Data & Knowledge Engineering 67 (2008) 504–516

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

Comparison of approaches for estimating reliability of individual regression predictions

Zoran Bosnić *, Igor Kononenko

University of Ljubljana, Faculty of Computer and Information Science, Tržaška 25, Ljubljana, Slovenia

ARTICLE INFO

Article history:
Received 2 April 2008
Received in revised form 5 August 2008
Accepted 11 August 2008
Available online 22 August 2008

PACS:
07.05.Kf
07.05.Mh

Keywords:
Reliability estimate
Regression
Sensitivity analysis
Prediction accuracy
Prediction error

ABSTRACT

The paper compares different approaches to estimate the reliability of individual predictions in regression. We compare the sensitivity-based reliability estimates developed in our previous work with four approaches found in the literature: variance of bagged models, local cross-validation, density estimation, and local modeling. By combining pairs of individual estimates, we compose a combined estimate that performs better than the individual estimates. We tested the estimates by running data from 28 domains through eight regression models: regression trees, linear regression, neural networks, bagging, support vector machines, locally weighted regression, random forests, and generalized additive model. The results demonstrate the potential of a sensitivity-based estimate, as well as the local modeling of prediction error with regression trees. Among the tested approaches, the best average performance was achieved by estimation using the bagging variance approach, which achieved the best performance with neural networks, bagging and locally weighted regression.

1. Introduction

When using supervised learning for modeling data, we aim to achieve the best possible prediction accuracy for the unseen examples which were not included in the learning process [1]. The most common measures for evaluating prediction accuracy are averaged accuracy measures such as the mean squared error (MSE) and the relative mean squared error (RMSE). Although these estimates evaluate model performance by summarizing the error contributions of all test examples, they provide no local information about the expected error of an individual prediction for a given unseen example. Having information about single prediction reliability [2] at our disposal could be an important benefit in risk-sensitive areas where acting upon predictions may have financial or medical consequences (e.g. medical diagnosis, stock market, navigation, control applications). For example, in medical diagnosis, physicians are not interested only in the average accuracy of the predictor, but expect the system to provide a prediction as well as an estimate of its reliability. Since the averaged accuracy measures do not fulfill this requirement, reliability measures for individual predictions are needed in this area.

Various methods have been developed to enable the users of classification and regression models to gain more insight into the reliability of individual predictions. Some of these methods were focused on extending formalizations of existing classification and regression models, enabling them to make predictions with their adjoined reliability estimates. Another group of methods focused on the development of model-independent approaches, which are more general but also harder to analytically evaluate with individual models.

* Corresponding author. Tel.: +386 14768459; fax: +386 14768498.
E-mail addresses: zoran.bosnic@fri.uni-lj.si (Z. Bosnić), igor.kononenko@fri.uni-lj.si (I. Kononenko).

0169-023X/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.datak.2008.08.001

Since some of the approaches from the first group were able to make use of probabilistic and model-specific properties, reliability estimates in this area were defined with a probabilistic interpretation, which makes them *confidence measures*. Confidence measures have values spanning the interval [0,1], where 0 represents the confidence of the most unreliable prediction and 1 the confidence of the most reliable one. Since estimates developed using other approaches do not necessarily have a probabilistic interpretation (and values are not confined to the interval [0,1]), we use the more general term *reliability estimate* to refer to measures that provide information about our level of trust in the accuracy of an individual prediction.

In this paper we compare five approaches to model-independent reliability estimation for individual examples in regression. We mostly focus on comparing the sensitivity-based estimates developed in our previous work to four other approaches implemented using ideas from related work. We evaluate the performance of selected reliability estimates on 28 testing domains 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. With the aim of improving the performance of the estimates, we test the performance of combined (averaged) pairs of estimates.

This paper is organized as follows. Section 2 summarizes previous work from related areas of individual prediction reliability estimation and Section 3 presents and proposes the reliability estimates. We describe the experiments and testing protocol, and interpret the results in Section 4. Section 5 provides conclusions and ideas for further work.

2. Related work

The most relevant related work is in the area of reliability estimation for individual examples. An appropriate criterion for differentiating between various approaches is whether they target a specific predictive model or whether they are model-independent. While the model-specific approaches are less general, they are usually founded on exact mathematical or probabilistic properties. Since the model-independent approaches are general, they cannot exploit parameters specific to a given predictive model (e.g. the sum of least squares in linear regression), but rather focus on influencing the parameters that are available in the standard supervised learning framework (e.g. the learning set and attributes). The reliability estimates based on these approaches are defined as metrics over the observed learning parameters. Since the reliability is not based on a particular probabilistic distribution and a chosen confidence interval, these metrics can take values from an arbitrary interval of numbers. As such, the metrics' values have no probabilistic interpretation.

The idea of reliability estimation for individual predictions originated in statistics, where confidence values and intervals are used to express the reliability of estimates. In machine learning, statistical properties of predictive models were used to extend predictions with reliability estimates. Research on support vector machines [3,4] introduced the notions of *confidence* and *credibility* and showed that the basic model can be successfully expanded with reliability estimates. Also focusing on particular predictive models, Nouretdinov et al. [5] expanded the ridge regression model, Weigend and Nix [6] the multilayer perceptron, and Heskes, Carney and Cunningham [7,8] the ensembles of neural networks. It is obvious that these approaches cannot be used with an arbitrary predictive model due to their definition, which is bound to the specific model formalism.

In contrast, the methods that are independent of the predictive model are also more generally applicable. These methods utilize approaches such as local modeling of prediction error based on input space properties and local learning [9,10], and meta-predicting the leave-one-out error of an individual example [11]. The latter work introduced a meta-level of reasoning and presented a meta-algorithm for predicting the leave-one-out error. The simulations showed that the proposed meta-learning approach compares favorably to the conventional theoretical leave-one-out error bounds.

The notion of reliability estimation has frequently appeared together with the notion of *transduction*, e.g. in [3,4]. Transduction is an inference principle that reasons from particular to particular [12], in contrast to inductive learning, which aims at inferring a general rule from a finite set of data. Transductive methods may therefore use only selected examples and not necessarily the whole input space. This locality enables the transductive algorithms to be used for making other inferences besides predictions. We find inferences of reliability measures to be of special interest. As an application of this principle, Kukar and Kononenko [13] proposed a transductive method for estimation of classification reliability. Their work introduced a set of reliability measures which successfully separate correct and incorrect classifications and are independent of the learning algorithm.

We later adapted this transductive approach to regression by proposing a simple linear model for predicting prediction reliability [14] and by evolving a basic sensitivity analysis approach, which we evaluated with regression trees and neural networks [15]. In our further work [16], we proposed a standardized framework for the use of sensitivity analysis to estimate prediction reliability, evaluating its suitability on three regression models (regression trees, neural networks and support vector machines) and motivating its development by the minimum description length principle [17]. In our previous work, we based our motivation for developing model-independent reliability estimates on the following ideas from related research fields:

- *Approaches which perturb data.* The design of these approaches implies that different subsets of learning examples cause different prediction accuracies for individual predictions [18–22]. Work in this field suggests that we should observe how the inclusion or removal of individual learning examples in the learning set influences the prediction accuracy of that particular example. Similar ideas of focusing on particular examples to perform learning are used also in the fields of active learning [23–25] and reinforcement learning [26,27].

