Estimation of individual prediction reliability using the local sensitivity analysis

Zoran Bosnić · Igor Kononenko

Published online: 3 August 2007
© Springer Science+Business Media, LLC 2007

Abstract For a given prediction model, some predictions may be reliable while others may be unreliable. The average accuracy of the system cannot provide the reliability estimate for a single particular prediction. The measure of individual prediction reliability can be important information in risk-sensitive applications of machine learning (e.g. medicine, engineering, business). We define empirical measures for estimation of prediction accuracy in regression. Presented measures are based on sensitivity analysis of regression models. They estimate reliability for each individual regression prediction in contrast to the average prediction reliability of the given regression model. We study the empirical sensitivity properties of five regression models (linear regression, locally weighted regression, regression trees, neural networks, and support vector machines) and the relation between reliability measures and distribution of learning examples with prediction errors for all five regression models. We show that the suggested methodology is appropriate only for the three studied models: regression trees, neural networks, and support vector machines, and test the proposed estimates with these three models. The results of our experiments on 48 data sets indicate significant correlations of the proposed measures with the prediction error.

1 Introduction

A key issue in determining the quality of learning algorithm is the measurement of its accuracy. Commonly used accuracy measures, as the mean squared error (MSE) and the relative mean squared error (RMSE), evaluate the model performance by summarizing the error contributions of all test examples. They nevertheless provide no local information about the expected error of individual prediction for a given unseen example.

Measuring the expected prediction error is very important in risk-sensitive areas where acting upon predictions may have financial or medical consequences (e.g. medical diagnosis stock market, navigation, control applications). In such areas, appropriate *local* accuracy measures may provide additional necessary information about the prediction confidence. For example, in medical diagnosis physicians are not interested only in the average accuracy of the predictor. When a certain patient is analyzed, the physicians expect from a system to be able to provide a prediction as well as the estimate of the reliability of that particular prediction. The average accuracy of the model cannot provide information whether some particular prediction is reliable or not.

The above described challenge is illustrated in Fig. 1. It illustrates a contrast between the average reliability estimate (e.g. MSE), and reliability estimates of individual predictions. Note that instead of estimating the aggregated accuracy of the whole prediction model, the individual reliability estimates enable the user to make a distinction between better and worse predictions. The idea of the latter approach has also an additional advantage as well. Namely, the calculation of individual predictions' reliability estimates does not require true label values. Unlike MSE estimate, which requires a testing data set, the idea of individual predictions'

Z. Bosnić (✉) · I. Kononenko
University of Ljubljana, Faculty of Computer and Information
Science, Tržaška 25, Ljubljana, Slovenia
e-mail: zoran.bosnic@fri.uni-lj.si

I. Kononenko
e-mail: igor.kononenko@fri.uni-lj.si

Fig. 1 Reliability estimate for the whole regression model (*above*) in contrast to reliability estimates for individual predictions (*below*)

estimates is that they can be calculated for arbitrary unseen examples.

The existing estimates of prediction errors are founded on the quantitative description of distribution of learning examples in the problem space, in which algorithms usually make the i.i.d. (independent and identically distributed) assumption. Noise in data and nonuniform distribution of examples represent a challenge for learning algorithms, leading to different prediction accuracies in various parts of the problem space. Apart from distribution of learning examples, there are also other causes that influence the inaccuracy of prediction models: their generalization ability, bias, resistance to noise, avoidance of overfitting, etc. Since these aspects cannot be measured quantitatively, they cannot be composed into a common formula and used to estimate the prediction error. Therefore, we focus on an approach which enables us to analyze the *local* particularities of learning algorithms.

Our method is based on the *sensitivity analysis* [1]. Sensitivity analysis aims at determining how much the variation of input can influence the output of a system. Our approach is to locally modify the learning set in a controlled manner in order to explore the sensitivity of the regression model in a particular part of the problem space. By doing so, we adapt the reliability estimate to the local particularities of data distribution and noise. The sensitivity is thus related to changes of the prediction of the regression model when the learning set is slightly changed. Since the true error of an unlabeled example is not known, we use more appropriate term, saying that we estimate *prediction reliability* rather than *prediction error*. This conforms with the definition of *reliability* which can be defined as *the ability to perform certain tasks conforming to required quality standards* [2]. Namely, the prediction accuracy in regression is considered the required quality standard.

The paper is organized as follows. Section 2 introduces previous work from three related areas which are later jointly combined by our approach. In Sect. 3 we present the motivation for this research using the minimum description length (MDL) principle formalization. Section 4 defines our

sensitivity analysis task and illustrates the expected output of five regression models. In Sect. 5 we define three reliability estimates. These are tested and compared to another reliability estimation approach in Sect. 6. Section 7 provides conclusions and ideas for further work.

2 Related work

Our paper is related to previous work within three different research areas of the learning algorithms' properties. These areas are: the use of sensitivity analysis, perturbations of learning data in order to improve accuracy, and the estimation of the prediction reliability for single examples.

Sensitivity analysis Sensitivity analysis is an approach which has been applied to many areas such as statistics and mathematical programming. In the context of analyzing stability of learning algorithms it has been discussed by Bousquet and Elisseeff [1]. They defined notions of stability for learning algorithms and showed how to derive the generalization error bounds based on the empirical error and the leave-one-out error. They have also introduced the concept of β -stable learner as one for which the expected loss function of the learned solution does not change more than β with small changes in the training set. Bousquet and Elisseeff [3] and Elisseeff and Pontil [4] applied this ideas to several learning models and showed how to obtain bounds on their generalization performance.

In a similar way Kearns and Ron [5] define the *hypothesis stability* as a quantity that measures how much the function learned by the algorithm will change when one point in the training set is removed. All mentioned studies focus on dependence of error-bounds either from the VC (Vapnik–Chervonenkis) theory [17] or from the way the learning algorithm searches the space. By proving theoretical usefulness of notion of *stability*, these approaches motivated us to explore the possibilities for empirical estimation of the individual prediction reliability based on the local stability of the model.

Perturbation of learning data The group of approaches from the second related area is more general. These approaches generate perturbations of initial learning set to improve accuracy of final aggregated prediction. Bagging [6] and boosting [7–9] are the most popular in this field. Besides improving the predictive accuracy, *pasting* [10] also solves the prediction problem in data sets which are too large to fit in memory.

Tibshirani and Knight [11] introduced the *covariance inflation criterion* (CIC), which they use to improve the learning error by iteratively generating perturbed versions of the learning set. In each iteration they measure a covariance between input and predictor response and perform the model selection accordingly. The studies [12] have shown that CIC is a suitable measure for model comparison, even if we do not use cross-validation to estimate the model accuracy.

Elidan, et al. [13] introduced a strategy for escaping local maxima that also perturbs the training data instead of perturbing the hypotheses directly. They use reweighting of the training examples to create useful ascent directions in the hypothesis space and look for optimal solutions in the sets of perturbed problems. Their results show that such perturbations allow one to overcome local maxima in several learning scenarios. On both synthetic and real-life data this approach significantly improved models in learning structure from complete data, and learning both parameters and structure from incomplete data.

