–326.

Merz, C. J. 1996. Dynamical selection of learning algorithms. In *Learning from Data: Artificial Intelligence and Statistics*, Fisher, D. & Lenz, H. J. (eds), 1–10. Springer-Verlag.

Michie, D., Spiegelhalter, D. J. & Taylor, C. C. (eds) 1994. Analysis of results. In *Machine Learning, Neural and Statistical Classification*, 176–212. Ellis Horwood.

Mitchell, T. 1999. The role of unlabelled data in supervised learning. In *Proceedings of the 6th International Colloquium of Cognitive Science*, San Sebastian, Spain.

Nouretdinov, I., Melluish, T. & Vovk, V. 2001. Ridge regression confidence machine. In *Proc. 18th International Conf. on Machine Learning*, Morgan Kaufmann, San Francisco, CA, 385–392.

Pratt, L. & Jennings, B. 1998. A survey of connectionist network reuse through transfer. *Learning to Learn*, Norwell, MA, USA, ISBN: 0-7923-8047-9, 19–43. Kluwer Academic Publishers.

R Development Core Team 2006. *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. 1986. *Learning Internal Representations by Error Propagation*. MIT Press, 318–362.

Saunders, C., Gammerman, A. & Vovk, V. 1999. Transduction with confidence and credibility. In *Proceedings of IJCAI'99*, **2**, 722–726.

Schaal, S. & Atkeson, C. G. 1994. Assessing the quality of learned local models. In *Advances in Neural Information Processing Systems*, Cowan, J. D., Tesauro, G. & Alspector, J. (eds), 160–167. Morgan Kaufmann Publishers.

Schaal, S. & Atkeson, C. G. 1998. Constructive incremental learning from only local information. *Neural Computation* **10**(8), 2047–2084.

Schaffer, C. 1993. Selecting a classification method by cross-validation. In *Fourth International Workshop on Artificial Intelligence & Statistics*, 15–25.

Schmidhuber, J. & Storck, J. 1993. *Reinforcement Driven Information Acquisition in Nondeterministic Environments*. Technical Report. Fakultat fur Informatik, Technische Universit at Munchen.

Schmidhuber, J, Zhao, J. & Wiering, M. 1996. *Simple principles of metalearning*, Technical Report IDSIA-69-96, Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale, 1–23.

Seeger, M. 2000. *Learning with Labeled and Unlabeled Data*. Technical report. `http://www.dai.ed.ac.uk/seeger/papers.html`.

Silverman, B. W. 1986. *Density Estimation for Statistics and Data Analysis. Monographs on Statistics and Applied Probability*. Chapman and Hall.

Smola, A. J. & Schölkopf, B. 1998. *A Tutorial on Support Vector Regression*. NeuroCOLT2 Technical Report NC2-TR-1998-030.

Tibshirani, R. & Knight, K. 1999. Model search and inference by bootstrap bumping. *Journal of Computational and Graphical Statistics* **8**, 671–686.

Torgo, L. 2003. *Data Mining with R: Learning by Case Studies*. University of Porto, LIACC-FEP.

Tsuda, K., Rätsch, G., Mika, S. & Müller, K. 2001. Learning to predict the leave-one-out error of kernel based classifiers. In *Lecture Notes in Computer Science*, 227–331. Springer Berlin/Heidelberg.

Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. Springer.

Vilalta, R. & Drissi, Y. 2002. A perspective view and survey of metalearning. *Artificial Intelligence Review* **18**(2), 77–95.

Wand, M. P. & Jones, M. C. 1995. *Kernel Smoothing*. Chapman and Hall.

Weigend, A. & Nix, D. 1994. Predictions with confidence intervals (local error bars). In *Proceedings of the International Conference on Neural Information Processing (ICONIP'94)*, Seoul, Korea, 847–852.

Whitehead, S. D. 1991. A complexity analysis of cooperative mechanisms in reinforcement learning. In *AAAI*, 607–613.

Wolpert, D. H. 1992. Stacked generalization. In *Neural Networks*, Amari S. Grossberg S. & Taylor J. G. (eds) **5**, 241–259. Pergamon Press.

Wood, S. N. 2006. *Generalized Additive Models: An Introduction with R*, Chapman & Hall/CRC.

Woods, K., Kegelmeyer, W. P. & Bowyer, K. 1997. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on PAMI* **19**(4), 405–410.