- *Usage of unlabeled examples in supervised learning.* The results in this field [28–32] indicate that additional learning examples, generated from the same original probabilistic distribution, can be beneficial for improving predictor accuracy. This idea suggests that we should observe how the presence of a particular new example in the learning set influences the model's predictive accuracy. The idea also offers the possibility of making inferences about prediction reliability for an example by observing the resulting change in the model.
- *Sensitivity analysis.* This area presents a possible general framework which can be used to design an approach that is independent of the predictive model [33–36]. Namely, this technique requires no knowledge of the model's mathematical properties. Instead, the model is considered as a black box, with inputs and outputs as the only parameters being influenced and observed. We focus on this research field in more detail in Section 3.1.

The work presented here extends our initial work [14–16] and compares the performance of the previously proposed estimates to four adapted approaches from related work. They are summarized in the following section.

3. Reliability estimates

3.1. Local sensitivity analysis reliability estimates

3.1.1. Sensitivity analysis

An approach which enables us to analyze the *local* particularities of learning algorithms is *sensitivity analysis* [35–38], which is used in statistics and mathematical programming [39]. Sensitivity analysis aims to determine how much the output of a system is affected by variations in input. We can adopt the sensitivity analysis framework for reliability estimation by observing the changes in model outputs (predictions) when we modify its input (the learning data set). Apart from modifying the learning set, the sensitivity analysis approach could also observe the output variation with respect to changes in other model parameters. Since such a technique depends on the specifics of a model, however, we lose the benefits of having a model-independent procedure. Treating the predictive model as a black box, the sensitivity analysis approach indirectly analyzes aspects of the model that cannot be measured quantitatively, such as generalization ability, bias, resistance to noise, and avoidance of overfitting.

For influencing the input of the system (regression model), we expanded the learning set with an additional learning example, as motivated by the ideas from Section 2. By making such a minor change to the learning set and inducing a model on it, one can expect a similarly small change in the predictions of the new model. Large changes in prediction may be a sign of instability in the induced model. The purpose of modifying the learning set locally is to explore the sensitivity of the regression model in that particular part of the problem space. By doing so, the reliability estimates were adapted to the local particularities of data distribution and noise, thus relating the sensitivity to changes in prediction when the learning set was perturbed.

Given an unlabeled example and the *initial prediction* (the prediction of the original regression model) for which the reliability was to be estimated, we therefore repeatedly modified the learning set to obtain a set of *sensitivity models*. Using these sensitivity models, the *sensitivity predictions* were computed on the same example. A set of these *sensitivity predictions* allowed us to design the reliability estimates which evaluated the *local* bias and *local* variance in the problem subspace, leading to information about prediction reliability.

3.1.2. Reliability estimates

In our previous work [16] we used the sensitivity analysis approach to develop two reliability estimates (RE_1 and RE_3), which estimate the local variance and the local bias for a given unlabeled example. For greater clarity, we refer to these two estimates as SAvar (Sensitivity Analysis – variance) and SAbias (Sensitivity Analysis – bias) in this paper. To estimate the reliability for a given example, we expanded the learning set with that particular example, labeling it with $K + \varepsilon \cdot (l_{\max} - l_{\min})$. Here, K denotes the *initial prediction* of the example, ε is a sensitivity parameter (which influences the label value of the additional learning example and therefore indirectly defines the magnitude of the induced change in the initial learning set), and $l_{\min}$ and $l_{\max}$ denote the lower and the upper label bounds of the learning examples, respectively. In the next step, a sensitivity model was induced on the modified learning set and a *sensitivity prediction* K_ε was computed for the same particular example.

After computing different sensitivity predictions using different values of parameter ε (in our previous work, we selected $\varepsilon \in E$, $E = \{0.01, 0.1, 0.5, 1.0, 2.0\}$), the predictions were combined into different reliability estimates. Using more values of ε allowed us to widen the observation window in the local problem space and to define more robust reliability estimates by averaging the individual components. The estimates were defined by observing the differences between the initial and the sensitivity predictions. The main idea behind their definition was to observe the differences between sensitivity predictions made using ε and $-\varepsilon$. The estimates were defined as follows:

$$\text{SAvar} = \frac{\sum_{\varepsilon \in E}(K_\varepsilon - K_{-\varepsilon})}{|E|} \tag{1}$$

and

$$\text{SAbias} = \frac{\sum_{\varepsilon \in E}(K_\varepsilon - K) + (K_{-\varepsilon} - K)}{2|E|} \tag{2}$$

In the above estimates, K represents the prediction of the initial regression model (the *initial prediction*), K_ε and $K_{-\varepsilon}$ denote the *sensitivity predictions*, and E denotes a set of used sensitivity parameters ε.

3.2. Variance of a bagged model

In related work, the variance of predictions in bagged aggregates [18] of artificial neural networks has been used to indirectly estimate the reliability of the aggregated prediction [7,8]. Since an arbitrary regression model can be used with the bagging technique, we generalize the proposed reliability estimate for use with other regression models.

Given a bagged aggregate of m predictive models, where each of the models yields a prediction K_i, $i = 1,\ldots,m$, the label of an example is predicted by averaging the individual predictions:

$$K = \frac{\sum_{i=1}^{m} K_i}{m} \tag{3}$$

and reliability estimate BAGV is defined as the prediction variance:

$$\mathrm{BAGV} = \frac{1}{m}\sum_{i=1}^{m}(K_i - K)^2 \tag{4}$$

The design of the BAGV estimate is illustrated in Fig. 1. In our experimental work, the tested bagged aggregates contained $m = 50$ model instances.

3.3. Local cross-validation reliability estimate

Existing work has demonstrated various uses of only local information to perform learning. For example, Schaal and Atkeson [40] perform learning using locally applied linear models, and Birattari et al. [9] proposed an approach to dynamically select the most appropriate local neighborhood size. In this context, some authors also perform or optimize local learning by using local reliability estimates. In this way, Giacinto and Roli [10] use local models to evaluate local classification accuracies and dynamically select the most appropriate classifier to classify each example correctly. Similarly, Woods et al. [41] combine local classifiers using estimates of local classifiers' accuracies. In regression, Schaal and Atkeson [42] propose a local application of the locally weighted regression, optimizing it by assessing local prediction accuracy using the mean squared local cross-validation error.