All mentioned approaches iteratively modify the learning set of examples in order to generate a series of learning models, on which they perform a model selection and choose the most accurate one. Though results in this field show that the perturbation approaches are justified in improving the general hypothesis accuracy score, the application to estimation/improvement of the single prediction accuracy was not yet studied. In our research we thus focus on this challenge.

Estimation of reliability for single examples Previous studies have referred to reliability of single predictions with different terms. Gammerman, Vovk, and Vapnik [14] and Saunders, Gammerman and Vovk [15] introduce the notions of *confidence* and *credibility*. *Confidence* indicates a confidence value for predicted classifications and *credibility* is an indicator of the reliability of the data upon which the prediction is made. Experiments with their modified Support Vector Machine algorithm showed that they successfully produced the confidence and credibility measures and outperformed other predictive algorithms.

Later, Nourtdinov, et al. [16] demonstrate the use of *confidence* value in the context of ridge regression. Using residuals of learning examples and a p-value function they improve an ordinary ridge regression with confidence regions.

The drawback of these approaches is that confidence estimations need to be specifically designed for each particular model and cannot be applied to other methods.

The notion of reliability estimation has most frequently appeared together with the notion of *transduction*, as for example in [14, 15]. Transduction is an inference principle that reasons from particular to particular [17] 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 of interest and not necessarily the whole input space. This locality enables the use of transductive algorithms for making other inferences besides predictions. We find inferences of reliability measures of special interest.

As an application of the above principle, Kukar and Kononenko [18] 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. Bosnić and Kononenko [19] later adapted the approach to regression. *Transductive predictions*, introduced by this technique, were used to model prediction error for each individual example. Initial results were promising and showed the potential for estimating and possibly correcting the prediction error.

Some other related theoretical work should also be mentioned. In the context of co-training, it has been shown that the unlabeled data can be used to improve the performance of a predictor [20]. It has also been shown that for every reasonable classifier (i.e., better than random) performance can be significantly boosted by utilizing only additional unlabeled data [21]. Both of these studies additionally encourage the use of utilizing additional learning examples, which we used in this research.

The work presented here extends the work of Kukar and Kononenko [18] to regression and places it in the context of the sensitivity analysis.

3 MDL based motivation

The dependence between input data and the prediction cannot be analytically expressed for most of the prediction models. This is true especially for models, which partition input space and establish local models for each separate partition. In contrast, the Minimum Description Length principle [22] offers a formalism based on the probabilistic and information theory. In this section we use MDL to introduce the motivation for our research. We show that it is possible to obtain additional information if we expand the learning data set with an additional example. This finding motivates us to use this information for estimation of the prediction reliability for individual examples.

In the following analysis we start by setting the basic MDL framework. We introduce the notions of absolute and relative reliability, which we define as the hypothesis' capability to achieve the largest compression of the data. Based on this general definition we express the reliability of a single example. Considering two extreme scenarios we demonstrate the behavior of the relative reliability. We conclude by showing that the expanding of learning set in the appropriate way may result in obtaining the extra information.

3.1 MDL preliminaries

Let us denote learning examples by E , background knowledge by B and a hypothesis by H . By definition of MDL, the optimal hypothesis maximizes the conditional probability, which is equal to minimizing the information (minus logarithm of probability) [22]:

$$H_{\text{opt}} = \arg \max_H P(H|E, B) = \arg \min_H -\log_2 P(H|E, B). \quad (1)$$

Using the Bayes theorem, the information $-\log_2 P(H|E, B)$ can be written as

$$-\log_2 P(H|E, B) = -\log_2 P(E|H, B) - \log_2 P(H|B) + \log_2 P(E|B) \quad (2)$$

and since $-\log_2 P(E|B)$ is independent of H , the optimal hypothesis H_{opt} can be obtained with:

$$H_{\text{opt}} = \arg \min_H [-\log_2 P(E|H, B) - \log_2 P(H|B)]. \quad (3)$$

The MDL criterion therefore tries to minimize the complexity of the hypothesis ($-\log_2 P(H|B)$) and the error ($-\log_2 P(E|H, B)$).

The problem with the above derivation is that in order to implement it one needs the optimal coding which, however, is not computable [22]. Therefore, in practice one introduces a practical coding and the MDL criterion (3) is transformed into

$$H_{\text{opt}} \sim \arg \min_H [I(E|H, B) + I(H|B)].$$

The term $I(H|B)$ represents the number of bits needed to encode the hypothesis (given the background knowledge) and the term $I(E|H, B)$ represents the number of bits needed to describe data E , given the hypothesis (and the background knowledge). However, note, that in this formula the practical (nonoptimal) coding is used. Therefore, the (nearly) optimal hypothesis is described with short hypothesis that explains a great portion of the data. The example of relationship between hypothesis complexity and prediction error is illustrated in Fig. 2.

Another point of view on the same situation is provided with the notion of compressivity. The hypothesis H is said

Fig. 2 Illustration of the MDL principle: the relation between the model complexity ($I(H|B)$) and the prediction error ($I(E|H, B)$). The model at the top has overly large error, while the model at the bottom has overly large complexity

to be compressive, if it allows for shorter description of data (in [22] this principle is formally described with the Occam's Razor Theorem):

$$I(E|H, B) + I(H|B) < I(E|B). \quad (4)$$

The term $I(E|B)$ represents the number of bits needed to encode the data (given the background knowledge) without the use of the hypothesis. The lesser the left-hand side term, the more compressive is the hypothesis H .

3.2 Definition of reliability

For a given set of learning data E and background knowledge B let the *most reliable hypothesis* be the one that *leads to the largest compression of the data*. Put differently, the most reliable hypothesis is the one that minimizes the left-hand side of the inequality (4).

Based on this definition we now introduce the quantitative measures of absolute and relative reliability.

Definition 1 Let H , E and B be defined as above. Hypothesis H has *reliability* R , defined as

$$R(H|E, B) = I(E|B) - I(E|H, B) - I(H|B). \quad (5)$$

Definition 2 Let H , E and B be defined as above. Hypothesis H has *relative reliability* R_{rel} , defined as

$$R_{\text{rel}}(H|E, B) = 1 - \frac{I(E|H, B) + I(H|B)}{I(E|B)}. \quad (6)$$

3.3 Reliability of a single prediction

Using the above definitions we now consider the reliability of a single example. We start by expanding the learning set $E = \{e_1, e_2, \dots, e_n\}$, $|E| = n$ with an additional learning example e_{n+1} . Inequality in (4) thus becomes

$$I(E \cup \{e_{n+1}\}|\bar{H}, B) + I(\bar{H}|B) < I(E \cup \{e_{n+1}\}|B) \quad (7)$$

where $\bar{H}$ stands for new, modified hypothesis, which also covers (predicts) example e_{n+1} . Since learning sets of H and $\bar{H}$ differ only in one learning example, we may assume that the both hypotheses are very similar. For sufficiently large data sets we therefore assume that

$$I(H|B) \approx I(\bar{H}|B), \quad (8)$$

$$I(e_i|H, B) \approx I(e_i|\bar{H}, B), \quad i = 1, \dots, n \quad (9)$$

and use these simplifications in the following definition.