In our work, we propose the LCV (Local Cross-Validation) reliability estimate, which is computed using the local leave-one-out (LOO) procedure. The estimate is similar to the mean squared local cross-validation error as in [42], except that it uses the absolute errors instead of the signed errors. We also use this estimate not only with neural networks, but with all eight regression models that we studied. Suppose that we are given an unlabeled example for which we wish to compute the prediction and the LCV estimate. Focusing on the subspace defined by k nearest neighbors (parameter k is selected in advance), we then generate k local models, each of them excluding one of the k nearest neighbors. Using the generated models, we compute the leave-one-out predictions K_i, $i = 1,\ldots,k$ for each of the nearest neighbors. Since the labels C_i, $i = 1,\ldots,k$ of the nearest neighbors are given, we are therefore able to calculate the absolute local leave-one-out prediction errors $E_i = |C_i - K_i|$. The prediction for the given local example is then computed using the k nearest neighbors (k-NN) algorithm and the LCV estimate is computed as the weighted average (weighted by distance using a distance metric $d()$) of the nearest neighbors' local errors E_i. The procedure is schematically illustrated in Fig. 2 and the algorithm pseudocode is shown in Fig. 3. In the algorithm, N denotes a set of the nearest neighbors, M denotes a local regression model, $d()$ denotes a distance metric and x an example for which we are computing the LCV estimate.

In our experimental work, we implemented the preceding algorithm to be adaptive to the size of the neighborhood with respect to the number of examples in the learning set. The parameter k was therefore not fixed, but assigned as $\lceil 1/20 \times |L| \rceil$, where $|L|$ denotes the number of learning examples.

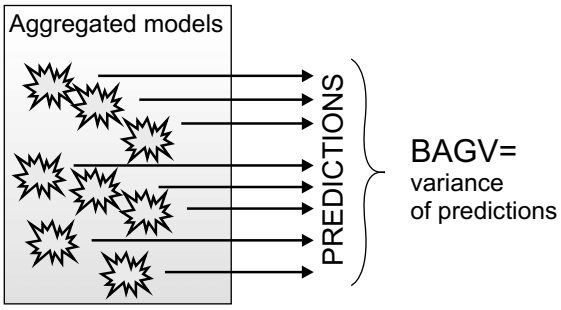

Fig. 1. Illustrated definition of the reliability estimate BAGV.

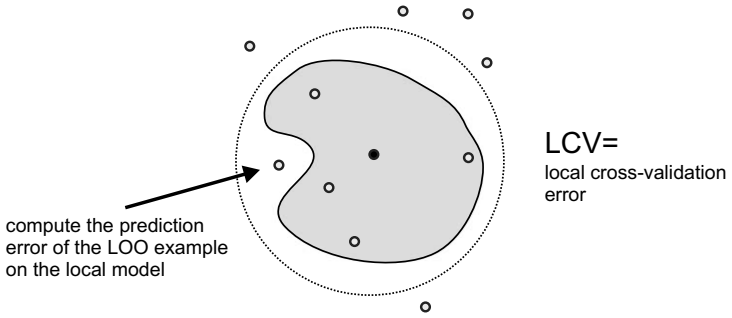

Fig. 2. Illustrated definition of the reliability estimate LCV.

```
1 PROGRAM LCV
2     define the set of k nearest neighbors N = {(x_1, C_1), ..., (x_k, C_k)}
3     FOR EACH (x_i, C_i) ∈ N
4         generate model M on N\(x_i, C_i)
5         for (x_i, C_i) compute LOO prediction K_i
6         for (x_i, C_i) compute LOO error E_i = |C_i − K_i|
7     END FOR EACH
8     LCV(x) = (Σ_{(x_i,C_i)∈N} d(x_i,x)·E_i) / (Σ_{(x_i,C_i)∈N} d(x_i,x))
9 END LCV
```

Fig. 3. The pseudocode for the computation of the LCV reliability estimate.

3.4. Density-based reliability estimate

One of the traditional approaches to estimating prediction reliability is based on the distribution of learning examples in the input space [43]. As illustrated in Fig. 4, the density-based estimation of prediction error assumes that error is lower for predictions which are made in *denser* problem subspaces (a portion of the input space with a more learning examples), and higher for predictions which are made in sparser subspaces (a portion of the input space with fewer learning examples). This means that we trust the prediction with respect to the *quantity of information* that is available for its computation.

A typical use of this approach is with decision and regression trees, where we trust each prediction according to the proportion of learning examples that fall in the same leaf of a tree as the predicted example. But although this approach considers the quantity of disposable information, it also has the disadvantage that it does not take into account the learning examples' labels. This causes the method to perform poorly with noisy data and in cases when distinct examples are not clearly separable.

We define the reliability estimate DENS as the value of the estimated probability density function for a given unlabeled example. To estimate the density, we use Parzen windows [44,45] with the Gaussian kernel. We reduced the problem of computing the multidimensional Gaussian kernel to computing the two-dimensional kernel by using a distance function ap-

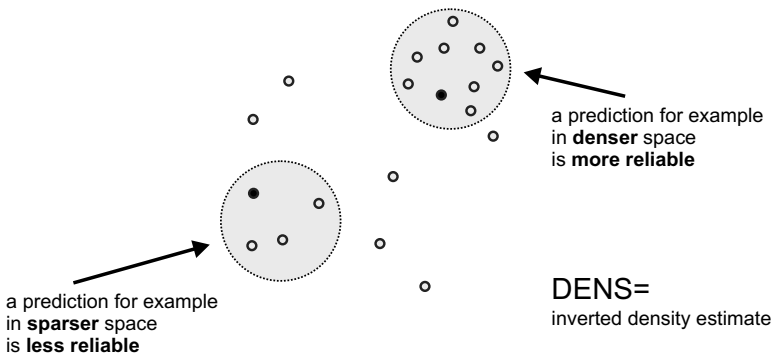

Fig. 4. Illustrated definition of the reliability estimate DENS.

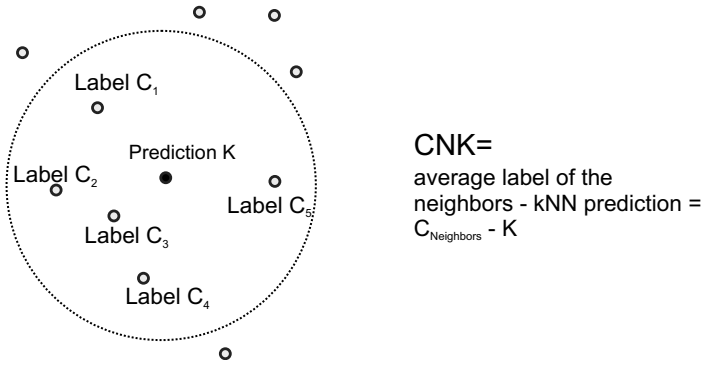

Fig. 5. Illustrated definition of the reliability estimate CNK.

plied to pairs of example vectors. Given the learning set $L=((x_1,y_1),\ldots,(x_n,y_n))$, the density estimate for unlabeled example $(x,\_)$ is therefore defined as

$$p(x) = \frac{\sum_{i=1}^{n} \kappa(D(x,x_i))}{n} \tag{5}$$

where D denotes a distance function, κ denotes a kernel function (in our case the Gaussian), x denotes an example for which we are estimating reliability and x_i, $i = 1,\ldots,n$ denotes the learning examples. Since we expect the prediction error to be higher in cases when the density is lower, this means that $p(x)$ correlates negatively with the prediction error. To establish the positive correlation, which can be used to compare our method with other reliability estimates, we need to invert $p(x)$, thus defining the reliability estimate as

$$\mathrm{DENS}(x) = \max_{i=1,\ldots,n}(p(x_i)) - p(x) \tag{6}$$

where $\max_{i=1,\ldots,n}(p(x_i))$ denotes the maximum value of estimated density over all learning examples in the learning set L.

3.5. Local modeling of prediction error

For comparison with existing reliability estimates, we also propose an approach to local estimation of prediction reliability using the nearest neighbors' labels. Given a set of nearest neighbors $N=\{(x_1,C_1),\ldots,(x_k,C_k)\}$, we define the estimate CNK ($C_{\mathrm{Neighbors}} - K$) for an unlabeled example $(x,\_)$ as the difference between the average label of the nearest neighbors and the example's prediction K (using the model that was generated on all learning examples):

$$CNK = \frac{\sum_{i=1}^{k} C_i}{k} - K \tag{7}$$

where k denotes the number of neighbors, C_i denotes neighbors' labels and K denotes the example's prediction. In our experiments we computed estimate CNK using five nearest neighbors. The design of the estimate CNK is illustrated in Fig. 5.

4. Experimental results

We tested and compared the estimates SAvar and SAbias, developed in our previous work, to the other estimates described in Sections 3.2–3.5. Testing was performed using the leave-one-out cross-validation procedure. For each learning example that was left out in the current iteration, we computed the prediction and all the reliability estimates. The performance of the reliability estimates was measured by computing the Pearson correlation coefficient between each reliability estimate and the prediction error. The significance of the correlation coefficient was then statistically evaluated using the two-sided t-test for correlation coefficients.

Note that all of the estimates are expected to correlate positively with the prediction error. This means that all the estimates are founded so that higher absolute values represent less reliable predictions and lower absolute values represent more reliable predictions (the value 0 represents the reliability of the most reliable prediction). Note also that all of the estimates except SAbias and CNK can take only positive values. Besides the magnitude (absolute value), which we interpret as the prediction reliability, these two estimates also provide additional information about the direction of error (whether the value of prediction was too high or too low). This holds potential for further work in correcting initial predictions using these two estimates. We therefore performed the experiments by correlating the magnitudes of estimates to the absolute prediction error of test examples. For estimates SAbias and CNK we also correlated their signed values to the signed prediction error of test examples. In this way, we actually tested the performance of eight estimates: SAvar, SAbias-s (signed), SAbias-a (absolute), BAGV, LCV, DENS, CNK-s and CNK-a.

The performance of reliability estimates was tested using eight regression model implemented in the statistical package R [46]. Here are some key properties of the models used:

Regression trees (RT): trees [47] with mean squared error used as the splitting criterion, the value in leaves represents the average label of examples, and trees were unpruned,
Linear regression (LR): linear regression with no explicit parameters,
Neural networks (NN): three-layered perceptron [48] with five hidden neurons and tanh activation function, learning was performed using backpropagation with adaptive gradient descent,
Bagging (BAG): bagging [18] with 50 regression trees,
Support vector machines (SVM): an implementation of the support vector ε-regression [12,49], implemented in the library for support vector machines (LIBSVM) [50,51]. We use a radial basis function (RBF) kernel with parameter $\gamma = 1/(\text{number of attributes})$, parameter $C = 1$ and the precision parameter $\epsilon = 0.1$,
Locally weighted regression (LWR): local regression with Gaussian kernel for weighting examples according to their distance,
Random forests (RF): random forests [52] with 100 trees,
Generalized additive model (GAM): linear model [53,54] with no special parameters.

The aim of our research was to evaluate the reliability estimates with models being treated as black boxes. Therefore, the focus of our research was not to optimize the above model parameters to improve prediction accuracy, but to evaluate the accuracy of reliability estimates. For the testing, 28 benchmark data sets were used, which are standard across the machine learning community. Each data set is a regression problem. The application domains vary, including medical, ecological, technical, mathematical and physical. Most of the data sets are available from the UCI machine learning repository [55] and from the StatLib DataSets Archive [56]. All data sets are available from the authors on request. A brief description of the data sets is given in Table 1.

4.1. Testing of individual estimates

The summarized results of the experiments are shown in Table 2 and in Fig. 6. The data in Table 2 indicate the percent of domains with a significant positive or negative correlation between the reliability estimates and the prediction error. The same results are presented graphically in Fig. 6 for each of the tested regression models. The graphs show the performance of reliability estimates (percentage of tests with significant positive/negative correlation with the prediction error), ranked in decreasing order with respect to the percentage of positive correlations.

Table 1
Basic characteristics of the testing data sets

Data set	Number of examples	Number of discrete attributes	Number 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

Table 2
Percentage of experiments exhibiting significant positive/negative correlations between *reliability estimates* and *prediction error*

Model	SAvar +/−	SAbias-s +/−	SAbias-a +/−	BAGV +/−	LCV +/−	DENS +/−	CNK-s +/−	CNK-a +/−
RT	46/0	82/0	50/0	64/0	36/0	36/4	86/0	68/0
LR	54/0	7/0	7/4	54/0	32/0	32/4	50/0	57/0
NN	39/4	18/4	29/4	50/0	36/0	25/4	36/4	39/4
BAG	46/4	21/0	11/0	57/0	50/0	36/7	25/0	46/0
SVM	46/4	36/7	25/0	46/0	61/0	39/4	29/11	36/0
LWR	39/7	4/7	11/7	46/0	46/0	43/7	25/11	32/0
RF	25/7	14/0	11/0	57/0	61/0	32/11	11/25	46/4
GAM	54/0	7/0	7/4	50/0	32/0	32/4	50/0	57/0
Average	**44/3**	**24/2**	**19/2**	**53/0**	**44/0**	**34/6**	**39/6**	**48/1**

The last line in Table 2 presents the results of the reliability estimates, averaged across all eight testing regression models. We can see that the best results were achieved using the estimates BAGV, CNK-a, LCV and SAvar (in decreasing order with respect to the percentage of significant positive correlations). The estimate SAbias-a achieved the worst average results.

Due to space limitations, the detailed results (the values of correlation coefficients between estimates and the prediction error for each of the reliability estimates) are not presented here and are available from the authors on request. A summary of the results is provided in Table 3, showing the correlation coefficients for individual domains, averaged across the eight tested regression models. The results confirm our expectation that the reliability estimates should positively correlate with the prediction error. Namely, we can see that the positive (desired) correlations outnumber the negative (non-desired) correlations in all regression model/reliability estimate pairs.

By examining the detailed results (Table 2) we can see that the performance of the estimate CNK-s with the regression trees stands out, being significantly positively correlated with the prediction error in 86% of the experiments and not negatively correlated in any experiment. Estimate SAbias-s achieved similar performance, with positive correlation in 82% of tests with regression trees and negative in no tests. The estimate CNK-a achieved better performance with linear models (linear regression, generalized additive model) and regression trees than the estimate BAGV, which was also the estimate with the highest average number of significant positive correlations (53%).

The results achieved by estimates SAbias-a and SAbias-s, also confirm our expectations and analysis from previous work [16] that estimates based on sensitivity analysis are not appropriate for models which do not partition the input space (in our case: linear regression, locally weighted regression, and generalized additive model). Namely, from Table 2 we can see that no other reliability estimate performed worse, indicating that the other tested estimates present a better choice for use on these models.