Definition 3 Let $R_1(e, H)$ denote the reliability of prediction of example e by hypothesis H . We define the reliability of a single prediction as the difference between reliabilities of hypotheses H and $\bar{H}$:

$$R_1(e_{n+1}, \bar{H}) = R(\bar{H}|E \cup \{e_{n+1}\}, B) - R(H|E, B). \quad (10)$$

The first term at the right hand side can be expanded as in (5). Therefore, by substituting

$$\begin{aligned} R(\bar{H}|E \cup \{e_{n+1}\}, B) \\ = I(E \cup \{e_{n+1}\}|B) - I(E \cup \{e_{n+1}\}|\bar{H}, B) - I(\bar{H}|B) \end{aligned} \quad (11)$$

and by using the following simplifications

$$I(E|B) = I(e_1|B) + \dots + I(e_n|B), \quad (12)$$

$$I(E \cup \{e_{n+1}\}|B) - I(E|B) = I(e_{n+1}|B), \quad (13)$$

$$I(E|H, B) = I(e_1|H, B) + \dots + I(e_n|H, B) \quad (14)$$

and the simplification from (8) we get the intermediate result

$$\begin{aligned} R_1(e_{n+1}, \bar{H}) &= I(e_{n+1}|B) - I(E \cup \{e_{n+1}\}|\bar{H}, B) \\ &\quad + I(E|H, B). \end{aligned} \quad (15)$$

Plugging (9) and (14) into the latter result we rewrite the definition of reliability in (10) as

$$R_1(e_{n+1}, \bar{H}) \approx I(e_{n+1}|B) - I(e_{n+1}|B, \bar{H}). \quad (16)$$

We see that this form expresses the reliability of a single prediction as the difference of information terms, which depends on the expanded hypothesis $\bar{H}$ only and does not depend on the original hypothesis H . Furthermore, note that $I(e_{n+1}|B)$ is independent of $\bar{H}$ and plays a role of a constant term. This form of definition is suitable for analysis for the two scenarios, which we analyze in the following subsection.

3.4 Analysis of reliability in extreme cases

Let us now consider two cases, which can occur when expanding the learning set with an additional example, followed by constructing the new hypothesis $\bar{H}$. In the two extreme situations, hypothesis $\bar{H}$ can either contain the *maximum* information for prediction of the new example, or *none* at all. In each of extreme cases, the relative reliability R_{rel} changes its value as follows:

1. If hypothesis $\bar{H}$ contains the maximum information for prediction of the new example, then the term at the right hand side of (16) achieves the maximum value, thus $I(e_{n+1}|B, \bar{H}) = 0$ and

$$R_1(e_{n+1}, \bar{H}) \approx I(e_{n+1}|B). \quad (17)$$

As it holds $R_{\text{rel}}(H|E, B) < 1$ (see (6)) and:

$$\begin{aligned} R_{\text{rel}}(\bar{H}|E \cup \{e_{n+1}\}, B) &= \frac{R(\bar{H}|E \cup \{e_{n+1}\}, B)}{I(E \cup \{e_{n+1}\}|B)} \\ &= \frac{R(H|E, B) + R_1(e_{n+1}, \bar{H})}{I(E|B) + I(e_{n+1}|B)} \end{aligned}$$

the comparison of the relative reliabilities of hypotheses H (before adding a new example e_{n+1}) and $\bar{H}$ (after adding e_{n+1}) gives:

$$R_{\text{rel}}(\bar{H}|E \cup \{e_{n+1}\}, B) > R_{\text{rel}}(H|E, B). \quad (18)$$

since

$$\frac{R(H|E, B) + I(e_{n+1}|B)}{I(E|B) + I(e_{n+1}|B)} > \frac{R(H|E, B)}{I(E|B)}.$$

Equation (18) shows that the relative reliability of the hypothesis has *increased*. In other words, we state that an optimal learning algorithm modifies hypothesis H into $\bar{H}$ by including the accurate knowledge for prediction of the new example. This leads to the conclusion that the former hypothesis was unreliable (i.e. it achieved insufficient compression of the data) in that particular part of the problem space.

2. If hypothesis $\bar{H}$ includes no information for prediction of the new example, then term (16) achieves its minimum, thus:

$$R_1(e_{n+1}, \bar{H}) \approx I(e_{n+1}|B) - I(e_{n+1}|B, \bar{H}) \approx 0. \quad (19)$$

Comparing the relative reliabilities of hypotheses H and $\bar{H}$ in this case gives

$$R_{\text{rel}}(\bar{H}|E \cup \{e_{n+1}\}, B) < R_{\text{rel}}(H|E, B) \quad (20)$$

since $R(\bar{H}|E \cup \{e_{n+1}\}, B) = R(H|E, B)$ and $I(E \cup \{e_{n+1}\}|B) > I(E|B)$, meaning that the relative reliability of the hypothesis has *decreased*. In contrast to conclusions in the former extreme case, this means that the hypothesis H was more reliable (i.e. achieved better compression of the data) in that part of the problem space. However, neither H nor $\bar{H}$ cover e_{n+1} (for H it was not available for learning and $\bar{H}$ considers it a noisy example).

Based on the above findings we now state a result about hypotheses H and $\bar{H}$ for some optimal learning algorithm.

Theorem 1 *Let e_{n+1} , H , $\bar{H}$ and B be as defined above. If $I(e_{n+1}|H, B) = 0$ for some optimal learning algorithm, then the following is true:*

$$I(e_{n+1}|\bar{H}, B) = 0. \quad (21)$$

Proof Recall the property of the optimal learning algorithm,

$$H = \arg \min_T [I(E|T, B) + I(T|B)], \quad (22)$$

$$\bar{H} = \arg \min_T [I(E \cup \{e_{n+1}\}|T, B) + I(T|B)]. \quad (23)$$

If we assume that $I(e_{n+1}|\bar{H}, B) > 0$, then

$$I(E|\bar{H}, B) + I(\bar{H}|B) < I(E|H, B) + I(H|B)$$

meaning that H is not the optimal hypothesis for E , leading to contradiction. $\square$

The negation of the above theorem states that if $I(e_{n+1}|\bar{H}, B) > 0$, then either the algorithm is suboptimal

or the hypothesis H does not contain the maximum knowledge for prediction of e_{n+1} , i.e. $I(e_{n+1}|H, B) > 0$. Taking into account that the choice of the algorithm is fixed for both H and $\bar{H}$, we leave aside the discussion about the algorithm's optimality and focus rather on emphasizing the use of this in the sensitivity analysis. The latter result namely states that *the most of information is achieved if we expand the learning set with an example that is not well covered by the initial hypothesis H* . I.e., by expanding the learning set with such example, we test the two extreme cases, described above. Either the optimal learning algorithm will modify the hypothesis to obtain good coverage (case 1) or it will leave the hypothesis unchanged therefore indicating that the new example is considered noisy (case 2). This finding motivates for putting this extra information in use for estimating the reliability of individual predictions.

4 Sensitivity of regression models

The aim of regression predictors is to model learning examples by minimizing error on the learning and test data. By adding or removing an example from the learning set, thus making a minimal change to the input of the learning algorithm, one can expect that the change in output prediction for the modified example will also be small. Big changes in output prediction that result by making small changes in learning data may be a sign of instability in the generated model. The magnitude of output change may therefore be used as a measure of model instability for a modified example. In our research we focus precisely on using these instability measures and combining them into reliability estimates.