The results indicate that the estimates SAbias, CNK, BAGV and LCV have good potential for estimating prediction reliability. However, these estimates performed differently with different regression models. This motivated us to combine pairs of estimates in order to compose a new estimate which would perform well with all models.

4.2. Combination of estimates

To create a new estimate that would perform well on all tested regression models, we combined pairs of estimates using a linear model:

$$\text{Estimate}_{\text{new}} = \gamma \cdot \text{Estimate}_1 + (1-\gamma) \cdot \text{Estimate}_2 \tag{8}$$

Different choices of the value γ give different weights to each of the combined estimates in Eq. (8), making the combined estimate behave more similarly to the estimate with the greater weight. Since we want to create a combined estimate that equally combines the performance characteristics of both reliability estimates, we selected weight $\gamma = 0.5$, hence averaging Estimate_1 and Estimate_2. In our experimental work, we tested all possible pairs from the eight individual estimates, subject to the limitation that they can be reasonably combined with respect to correlation to signed or absolute prediction error. Thus, individual estimates which were correlated to the signed prediction error were combined only with other such estimates; the estimates which were correlated to the absolute prediction error were combined only with other such estimates. The testing of combined estimates was performed using the same experimental protocol as the testing of individual estimates. The results are shown in Table 4.

From the results on combined pairs of reliability estimates, we can see that the combination of estimates BAGV and CNK-a achieved the best performance. Comparing the results in Tables 2 and 4 reveals that the combined estimate achieved better results (higher percentage of experiments with significant positive correlations with the prediction error) with neural networks and with bagging. However, the combination of BAGV and CNK-a on average had significant positive correlation with the prediction error in 54% of tests (and negatively in 1%), which means that it did not achieve a signif-

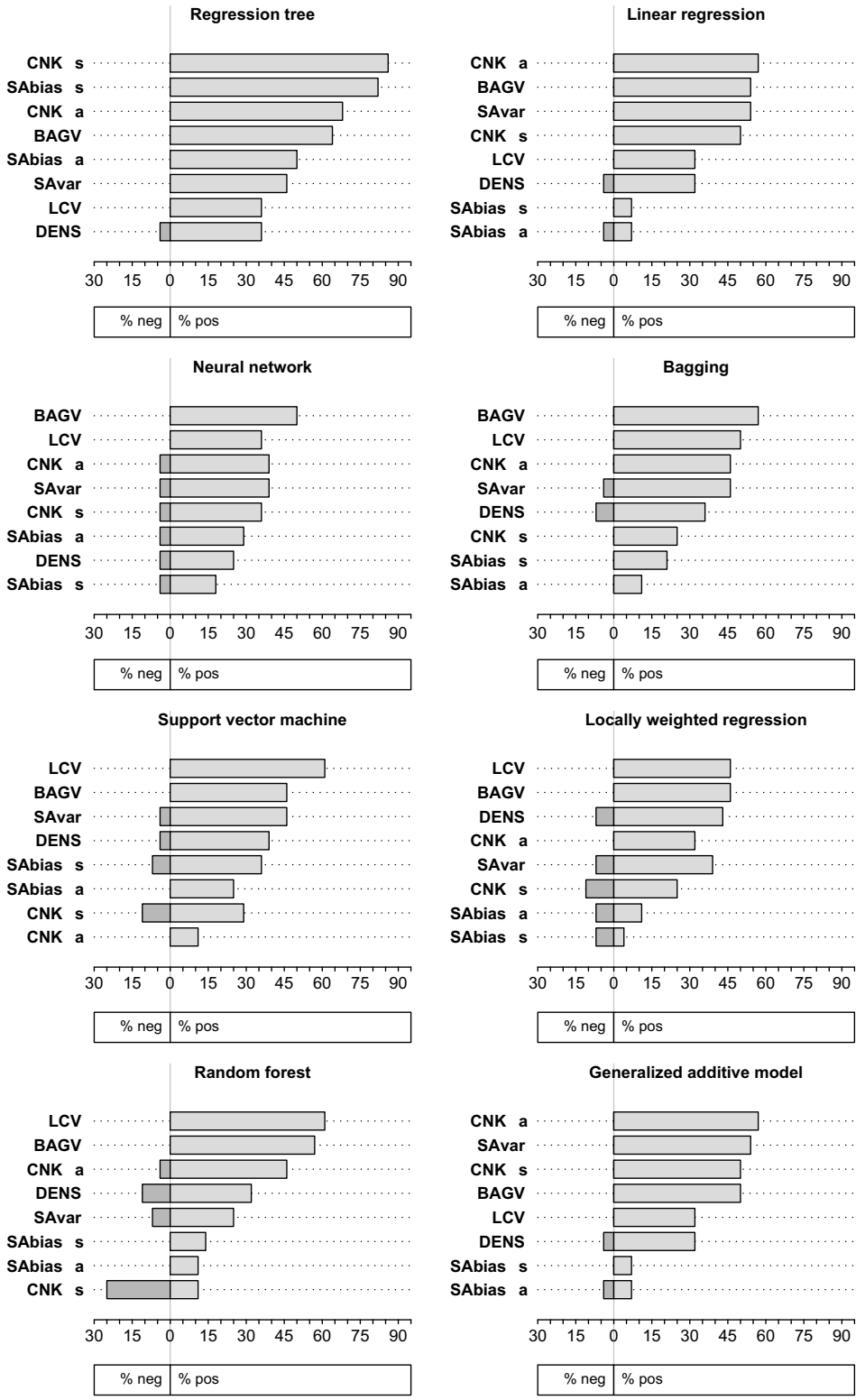

Fig. 6. Ranking of reliability estimates by percentage of significant positive and negative correlations with prediction error.

icant advantage over the estimate BAGV, which had significant positive correlation with the prediction error in 53% of tests.

Table 3
Average correlation coefficients, achieved by the reliability estimates in the individual domains