4.1 Modification of input

There are several possible ways to modify the learning data set. The theoretical result from Sect. 3 indicates that for the analysis we should use examples that are not well covered by the hypothesis. We decided to expand it with the additional learning example as follows. Let x represent the example and let y be its label. Therefore, with (x, y) we denote a learning example with known/assigned label y and with $(x, _)$ we denote an unseen example with unknown label. Let $(x, _)$ be the unseen and unlabeled example, for which we wish to estimate the reliability of prediction. The prediction K , which we are estimating, is made by regression model M , therefore $f_M(x) = K$. Since K was predicted at the beginning of the sensitivity analysis using an unmodified learning set of size n , we will refer to it as *initial prediction*. To expand the learning set with example $(x, _)$ we first label it with

$$y = K + \delta \quad (24)$$

Fig. 3 The sensitivity analysis process. The figure illustrates the obtaining of initial prediction (phase 1) and the sensitivity model with sensitivity prediction K_ε (phase 2)

where δ denotes some small change. Note, that we use the initial prediction K as a central value for y , which is afterwards incremented by term δ (which may be either positive or negative). We define such δ , which is proportionate to known bounds of label values. In particular, if the interval of learning examples' labels is denoted by $[a, b]$ and if ε denotes a value that expresses the relative portion of this interval, then $\delta = \varepsilon(b - a)$.

After selecting ε and labeling the new example, we expand the learning data set with example (x, y) . Based on this modified learning set with $n + 1$ examples we build a model, to which we refer to as the *sensitivity regression model* M' . We use M' to predict $(x, -)$, thus having $f_{M'}(x) = K_\varepsilon$. Let us refer to K_ε as the *sensitivity prediction*. The described procedure is illustrated in Fig. 3. By selecting different $\varepsilon_k \in \{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m\}$ to obtain y , we in fact iteratively obtain a set of sensitivity predictions

$$K_{\varepsilon_1}, K_{-\varepsilon_1}, K_{\varepsilon_2}, K_{-\varepsilon_2}, \dots, K_{\varepsilon_m}, K_{-\varepsilon_m}. \quad (25)$$

The obtained sensitivity predictions serve as output values in the sensitivity analysis process. As mentioned above, we use all of above differences $K_\varepsilon - K$ to observe the model stability and combine them into different reliability measures. Having introduced the needed terminology, we now illustrate in Fig. 4 our expected assumptions of how do differences $K_\varepsilon - K$ characterize *reliable* and *unreliable* predictions of K .

At this point let us make a remark that in the process of assigning a label one cannot take (local) bias and variance into account as one does not have their local estimates for a single new data point. In fact, local bias and variance are precisely what we expect to capture by locally testing the model sensitivity. In the remainder, we aim to estimate the local variance (reliability measures RE_1 and RE_2 , described later) and bias (measure RE_3). Before introducing these measures, let us take a brief look at output characteristics of some common regression models.

Fig. 4 Reliability defined as prediction sensitivity (three examples)

4.2 Sensitivity of the output

Based on the procedure for acquiring a *sensitivity prediction*, one could intuitively expect the following to hold:

$$\varepsilon_1 < \varepsilon_2 \Rightarrow K_{\varepsilon_1} < K_{\varepsilon_2} \quad (26)$$

and therefore

$$K_{-\varepsilon_m} < \dots < K_{-\varepsilon_1} < K < K_{\varepsilon_1} < K_{\varepsilon_2} < \dots < K_{\varepsilon_m}$$

for $\varepsilon_k \in \{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_m\}$, $\varepsilon_1 < \varepsilon_2 < \dots < \varepsilon_m$, $k = 1, \dots, m$. Although this empirically holds for the majority of cases, the relative ordering of initial and sensitivity predictions depends on the regression model itself.

We tested the proposed technique with five regression predictors: regression trees, locally weighted regression, neural networks, support vector machines and linear regression. These learning algorithms might be divided to *simple* and *complex*, regarding whether they divide the input space prior to modeling data or not. We consider linear regression and locally weighted regression to be simple models since they model the whole input data at once. In contrast, prior to

Fig. 5 Changes of regression tree after modifying the learning data set

modeling, other three models (regression trees, neural networks and support vector machines) perform either partitioning of space or other example selection: regression trees partition according to attribute values, series of neurons act like a complex discriminant function which partitions examples and SVM model depends only on examples at the margin.

For sensitivity analysis approach, the complex algorithms are more interesting than simple algorithms. Instead of a slight change in the output of complex models, the additional example may cause different partitioning of the input space, thus leading to a different hypothesis. This may also result in a big difference between initial and sensitivity predictions, which indicates that the initial hypothesis for tested example is unstable or unreliable. We illustrate this phenomenon in Fig. 5 for regression trees on a benchmark data set *pollution* [24]. Figure 5(a) displays the initial regression tree built on the original learning set. The first possible scenario is that by utilizing an additional example the model changes slightly, as shown in Fig. 5(b). The difference between initial and second model is a changed prediction value in the third leaf of the tree (prediction for utilized example changes

from 922.8 to 924.1). The other alternative is that the utilization of additional example causes a change in the structure of the hypothesis, as shown in Fig. 5(c). We can observe that the splitting criteria and predictive values in the subtree at the right-hand side change substantially in this case. Although the positive δ was used, the prediction for utilized example decreases from 922.8 to 906.4. This example illustrates that the rule (26) does not hold for cases, when the structure of the hypothesis changes. We do not consider this as a drawback for our approach, but rather as an indicator of the hypothesis unreliability, which we try to estimate.

Simple and complex models also differ in a way that in simple models, *sensitivity prediction* can be expressed as a functional dependency of *initial prediction* and the new example. Let us now take a look at two examples, one for each of the *simple* models.

Example 1 The prediction of example x using the locally weighted regression is:

$$K = \frac{\sum_j^N \kappa(D(\mathbf{x}, \mathbf{x}_j)) \cdot C_j}{\sum_j^N \kappa(D(\mathbf{x}, \mathbf{x}_j))},$$

where D represents a distance function, κ a kernel function, C_j the true label of example j and $\mathbf{x}$ the attribute vector of our example. The sensitivity prediction K_ε can then be expressed as:

$$\begin{aligned} K_\varepsilon &= \frac{\sum_j^N [\kappa(D(\mathbf{x}, \mathbf{x}_j)) \cdot C_j] + [\kappa(D(\mathbf{x}, \mathbf{x})) \cdot (K + \delta)]}{\sum_j^N [\kappa(D(\mathbf{x}, \mathbf{x}_j))] + [\kappa(D(\mathbf{x}, \mathbf{x}))]} \\ &= \frac{\sum_j^N [\kappa(D(\mathbf{x}, \mathbf{x}_j)) \cdot C_j] + [\kappa(0) \cdot (K + \delta)]}{\sum_j^N [\kappa(D(\mathbf{x}, \mathbf{x}_j))] + [\kappa(0)]}. \end{aligned}$$

Example 2 Let us assume the case of linear regression in two-dimensional space. We are thus trying to model dependency $y = kx + n$, where k is a regression line slope and n its intercept. Then,

$$\begin{aligned} K &= kx + n, \\ k &= \frac{\sum_k (x_k - \bar{x})(y_k - \bar{y})}{\sum_k (x_k - \bar{x})^2} = \frac{n \sum_k x_k y_k - \sum_k x_k \sum_k y_k}{n \sum_k x_k^2 - (\sum_k x_k)^2}, \\ n &= \bar{y} - k\bar{x} \end{aligned}$$

where n represents the number of examples and $\bar{x}$ denotes the average value. Sensitivity prediction K_ε can then be expressed as:

$$\begin{aligned} K_{i,\varepsilon} &= k_\varepsilon x_i + n_\varepsilon, \\ k_\varepsilon &= \frac{(n+1)[\sum_{k=1}^n x_k y_k + x(K+\delta)] - [\sum_{k=1}^n x_k + x][\sum_{k=1}^n y_k + K + \delta]}{(n+1)[\sum_{k=1}^n x_k^2 + x^2] - (\sum_{k=1}^n x_k + x)^2}, \\ n_\varepsilon &= \frac{\bar{y}n + K + \delta}{n+1} - k_\varepsilon \frac{\bar{x}n + x}{n+1}. \end{aligned}$$

Here x_k , $k = 1, \dots, n$ represent learning examples and x represent a new example.

For *complex* models it is not possible to express such functional dependency.

The magnitude of change in model depends also on the size of δ (or ε) used in (24). Using different values of δ we can depict model sensitivity in plots, shown in Fig. 6. The plots show empirical dependency of $K_\varepsilon - K$ (vertical axis) on $\varepsilon \in [-5, 5]$ (horizontal axis) for a typical example, using the benchmark data set *strikes*. It is obvious that the plots for locally weighted regression (Fig. 6(d)) and linear regression (Fig. 6(e)) do not show any local anomalies which could be captured by the reliability measures.

The critical local regions, which emphasize the sensitivity, can be identified for *complex* models. For regression trees (Fig. 6(a)), the local regions are limited to single leaf of the tree. For support vector machines (Fig. 6(c)), the critical local region is defined with the hyperplane margin. And for neural networks (Fig. 6(b)), the critical local regions are not so obvious. When the label is approaching the extreme

values, the added example is more and more considered as an outlier, and the prediction therefore becomes more dependent on the other learning examples.

Based on the above examples and plots in Fig. 6 we can conclude that the determinism in *simple* models does not allow us to capture what we defined as the unreliable model behavior. The tests, performed using linear regression and locally weighted regression, also confirmed this conclusion. In Sect. 6 we therefore focus on testing our technique with *complex* predictors, i.e. regression trees, neural networks and SVMs.

5 Reliability estimates

We use the differences between predictions of initial and sensitivity models as an indicator of the prediction reliability. At this point we combine these differences into three reliability estimates.

The calculation of the sensitivity prediction requires the selection of particular ε , as defined in (24). To avoid selecting a particular ε , we define measures to use an arbitrary number of sensitivity predictions (defined by using different ε parameters). In this way we widen the observation window for observing model instabilities and make the measures robust to local anomalies in the problem space. The number of used ε parameters therefore represents a trade-off between gaining more stable reliability estimates and the total computational time. Since we assumed that the zero-difference between predictions represents the maximum reliability, we also define the reliability measures so that value 0 indicates the most reliable prediction.

Let us assume we have a set of non-negative ε values $E = \{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_{|E|}\}$. We define the estimates as follows:

1. Estimate RE_1 (local variance):

$$RE_1 = \frac{\sum_{\varepsilon \in E} (K_\varepsilon - K_{-\varepsilon})}{|E|}. \quad (27)$$

In the case of reliable predictions we expect that the change in sensitivity model for K_ε and $K_{-\varepsilon}$ would be minimal (0 for the most reliable predictions). We define this reliability measure using both, K_ε and $K_{-\varepsilon}$, to capture the model instabilities not regarding the sign of δ in (24). The measure takes the average of differences across all values of ε . In Fig. 4, the RE_1 estimate represents the wideness of the whole interval, therefore it corresponds to local variance.

Fig. 6 Sensitivity of regression prediction to changes in a single learning example. The plots show empirical dependency of $K_\varepsilon - K$ (vertical axis) on $\varepsilon \in [-5, 5]$ (horizontal axis) for a typical example, using the benchmark data set *strikes*

2. Estimate RE_2 (local absolute variance):

$$RE_2 = \frac{\sum_{\varepsilon \in E} |K_\varepsilon - K| + |K_{-\varepsilon} - K|}{2|E|}. \quad (28)$$

In contrast to RE_1 , RE_2 measures the difference between predictions of initial and sensitivity models. The estimate takes the average of $|K_\varepsilon - K|$ and $|K_{-\varepsilon} - K|$ and therefore measures the average change of the prediction using positive

and negative δ . This measure also takes the average across all ε parameters and is defined as non-negative.

3. Estimate RE_3 (local bias):

$$RE_3 = \frac{\sum_{\varepsilon \in E} (K_\varepsilon - K) + (K_{-\varepsilon} - K)}{2|E|}. \quad (29)$$

We define RE_3 in a similar way as RE_2 . In contrast to RE_2 , RE_3 can be either positive or negative. Its sign carries in-

formation about the direction in which the predictor is more unstable. The measure is also averaged across all ε . In Fig. 4, the RE_3 estimates the skewness (the asymmetry between left and right subintervals) and therefore corresponds to local bias.

Note that all three reliability estimates are very similar to the symmetrized form of formula for calculation of the numerical derivative $K'(\varepsilon)$ [23]. The function derivative, i.e. the slope of a function, is an indicator of the function sensitivity at the given point, which is in accordance with our definition of reliability estimates.

In the experiments we correlate reliability measures RE_1 and RE_2 with the absolute value of the prediction error (residual). Since the value of RE_3 contains significant information about the direction of the model sensitivity, its value is correlated to the signed value of the prediction error.

6 Experimental results

6.1 Sensitivity-based estimation of prediction error

The reliability estimates were tested on 48 standard benchmark data sets, which are used 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. The number of examples in this domains varies from 20 to over 6500. Most of the data sets are available from UCI Machine Learning Repository [24] and from StatLib DataSets Archive [25]. All data sets are available from authors upon request. The brief description of data sets is given in Table 1.

As explained in Sect. 4.2 we experimented with five regression models. We present results only for regression trees, neural networks and SVMs, as our technique is inadequate for linear regression and locally weighted regression. Some key properties of used models are:

- Regression trees: the mean squared error is used as the splitting criterion, the value in leaves represents the average label of examples, trees are pruned using the m-estimate [26].
- Neural networks: one hidden layer of neurons, the learning rate was 0.5, the stopping criterion for learning is based on the change of MSE between two backprop iterations.
- Support vector machines: The ϵ -support vector regression algorithm from the LIBSVM library is used [27], we use the third-degree RBF kernel, the precision parameter was $\epsilon = 0.001$.