Domain	SAvar	SAbias-s	SAbias-a	BAGV	LCV	DENS	CNK-s	CNK-a
autoprice	**0.451**	0.108	0.099	**0.459**	**0.306**	**0.285**	0.153	**0.428**
auto93	0.189	0.087	−0.026	**0.267**	**0.236**	0.085	**0.243**	**0.350**
autohorse	**0.458**	−0.029	**0.144**	**0.314**	0.111	**0.254**	0.090	**0.300**
baskball	0.017	0.082	−0.001	0.036	0.067	0.075	0.143	0.015
bodyfat	**0.152**	0.085	0.071	**0.237**	0.092	**0.165**	**0.173**	**0.290**
brainsize	−0.048	0.046	−0.129	−0.073	−0.118	−0.194	0.053	−0.164
breasttumor	−0.080	0.104	−0.034	−0.027	−0.017	−0.084	0.039	0.003
cloud	**0.427**	0.036	0.112	**0.414**	**0.233**	**0.442**	0.083	**0.318**
cpu	**0.523**	**0.150**	**0.278**	**0.628**	**0.517**	**0.363**	0.068	**0.670**
diabetes	0.210	0.055	−0.009	0.175	0.017	**0.396**	0.214	−0.032
echomonths	−0.003	0.069	0.062	0.131	0.136	−0.162	0.145	0.169
elusage	0.194	0.039	0.123	**0.293**	**0.325**	0.205	0.125	0.104
fishcatch	**0.246**	0.137	0.125	**0.596**	**0.466**	**0.341**	**0.421**	**0.499**
fruitfly	−0.028	0.080	−0.008	−0.048	−0.071	−0.060	−0.149	0.022
grv	**0.187**	0.120	0.076	**0.195**	**0.183**	0.014	0.170	0.124
hungarian	**0.247**	0.053	**0.131**	**0.380**	**0.193**	0.101	**0.222**	**0.387**
lowbwt	0.079	0.048	−0.014	0.091	0.011	0.053	0.071	0.083
mbagrade	0.004	−0.007	0.001	0.019	0.053	−0.059	−0.040	0.031
pharynx	**0.164**	0.109	0.058	**0.236**	**0.170**	0.134	−0.034	**0.166**
pollution	0.115	0.042	−0.074	0.083	0.007	0.105	**0.258**	0.157
pwlinear	**0.191**	0.017	0.015	**0.233**	0.029	0.137	0.102	**0.166**
pyrim	**0.317**	0.097	−0.007	**0.325**	0.217	0.097	**0.343**	**0.504**
servo	0.040	0.048	**0.163**	**0.543**	**0.452**	**−0.245**	−0.098	**0.226**
sleep	0.118	0.221	0.033	**0.280**	0.190	−0.037	0.164	0.170
transplant	**0.302**	**0.182**	0.162	**0.434**	**0.482**	**0.477**	**0.280**	**0.380**
triazines	**0.187**	0.070	0.038	**0.284**	**0.147**	**0.153**	**0.228**	**0.269**
tumor	−0.075	0.143	0.077	0.105	0.136	−0.104	0.146	−0.007
wpbc	0.042	0.101	0.059	0.042	0.080	−0.052	0.139	0.048

Statistically significant values are denoted with bold type. Significant negative values are also underlined.

Table 4
Percentage of experiments exhibiting significant positive/negative correlations between the *combinations of the reliability estimates* and the *prediction error*

Model	SAvar & SAbias-a +/−	SAvar & BAGV +/−	SAvar & LCV +/−	SAvar & DENS +/−	SAvar & CNK-a +/−	SAbias-s & CNK-s +/−	SAbias-a & BAGV +/−	SAbias-a & LCV +/−
RT	50/0	61/0	43/0	46/0	61/0	89/0	61/0	50/0
LR	54/0	54/0	39/4	54/0	61/0	50/0	54/0	32/0
NN	39/0	54/0	32/0	36/4	50/0	36/4	46/0	32/0
BAG	46/4	54/0	39/0	43/4	54/0	25/0	54/0	39/0
SVM	46/4	46/0	46/0	50/4	46/4	32/11	46/0	61/0
LWR	39/7	43/4	46/4	39/7	43/7	29/7	43/0	46/0
RF	25/7	39/4	25/4	25/7	36/4	14/18	50/0	50/0
GAM	54/0	57/0	39/4	54/0	61/0	50/0	50/0	32/0
Average	**44/3**	**51/1**	**39/2**	**43/3**	**52/2**	**41/5**	**51/0**	**43/0**

Model	SAbias-a & DENS +/−	SAbias-a & CNK-a +/−	BAGV & LCV +/−	BAGV & DENS +/−	BAGV & CNK-a +/−	LCV & DENS +/−	LCV & CNK-a +/−	DENS & CNK-a +/−
RT	54/0	68/4	61/0	57/0	71/4	36/0	57/0	64/0
LR	32/4	57/0	43/0	50/0	50/0	32/0	39/0	57/0
NN	29/4	43/4	39/0	54/0	54/0	32/0	39/0	39/4
BAG	18/0	46/0	54/0	54/0	61/0	50/0	50/0	46/0
SVM	29/0	36/0	54/0	54/0	46/0	61/0	50/0	36/0
LWR	7/7	32/0	46/0	46/0	43/0	46/0	36/0	36/0
RF	11/0	46/4	61/0	50/0	50/0	61/0	61/0	46/4
GAM	32/4	57/0	46/0	54/0	54/0	32/0	39/0	57/0
Average	**27/1**	**48/2**	**51/0**	**52/0**	**54/1**	**44/0**	**46/0**	**48/1**

Since the combination of the estimates BAGV and CNK-a performed well with some regression models and achieved good average results, we selected it for further evaluation. Let us therefore define the estimate BVCK (Bagging Variance – $(C_{\text{neighbors}} - K)$) which is defined as

$$\text{BVCK}(x) = \frac{\text{BAGV}(x) + \text{CNK}(x)}{2}. \tag{9}$$

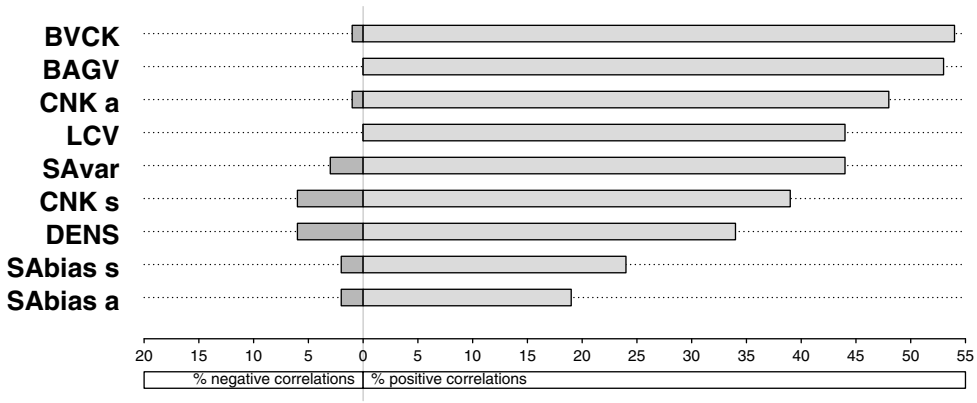

Fig. 7. Ranking of the reliability estimates by the average percentage of significant positive and negative correlations with the prediction error.

Table 5
An example of analyzing prediction reliability

	Prediction	SAvar	SAbias-s	BAGV	LCV	DENS	CNK-s	BVCK
$(x_1,\_)$	101.26	10.69	−8.50	8.78	−14.00	0.00	1.94	5.36
$(x_2,\_)$	99.63	12.13	0.86	10.94	−16.00	0.01	−2.63	6.79

The table shows predictions and reliability estimates' values for two unlabeled examples. Due to space limitations, absolute versions of SAbias and CNK were omitted.

where x denotes the example for which we are estimating reliability and BAGV(x) and CNK(x) denote the corresponding estimates' values for that particular example.

Fig. 7 presents a comparison of the results, averaged across all regression models for all reliability estimates.