The testing of model sensitivity was performed similarly to cross-validation. The data set was divided into ten subsets and the sensitivity predictions were calculated for

each of the examples in the excluded subset. This was repeated for all ten data folds, thus obtaining initial and sensitivity predictions for all examples in the testing data set. When calculating a sensitivity prediction for each example, the learning set was expanded with the additional learning example. This change in the learning set was not permanent, but the changes were discarded before the calculation of the new sensitivity prediction (original data set with excluded subset was used). For calculation of reliability estimates, five different values of ε parameter were used: $E = \{0.01, 0.1, 0.5, 1.0, 2.0\}$. For each of the examples in the data set, the reliability estimates were correlated with their prediction error using the Pearson's correlation coefficient. The significance of correlation coefficient was evaluated using the t -test for correlation coefficients. The described testing procedure is illustrated in Fig. 7.

The summarized results are shown in Table 2. The table displays percentages of experiments in which we achieved significant results. Detailed results for individual domains are shown in Table 3. Results confirm our expectations that the reliability estimates should positively correlate with the prediction error. We can see that the positive correlations highly outnumber the negative correlations with all regression models and reliability estimates.

The best summarized results were achieved using estimate RE_3 (local bias), and the worst, although still good with RE_1 (local variance). The best estimate, RE_3 , significantly positively correlated with prediction error in 48% of tests (negatively in 3%). The result, which stands out the most, is the performance of RE_3 with regression trees. In this case, RE_3 significantly positively correlated in 75% of tests and negatively in none. Analyzing from different perspective, we can also see that the estimates perform best with regression trees and the worst, although still quite well with SVMs.

Summarized results show the potential of using the proposed reliability estimates for estimation of the prediction error. We proceed by comparing our results to another approach.

6.2 Density-based estimation of prediction error

A traditional approach to estimation of the prediction confidence/reliability is based on distribution of learning examples. In this section we use term *subspace* which refers to the subset of learning examples, which are related by locality regarding their attributes' values. If two subspaces of the same size are compared to one another, the first one having the greater number of learning examples than the other, we refer to the first one as the *denser* subspace.

The density-based estimation of prediction error assumes that error is lower for predictions which are made in *denser* problem subspaces, and higher for predictions which are

Table 1 Basic characteristics of testing data sets

Data set	# examples	# disc.attr.	# cont.attr.
abalone	4177	1	7
audio	200	69	0
auto_price	159	1	14
auto93	93	6	16
autohorse	203	8	17
autompg	398	1	6
balance	625	0	4
basketball	96	0	4
bhouse	506	1	12
bodyfat	252	0	14
brainsize	20	0	8
breasttumor	286	1	8
cholesterol	303	7	6
cleveland	303	7	6
cloud	108	2	4
cos4	1000	0	10
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
meta	528	2	19
pbcc	418	8	10
pharynx	195	4	7
photo	858	2	3
places	329	0	8
plasma_carotene	315	3	10
plasma_retinol	315	3	10
pollution	60	0	15
pwlin2	200	0	10
pwlinear	200	0	10
pyrim	74	0	27
quake	2178	0	3
sensory	576	0	11
servo	167	2	2
sleep	58	0	7
stock	950	0	9
strikes	625	0	5
transplant	131	0	2
triazines	186	0	60
tumor	86	0	4
wind	6574	0	11
wpbc	198	0	32

Fig. 7 Illustration of the testing procedure

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

Correlation	RE_1		RE_2		RE_3		Average	
	+	−	+	−	+	−	+	−
Regression tree	31%	4%	52%	2%	75%	0%	53%	2%
Neural network	42%	2%	46%	0%	44%	4%	44%	2%
SVM	35%	4%	35%	4%	25%	4%	32%	4%
Average	36%	3%	44%	2%	48%	3%	43%	3%

made in sparser subspaces. This means that we trust the prediction with regard to the *quantity of information* at disposal for calculation of the prediction. A typical use of this approach is, for example, with decision and regression trees, where we trust each prediction according to proportion of learning examples that fall in the same leaf of a tree as the predicted example.

We estimated density using a nonparametric estimation of probability distribution [28]. This approach is called *the kernel estimator* or *Parzen windows*. Similar to experiments in Sect. 6.1, we correlated the density estimates with absolute prediction errors (absolute residuals). With regard to definition of the density-based error estimation we expected error to negatively correlate with density estimate. For calculating densities we used ten-fold cross-validation procedure in a similar manner as in Sect. 6.1. For each testing example the local density was estimated and correlated with its prediction error.

The summarized results are shown in Table 4. The table displays percentages of experiments in which we achieved significant results. Detailed results for individual domains are shown in Table 5. Results confirm that the negative correlations outnumber positive correlations. Nevertheless, results of density estimates are worse than those for estimates RE_1 , RE_2 and RE_3 . Note, that for our measures, the desired correlations are positive, while for the density estimates, the desired correlations are negative. The number of desired correlations obtained by our approach is greater (43% ver-

sus 33%) and the number of undesired correlations much smaller (3% versus 8%).

Better correlations of the proposed estimates with prediction error show that they provide more information than the probabilistic distribution of examples. They measure the prediction sensitivity, which also depends on learning and on predictive algorithms themselves. We can conclude that estimates RE_1 , RE_2 and RE_3 are more suitable for estimation of the prediction error than density estimates. However, it remains an open question how to adapt the three proposed measures in order to achieve the probabilistic interpretability as offered by the density function. We leave this for further work.

7 Conclusion

The paper presents a new method for reliability estimation of individual predictions. Our method is based on the sensitivity analysis, which is an approach that observes the output response with respect to small changes in the input data set.

Previous work in this field inspired us to modify the learning set by expanding it with an additional example. This is similar to ideas of Bousquet and Elisseeff [1] and Kearns and Ron [5]. Using the difference in predictions of initial and sensitivity models, we compose the reliability estimates RE_1 (local variance), RE_2 (local absolute variance) and RE_3 (local bias). We base these estimates with an effort

Table 3 Correlation coefficients between reliability estimates and prediction error. Cell shading represents the p-values. The data with significance level $\alpha \leq 0.05$ is marked by light grey (significant positive correlation) and dark grey (significant negative correlation) background