Example 1. Analyzing prediction reliability Suppose we are given a learning set L and two unlabeled examples $(x_1,\_)$ and $(x_2,\_)$ for which we wish to estimate and compare reliability. After computing the predictions and nine reliability estimates using a selected regression model for each unlabeled example, we obtain the values shown in Table 5. Five out of seven displayed estimates indicate that the prediction for the first example is more reliable (since all the estimates negatively correlate with the prediction error, lower values therefore represent lower estimated error). Based on this finding, we might therefore decide to trust only the prediction of the first example and potentially reject the prediction of the second example.

However, since values of the each estimate belong to the estimate-specific interval, they are difficult to compare with other reliability estimates. If one knew whether the values of each estimate are low or high with respect to the entire estimate's target interval, one would be able to interpret the reliabilities for the examples in Table 5 differently. Also, the discrepancy between the values of DENS and the other estimates reveals that some estimates may be more suitable for a given problem than others. Both these issues provide motivation for further work in improving estimates' interpretability and selecting the most appropriate reliability estimate for a given problem (see Section 5).

5. Conclusion and further work

In this paper, we focused on the importance of reliability and prediction accuracy in supervised learning, as already discussed in many of the related work in this area [57–61]. We compared the sensitivity-based reliability estimates proposed in our previous work with four other approaches to estimating the reliability of individual predictions, namely variance of bagged predictors, local cross-validation, density-based estimation and local error estimation. Testing on 28 testing domains and eight regression models showed promising results for using estimates SAbias-s and CNK-s with regression trees. These estimates significantly positively correlated with the signed prediction error in 82% and 86% of tests, respectively. The best average performance was achieved by the estimate BAGV, which turned out to be the best choice for use with neural networks, bagging and locally weighted regression. With linear models (linear regression and generalized additive model), the estimate CNK-a performed better than BAGV, thus improving on the performance of all other tested approaches.

With the aim of improving the estimates' performance, we combined pairs of reliability estimates, which perform differently with different regression models. The combination of the estimates BAGV and CNK-a performed better than any other estimate with neural networks and with bagging. On average, this combination performed comparably to the most successful individual estimate BAGV, with significant positive correlation with prediction error in 54% of tests, and negative in 1%. Based on these results, we selected this estimate for further evaluation and denoted it BVCK.

The results show the potential of using the sensitivity-based reliability estimates for estimation of prediction error with selected regression predictors. They also show the favorable performance of the newly proposed local error modeling estimate CNK, as compared with the other estimates. These results and new ideas offer challenges for further work:

- Good performance of the signed reliability estimates (SAbias-s with regression trees and CNK-s) with the signed prediction error implies the potential to use these reliability estimates for correction of regression predictions. We shall therefore explore whether these two reliability estimates can be used to significantly reduce the error of regression predictions.
- Different magnitudes of the correlation coefficients in different testing domains (see Table 3) indicate that the potential for estimation of prediction reliability is more feasible in some domains than in others. The domain and model characteristics which lead to good performance of the reliability estimates shall also be analyzed.
- Since different estimates performed differently with different domain/model pairs, an approach to automatic selection of the optimal reliability estimate for a given domain, model and example shall be developed. For example-based selection of the most appropriate estimate, a mapping of estimates' values onto the same interval is necessary to ensure comparability of estimates' across different examples.

References

[1] I. Kononenko, M. Kukar, Machine Learning and Data Mining: Introduction to Principles and Algorithms, Horwood Publishing Limited, UK, 2007.
[2] M.J. Crowder, A.C. Kimber, R.L. Smith, T.J. Sweeting, Statistical concepts in reliability, Statistical Analysis of Reliability Data, Chapman & Hall, London, UK, 1991.
[3] A. Gammerman, V. Vovk, V. Vapnik, Learning by transduction, in: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, Madison, Wisconsin, 1998, pp. 148–155.
[4] C. Saunders, A. Gammerman, V. Vovk, Transduction with confidence and credibility, in: Proceedings of IJCAI'99, vol. 2, 1999, pp. 722–726.
[5] I. Nouretdinov, T. Melluish, V. Vovk, Ridge regression confidence machine, in: Proc. 18th International Conf. on Machine Learning, Morgan Kaufman, San Francisco, CA, 2001, pp. 385–392.
[6] A. Weigend, D. Nix, Predictions with confidence intervals (local error bars), in: Proceedings of the International Conference on Neural Information Processing (ICONIP'94), Seoul, Korea, 1994, pp. 847–852.
[7] T. Heskes, Practical confidence and prediction intervals, in: M.C. Mozer, M.I. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing Systems, vol. 9, The MIT Press, 1997, pp. 176–182.
[8] J. Carney, P. Cunningham, Confidence and prediction intervals for neural network ensembles, in: Proceedings of IJCNN'99, The International Joint Conference on Neural Networks, Washington, USA, 1999, pp. 1215–1218.
[9] M. Birattari, H. Bontempi, H. Bersini, Local learning for data analysis, in: Proceedings of the Eighth Belgian-Dutch Conference on Machine Learning, 1998, pp. 55–61.
[10] G. Giacinto, F. Roli, Dynamic classifier selection based on multiple classifier behaviour, Pattern Recognition 34 (9) (2001) 1879–1881.
[11] K. Tsuda, G. Rätsch, S. Mika, K. Müller, Learning to predict the leave-one-out error of kernel based classifiers, Lecture Notes in Computer Science (2001) 331.
[12] V. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.
[13] M. Kukar, I. Kononenko, Reliable classifications with machine learning, in: T. Elomaa, H. Manilla, H. Toivonen (Eds.), Proc. Machine Learning: ECML-2002, Springer-Verlag, Helsinki Finland, 2002, pp. 219–231.
[14] Z. Bosnić, I. Kononenko, M. Robnik-Šikonja, M. Kukar, Evaluation of prediction reliability in regression using the transduction principle, in: B. Zajc, M. Tkalčič (Eds.), Proceedings of Eurocon 2003, Ljubljana, 2003, pp. 99–103.
[15] Z. Bosnić, I. Kononenko, Estimation of regressor reliability, Journal of Intelligent Systems 17 (1/3) (2008) 297–311.
[16] Z. Bosnić, I. Kononenko, Estimation of individual prediction reliability using the local sensitivity analysis, Applied Intelligence, in press [Online edition] http://www.springerlink.com/content/e27p2584387532g8/.
[17] M. Li, P. Vitányi, An Introduction to Kolmogorov Complexity and its Applications, Springer-Verlag, New York, 1993.
[18] L. Breiman, Bagging predictors, Machine Learning 24 (2) (1996) 123–140.
[19] D. Wolpert, Stacked generalization, Neural Networks, vol. 5, Pergamon Press, 1992, pp. 241–259.
[20] R. Tibshirani, K. Knight, Model search and inference by bootstrap bumping, Journal of Computational and Graphical Statistics 8 (1999) 671–686.
[21] Y. Freund, R. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences 55 (1) (1997) 119–139.
[22] G. Elidan M. Ninio N. Friedman D. Schuurmans, Data perturbation for escaping local maxima in learning, 2002.
[23] D.A. Cohn, Z. Ghahramani, M.I. Jordan, Active learning with statistical models, in: G. Tesauro, D. Touretzky, T. Leen (Eds.), Advances in Neural Information Processing Systems, vol. 7, The MIT Press, 1995, pp. 705–712. citeseer.ist.psu.edu/cohn95active.html.
[24] L.A.D.A. Cohn and, R. Ladner, Training connectionist networks with queries and selective sampling, in: M.K.D. Touretzky, (Ed.), Advances in Neural Information Processing Systems, vol. 2, 1990, pp. 566–573.
[25] A. Linden, F. Weber, Implementing inner drive by competence reflection, in: Proceedings of the Second International Conference on Simulation of Adaptive Behavior, Hawaii, 1992, pp. 321–326.
[26] J. Schmidhuber, J. Storck, Reinforcement driven information acquisition in nondeterministic environments, Fakultat fur Informatik, Technische Universit at Munchen, Technical Report, 1993.
[27] S.D. Whitehead, A complexity analysis of cooperative mechanisms in reinforcement learning, in: AAAI, 1991, pp. 607–613.
[28] M. Seeger, Learning with labeled and unlabeled data, Technical Report, http://www.dai.ed.ac.uk/seeger/papers.html, 2000.
[29] A. Blum, T. Mitchell, Combining labeled and unlabeled data with co-training, in: Proceedings of the 11th Annual Conference on Computational Learning Theory, 1998, pp. 92–100.
[30] T. Mitchell, The role of unlabelled data in supervised learning, in: Proceedings of the Sixth International Colloquium of Cognitive Science, San Sebastian, Spain, 1999.
[31] V. de Sa, Learning classification with unlabeled data, in: J.D. Cowan, G. Tesauro, J. Alspector (Eds.), Proc. NIPS'93, Neural Information Processing Systems, Morgan Kaufmann Publishers, San Francisco, CA, 1993, pp. 112–119.
[32] S. Goldman, Y. Zhou, Enhancing supervised learning with unlabeled data, in: Proc. 17th International Conf. on Machine Learning, Morgan Kaufman, San Francisco, CA, 2000, pp. 327–334.
[33] L. Breierova, M. Choudhari, An introduction to sensitivity analysis, MIT System Dynamics in Education Project, September 1996.
[34] J. Kleijnen, Experimental designs for sensitivity analysis of simulation models, in: Tutorial at the Eurosim 2001 Conference.
[35] O. Bousquet, A. Elisseeff, Stability and generalization, Journal of Machine Learning Research 2 (2002) 499–526.
[36] M.J. Kearns, D. Ron, Algorithmic stability and sanity-check bounds for leave-one-out cross-validation, Computational Learning Theory (1997) 152–162.

[37] O. Bousquet, A. Elisseeff, Algorithmic stability and generalization performance, in: NIPS, 2000, pp. 196–202.
[38] O. Bousquet, M. Pontil, Leave-one-out error and stability of learning algorithms with applications, in: J.A.K. Suykens et al. (Eds.), Advances in Learning Theory: Methods, Models and Applications, IOS Press, 2003.
[39] A. Saltelli, S. Tarantola, F. Campolongo, M. Ratto, Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models, John Wiley & Sons Ltd., England, 2003.
[40] S. Schaal, C.G. Atkeson, Constructive incremental learning from only local information, Neural Computation 10 (8) (1998) 2047–2084.
[41] K. Woods, W.P. Kegelmeyer, K. Bowyer, Combination of multiple classifiers using local accuracy estimates, IEEE Transactions on PAMI 19 (4) (1997) 405–410.
[42] S. Schaal, C.G. Atkeson, Assessing the quality of learned local models, in: J.D. Cowan, G. Tesauro, J. Alspector (Eds.), Advances in Neural Information Processing Systems, vol. 6, Morgan Kaufman Publishers, Inc., 1994, pp. 160–167.
[43] M.P. Wand, M.C. Jones, Kernel Smoothing, Chapman and Hall, London, 1995.
[44] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Monographs on Statistics and Applied Probability, Chapman and Hall, London, 1986.
[45] B. Jeon, D.A. Landgrebe, Parzen density estimation using clustering-based branch and bound, Transactions on Pattern Analysis and Machine Intelligence (1994) 950–954.
[46] R Development Core Team, A Language and Environment for Statistical Computing, R Foundation for Statistical Computing,Vienna, Austria, 2006.
[47] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and Regression Trees, Wadsworth International Group, Belmont CA, 1984.
[48] W.S. McCulloch, W. Pitts, A logical calculus of the ideas imminent in nervous activity, Bulletin of Mathematical Biophysics 5 (1943) 115–133.
[49] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, NeuroCOLT2 Technical Report NC2-TR-1998-030, 1998.
[50] N. Christiannini, J. Shawe-Taylor, Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, 2000.
[51] C. Chang, C. Lin, LIBSVM: a library for support vector machines, software available at http://www.csie.ntu.edu.tw/cjlin/libsvm, 2001.
[52] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.
[53] S.N. Wood, Generalized Additive Models: An Introduction with R, Chapman & Hall, CRC, 2006.
[54] T. Hastie, R. Tibshirani, Generalized Additive Models, Chapman and Hall, London, 1990.
[55] A. Asuncion D.J. Newman, UCI machine learning repository, 2007.
[56] Department of Statistics at Carnegie Mellon University, Statlib – data, software and news from the statistics community, http://lib.stat.cmu.edu/, 2005.
[57] J. Mendling, H.M.W. Verbeek, B.F. van Dongen, W.M.P. van der Aalst, Detection and prediction of errors in epcs of the sap reference model, Data and Knowledge Engineering 64 (1) (2008) 312–329.
[58] T.H. Fan, K.F. Cheng, Tests and variables selection on regression analysis for massive datasets, Data and Knowledge Engineering 63 (4) (2007) 811–819.
[59] Y. Han, W. Lam, Utilizing hierarchical feature domain values for prediction, Data and Knowledge Engineering 61 (3) (2007) 540–553.
[60] Y.J. Lee, W.F. Hsieh, C.M. Huang, epsilon-ssvr: A smooth support vector machine for epsilon-insensitive regression, IEEE Transactions of Knowledge and Data Engineering 17 (5) (2005) 678–685.
[61] Z.H. Zhou, M. Li, Semi-supervised regression with co-training style algorithms, IEEE Transactions of Knowledge and Data Engineering 19 (11) (2007) 1479–1493.

Zoran Bosnić obtained his Master and Doctor degrees in Computer Science at University of Ljubljana (Slovenia) in 2003 and 2007, respectively. Since 2006 he has been working as an assistant at Faculty of Computer and Information Science (Laboratory of Cognitive Modelling). His research interests include artificial intelligence, machine learning, regression, and reliability estimation for individual predictions, as well as applications in these areas.

Igor Kononenko received his PhD in 1990 from University of Ljubljana, Slovenia. He is a professor at the Faculty of Computer and Information Science in Ljubljana and the head of the Laboratory for Cognitive Modeling. His research interests include artificial intelligence, machine learning, neural networks and cognitive modeling. He is a member of the editorial board of Applied Intelligence Journal and Informatica Journal. He is a (co)author of 180 papers and 10 textbooks. Recently he co-authored the book "Machine Learning and Data Mining: Introduction to Principles and Algorithms" (Hoorwood, 2007).