	REG. TREES			NEURAL NET.			SVM		
	RE ₁	RE ₂	RE ₃	RE ₁	RE ₂	RE ₃	RE ₁	RE ₂	RE ₃
abalone	0.096	0.157	0.256	0.130	0.135	-0.019	0.061	0.060	-0.063
audio	0.244	0.223	0.105	0.269	0.272	0.098	0.373	0.373	-0.006
auto_price	0.276	0.307	0.404	0.402	0.402	0.052	-0.030	-0.030	0.161
auto93	0.020	0.095	0.423	0.206	0.206	-0.041	-0.012	-0.012	0.126
autohorse	-0.152	0.105	0.270	0.269	0.269	0.072	0.009	0.009	0.397
autompg	0.101	0.159	0.379	0.378	0.378	0.232	0.149	0.149	0.100
balance	0.502	0.508	0.290	0.336	0.485	0.089	0.050	0.049	-0.035
basketball	-0.129	-0.087	0.218	0.184	0.177	0.130	0.191	0.191	0.129
bhouse	0.169	0.381	0.512	0.218	0.251	0.225	0.260	0.261	-0.031
bodyfat	-0.181	-0.181	0.281	0.006	0.030	0.380	-0.080	-0.080	0.034
brainsize	-0.024	-0.257	0.043	0.041	0.023	0.318	-0.225	-0.225	-0.272
breasttumor	-0.054	-0.025	0.173	0.069	0.069	0.251	-0.067	-0.067	0.160
cholesterol	0.091	0.145	0.331	0.046	0.058	0.203	0.023	0.023	0.092
cleveland	0.33	0.344	0.383	0.294	0.303	0.160	0.173	0.173	0.047
cloud	0.289	0.410	0.378	0.326	0.327	0.054	0.228	0.228	-0.160
cos4	0.05	0.233	0.430	0.239	0.298	0.273	0.159	0.159	0.012
cpu	0.032	0.016	0.148	0.557	0.56	0.222	0.497	0.497	0.486
diabetes	0.185	0.327	0.434	0.178	0.168	-0.155	0.123	0.124	-0.148
echomonths	0.245	0.235	0.097	0.020	0.103	0.206	0.038	0.038	0.096
elusage	0.191	0.367	0.358	0.275	0.275	0.065	0.066	0.066	0.147
fishcatch	0.001	-0.057	0.272	-0.612	-0.079	-0.277	-0.067	-0.067	0.380
fruitfly	-0.14	-0.086	0.258	-0.101	-0.100	-0.004	-0.067	-0.067	-0.006
grv	0.044	0.089	0.319	0.197	0.242	0.273	0.137	0.137	0.081
hungarian	0.382	0.179	0.208	-0.092	-0.106	-0.016	0.170	0.170	0.142
lowbwt	-0.038	-0.038	0.022	0.058	0.067	0.245	-0.006	-0.006	0.142
mbagrade	0.056	0.085	0.029	0.008	0.016	0.334	-0.126	-0.125	0.063
meta	-0.023	-0.019	0.059	-0.018	-0.017	-0.018	-0.210	-0.210	0.088
pbc	0.045	0.085	0.224	0.058	0.025	0.112	0.020	0.020	-0.007
pharynx	0.188	0.215	0.321	0.138	0.153	0.231	0.189	0.189	0.138
photo	0.135	0.158	0.191	0.117	0.441	0.392	0.069	0.071	0.051
places	0.035	0.145	0.402	-0.045	-0.044	0.009	-0.051	-0.051	0.011
plasma_carotene	-0.037	-0.022	0.253	0.099	0.125	0.110	-0.049	-0.049	-0.035
plasma_retinol	0.102	0.154	0.351	0.172	0.176	-0.041	0.107	0.107	0.224
pollution	-0.137	-0.211	0.083	0.153	0.155	0.060	0.127	0.127	0.222
pwlin2	0.036	0.055	0.444	0.010	0.010	0.002	0.160	0.160	0.057
pwnlinear	0.069	0.067	0.392	0.052	0.053	0.076	0.181	0.181	0.205
pyrim	-0.161	-0.004	0.099	0.005	0.008	0.165	0.317	0.317	0.178
quake	0.009	0.029	0.024	0.011	0.011	0.037	0.054	0.051	0.063
sensory	0.124	0.124	0.224	0.015	0.040	0.334	-0.069	-0.069	0.101
Servo	0.506	0.525	0.482	0.493	0.494	0.098	0.089	0.089	0.039
sleep	-0.187	-0.232	0.185	-0.098	-0.098	-0.007	0.066	0.066	0.085
stock	0.04	0.088	0.336	0.180	0.147	-0.229	-0.379	-0.380	0.045
strikes	0.004	0.014	0.003	0.006	0.031	0.036	0.019	0.020	0.026
transplant	0.334	0.219	0.180	0.530	0.533	0.234	0.604	0.605	-0.303
triazines	0.11	0.184	0.316	0.128	0.138	0.106	0.022	0.022	0.066
tumor	0.111	0.014	0.189	-0.035	0.019	0.390	0.027	0.027	-0.008
wind	0.015	0.057	0.255	0.086	0.108	0.070	0.057	0.057	0.020
wdbc	0.101	0.255	0.490	0.095	0.139	0.177	-0.014	-0.014	0.014
+	15 (31%)	25 (52%)	36 (75%)	20 (42%)	22 (46%)	21 (44%)	17 (35%)	17 (35%)	12 (25%)
-	2 (4%)	1 (2%)	0 (0%)	1 (2%)	0 (0%)	2 (4%)	2 (4%)	2 (4%)	2 (4%)

Table 4 Percentage of positive and negative correlations between *density estimates* and *prediction error*

Correlation	Density	
	+	–
Regression tree	10%	35%
Neural network	8%	29%
SVM	4%	35%
Average	8%	33%

to measure instabilities of regression models that arise from the learning algorithm itself. We explain why this methodology is not appropriate for *simple* algorithms (linear regression, locally weighted regression) and focus on testing it with three *complex* regression models: regression trees, neural networks and support vector machines. We assume that the proposed approach shall perform well with other *complex* regression models, which have to be empirically verified in future.

Experiments show that the proposed estimates better correlate with prediction error than common density estimates. The most promising results were achieved using RE_3 (local bias) which seems to be a good candidate for a general reliability measure. This estimate has an additional advantage, as it was correlated to the non-absolute value of the prediction error. It therefore holds a potential for correction of prediction errors.

Compared to a traditional approach, which estimates reliability based only on the distribution of examples in the problem space, the proposed approach implicitly considers local particularities of the learning problem, including the learning algorithm's generalization ability, bias, resistance to noise, amount of noise in data, avoidance of overfitting, etc. These aspects, most of which cannot be measured quantitatively, are analyzed implicitly by applying the sensitivity analysis approach and thus considering the learning problem as a black box. We can conclude that this is the reason why the proposed sensitivity estimates compare favorably to density estimates and also achieve better experimental results.

Better correlations of the proposed estimates with prediction error show that they provide more information than the probabilistic distribution of examples. They measure the prediction sensitivity, which also depends on learning and on predictive algorithms themselves. We can conclude that estimates RE_1 , RE_2 and RE_3 are more suitable for estimation of the prediction error than density estimates. However, it remains an open question how to adapt the three proposed measures in order to achieve the probabilistic interpretability as offered by the density function. We leave this for further work.

Related work [18] proposed estimates for classification reliability, which are based on the change of the posterior class distribution. In contrast to this approach, our estimates

are based solely on the outputs, given by the prediction system, and therefore do not require any estimations of distribution functions. The use of this approach is possible due to the continuous nature of predicted values in regression. Namely, this enables us to numerically express the difference between two regression predictions, in contrast to classification, where one can only observe, whether the predicted class was the same or different.

Besides additional comparisons with other techniques, the ideas for further work in this field include the improvement of interpretability of the proposed estimates. Namely, the values of proposed estimates are not bounded to any particular interval, and since they are based on prediction of dependent variable, their numerical values depend on that variable domain. This consequently means that the reliability values for predictions of two different data sets cannot be compared to each other. It shall therefore be appropriate to transform the estimates to a unique interval with (hopefully) a probabilistic interpretation. The estimates shall be appropriately mapped into interval $[0, 1]$, with value 0 representing an unreliable prediction and 1 representing the most reliable one. The notion of *prediction reliability* shall be expanded to *confidence interval* and it shall also be tested whether the reliability estimates can be used to correct the initial predictions and thus improve their accuracy.

The proposed method 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 research showed that this is a difficult prediction problem, because the possibility for recurrence is continuously present for almost 20 years after the treatment. The bare recurrence predictions were therefore complemented with our reliability estimates, helping the doctors with the additional validation of predictions' accuracy. The promising preliminary results confirm the usability of our approach.

Acknowledgements We thank Matjaž Kukar and Marko Robnik-Šikonja for their contribution to this study.

Table 5 Correlation coefficients between density estimates and prediction error. Cell shading represents the p-values. The data with significance level $\alpha \leq 0.05$ is marked by light grey (significant positive correlation) and dark grey (significant negative correlation) background

	REG. TREES	NEURAL NET.	SVM
abalone	0,053	0,060	0,050
audio	-0,199	-0,269	-0,262
auto93	-0,036	-0,061	-0,105
auto_price	-0,374	-0,392	-0,377
autohorse	-0,486	-0,320	-0,234
autompg	0,006	0,044	-0,044
balance	0,084	0,028	0,050
basketball	-0,172	-0,116	-0,124
bhouse	-0,209	-0,118	-0,216
bodyfat	-0,280	-0,143	-0,035
brainsize	0,024	0,155	0,248
breasttumor	-0,090	-0,106	-0,079
cholesterol	-0,052	-0,156	-0,126
cleveland	-0,210	-0,173	-0,205
cloud	-0,651	-0,300	-0,304
cos4	-0,183	-0,166	-0,157
cpu	-0,522	-0,559	-0,547
diabetes	-0,414	-0,233	-0,270
echomonths	0,178	0,096	0,114
elusage	-0,165	0,019	-0,305
fishcatch	-0,590	0,236	-0,328
fruitfly	0,115	0,080	0,043
grv	-0,055	0,164	-0,045
hungarian	0,010	-0,252	0,034
lowbwt	-0,118	-0,044	-0,178
mbagrade	-0,174	-0,083	0,005
meta	-0,150	-0,100	-0,146
pbc	-0,028	0,080	0,025
pharynx	0,004	0,071	0,012
photo	0,142	0,118	0,123
places	0,064	0,077	0,075
plasma_carotene	-0,010	-0,074	0,015
plasma_retinol	-0,048	-0,077	-0,067
pollution	-0,229	-0,070	-0,211
pwlinear	-0,108	-0,087	-0,217
pwlin2	-0,147	-0,084	-0,082
pyrim	-0,552	-0,563	-0,417
quake	-0,005	-0,015	-0,021
sensory	0,011	0,051	0,063
servo	0,110	0,130	0,071
sleep	0,260	0,140	0,080
stock	-0,208	0,215	-0,001
strikes	-0,031	-0,034	0,065
transplant	-0,589	-0,506	-0,534
triazines	-0,063	0,080	0,050
tumor	0,098	0,149	0,108
wind	-0,121	0,009	-0,107
wdbc	0,000	-0,079	0,033
+	5 (10%)	4 (8%)	2 (4%)
-	17 (35%)	14 (29%)	17 (35%)

References

- Bousquet O, Elisseeff A (2002) Stability and generalization. *J Mach Learn Res* 2:499–526
- Crowder MJ, Kimber AC, Smith RL, Sweeting TJ (1991) Statistical concepts in reliability. *Statistical analysis of reliability data*. Chapman & Hall, London, pp 1–11
- Bousquet O, Elisseeff A (2000) Algorithmic stability and generalization performance. In: *NIPS*, pp 196–202
- Bousquet O, Pontil M (2003) Leave-one-out error and stability of learning algorithms with applications. In: *Suykens JAK et al, Advances in learning theory: methods, models and applications*. IOS Press, Amsterdam
- Kearns MJ, Ron D (1997) Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In: *Computational learning theory*, pp 152–162
- Breiman L (1996) Bagging predictors. *Mach Learn* 24:123–140

7. Schapire RE (1999) A brief introduction to boosting. In: Proceedings of IJCAI, pp 1401–1406
8. Drucker H (1997) Improving regressors using boosting techniques. In: Machine learning: proceedings of the fourteenth international conference, pp 107–115
9. Ridgeway G, Madigan D, Richardson T (1999) Boosting methodology for regression problems. In: Proceedings of the artificial intelligence and statistics, pp 152–161
10. Breiman L (1997) Pasting bites together for prediction in large data sets and on-line. Department of Statistics technical report, University of California, Berkeley
11. Tibshirani R, Knight K (1999) The covariance inflation criterion for adaptive model selection. *J Roy Stat Soc Ser B* 61:529–546
12. Rosipal R, Girolami M, Trejo L (2000) On kernel principal component regression with covariance in action criterion for model selection. Technical report, University of Paisley
13. Elidan G, Ninio M, Friedman N, Shuurmans D (2002) Data perturbation for escaping local maxima in learning. In: Proceedings of AAAI/IAAI, pp 132–139
14. Gammerman A, Vovk V, Vapnik V (1998) Learning by transduction. In: Proceedings of the 14th conference on uncertainty in artificial intelligence, Madison, WI, pp 148–155
15. Saunders C, Gammerman A, Vovk V (1999) Transduction with confidence and credibility. In: Proceedings of IJCAI, vol 2, pp 722–726
16. Nourtdinov I, Melliush T, Vovk V (2001) Ridge regression confidence machine. In: Proceedings of the 18th international conference on machine learning. Kaufmann, San Francisco, pp 385–392
17. Vapnik V (1995) The nature of statistical learning theory. Springer, Berlin
18. Kukar M, Kononenko I (2002) Reliable classifications with machine learning. In: Proceedings of the machine learning: ECML-2002. Springer, Helsinki, pp 219–231
19. 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, Ljubljana, pp 99–103
20. Mitchell T (1999) The role of unlabelled data in supervised learning. In: Proceedings of the 6th international colloquium of cognitive science, San Sebastian, Spain
21. Blum A, Mitchell T (1998) Combining labeled and unlabeled data with co-training. In: Proceedings of the 11th annual conference on computational learning theory, pp 92–100
22. Li M, Vitányi P (1993) An introduction to Kolmogorov complexity and its applications. Springer, New York
23. Press WH et al (2002) Numerical recipes in C: the art of scientific computing. Cambridge University Press, Cambridge
24. Newman DJ, Hettich S, Blake CL, Merz CJ (1998) UCI repository of machine learning databases. Department of Information and Computer Sciences, University of California, Irvine
25. Department of Statistics at Carnegie Mellon University (2005) StatLib—data, software and news from the statistics community
26. Cestnik B, Bratko I (1991) On estimating probabilities in tree pruning. In: Proceedings of European working session on learning (EWSL-91), Porto, Portugal, pp 138–150
27. Chang C, Lin C (2001) LIBSVM: a library for support vector machines. <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
28. Alpaydin E (2004) Introduction to machine learning. MIT Press, Cambridge