

An Overview of Advances in Reliability Estimation of Individual Predictions in Machine Learning

Technical Report T-03/08

Zoran Bosnić and Igor Kononenko
University of Ljubljana
Faculty of Computer and Information Science
Tržaška 25, Ljubljana, Slovenia
zoran.bosnic@fri.uni-lj.si, igor.kononenko@fri.uni-lj.si

June 21, 2008

Abstract

In Machine Learning, estimation of the predictive accuracy for a given model is most commonly approached by analyzing the average accuracy of the model. In general, the predictive models do not provide accuracy estimates for their individual predictions. The reliability estimates of individual predictions require the analysis of various model and instance properties. In the paper we make an overview of the approaches for estimation of individual prediction reliability. We start by summarizing three research fields, that provided ideas and motivation for our work: (a) approaches to perturbing learning data, (b) the usage of unlabeled data in supervised learning, and (c) the sensitivity analysis. The main part of the paper presents two classes of reliability estimation approaches and summarizes the relevant terminology, which is often used in this and related research fields.

Keywords: predictions, reliability, prediction accuracy, data perturbation, unlabeled examples, supervised learning.

1 Introduction

In Machine Learning, various criteria can be used to assess the predictive model quality, such as interpretability and computational complexity. However, predictive accuracy is usually considered the most important criterion [39]. When using supervised learning, we aim to achieve the best possible accuracy for modeling learning data and for making predictions for new examples that were not included in the learning process [1, 39]. As a result of these efforts, one of the

mainstreams of the research proposed a variety of predictive models, each of them featuring different properties.

Another line of research focused on developing techniques for achieving better accuracy by stabilizing predictive bias and variance properties of existing predictive models. These approaches use predictive models as their parameters and are therefore model-independent. Their strategies for improvement of the accuracy usually consist of iterative re-building of predictive models and modifications of learning set or other learning parameters. Bagging [14] and boosting [25] are the examples of such approaches.

All these research fields positively contributed towards gaining more accurate predictive methods and formalizing their evaluation approaches. However, the averaged accuracy measures which are the most commonly used for the evaluation of the model accuracy, provide no local information about the expected error of individual prediction for a given unseen example. For example, frequently used estimates in supervised learning are the mean squared error (MSE) and the relative mean squared error (RMSE); in unsupervised learning, various cluster quality measures and indices are used, (e.g. cluster homogeneity, isolation distance, etc.) [23, 46]. This became a challenge for the third related research field, whose challenge was to estimate the reliability (prediction error, confidence, accuracy¹) of individual predictions. The application of individual prediction reliability [19] estimates is important for different fields of machine learning, such as:

- risk-sensitive decision systems, where acting upon predictions may have significant consequences (e.g. medical diagnosis stock market, navigation, control applications). In such areas, the appropriate *local* accuracy measures may provide additional 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. Therefore, reliability estimates for individual predictions need to be developed.
- statistical simulations of the real data, which are used to study the properties of the applied statistical methods. In the area of real data simulations, the model parameters are set to describe the true model parameters as best as possible [28]. By replicating data using the simulation model and comparing it to the actual data, one can evaluate the used statistical method and its fit [47]. However, in cases when it is expensive and cost-consuming (e.g. in medicine, using expensive equipment, etc.) or impossible (e.g. in regression) to gather all outputs of the true model to perform the comparison, or when only the accuracy of some specific simulated output values

¹the terminology is explained in following sections and in Appendix A

of interest needs to be evaluated, the local reliability estimates are of a greater benefit than the measures of the overall model’s performance.

In this paper we summarize the main contributions of the research in the area of evaluation of reliability of individual predictions in supervised learning. The paper is organized as follows. In Section 2 we describe the work in the fields, which provided motivating ideas for the research which is the main subject of the paper: (a) the approaches to perturbing learning data, (b) the usage of unlabeled data, and (c) the sensitivity analysis. We explain the motivating ideas stemming from these fields. Section 3 summarizes the general ideas and two classes of approaches in the field of estimating the individual prediction reliability. Section 4 provides the conclusions drawn from the presented work and Appendix A summarizes the relevant terminology in the field.

2 Motivations in the Field of Model Analysis

In the following we summarize the work in three related research fields, which form the motivation for the development of the model-independent approaches, which are summarized in Section 3.2. The fields which deal with perturbing data and the usage of unlabeled examples in supervised learning are generally concerned with the accuracy performance and evaluation of the whole predictive model. Both of these fields exploit the variations of the original learning set to improve the general model accuracy. Since some of these methods also focus on weighing and analyzing the role of included individual examples in the learning set variations, an inspiration arises to explore the usage of these ideas further. A possible way to apply these ideas is using the sensitivity analysis as a general framework, which is the third field, summarized in this section.

2.1 Perturbations of Learning and Test Examples

The group of approaches aiming at improving accuracy of predictive systems by perturbing learning data is general and model-independent. These approaches generate the perturbations of learning data either by creating a new learning set by selecting learning examples with replacement or by assigning weights to particular examples. The approaches which perturb the test examples perform such perturbing by modifying attribute vectors with an objective to estimate or improve accuracy for a particular test example. In the following we make an overview for each of these groups of approaches.

One of the most well-known methods, which generates multiple learning sets by sampling with replacement the original learning set, is *bagging* [14]. On each of generated versions of the learning set bagging builds a separate predictive model and uses it to calculate a prediction which represents a solution to a partial problem. The final solution is achieved by combining individual prediction into the aggregated one. Bagging has been shown to strongly reduce the variance while in most cases leaving the bias unchanged. Thus it is mostly effective in conjunction with decision and regression trees which exhibit high variance.

Focusing on a model as a discriminant or predictive function, bagging works as smoothing of the classification frontier or the regression function. Averaging individual predictions in an aggregate therefore gives a smoothed prediction which is more stable. Although the approach has proven to be effective, its downside is that the construction of the whole set of models is time and memory consuming.

Similar to bagging, other methods like *stacking* [68] and *bumping* [61] attempt to decrease the prediction bias besides reducing the variance. Researchers found that in all these cases a viable solution involved fitting several models and merging the predictions that each model produced [48]. One of the latest methods that involved model mixing, is *boosting* [25, 56]. Its authors developed an algorithm that sequentially fits weak classifiers to different weightings of the examples in a dataset. Those observations which the previous classifier poorly predicts receive greater weight in the next iteration. The final classifier is defined as a weighted average of all the weak classifiers. The final classifier merge has proven to be an effective method for reducing bias and variance, and for improving misclassification rates. Empirical evidence has shown that the base classifier can be fairly simplistic (shallow classification trees) and yet, when boosted, can capture complex decision boundaries. In the later work, the approach was adapted and evaluated also for regression problems [22, 49].

Tibshirani and Knight [60] 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 [50] 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. [24] 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.

The *dual perturb and combine* algorithm [29] is an approach which, in contrast to previously described approaches, perturbs test examples. In the first stage of the algorithm, a single prediction model, which remains unchanged through the whole procedure, is generated. In the prediction stage, attribute vector of a test example is perturbed several times by an additive random noise. The predictions calculated for each of these perturbed test examples are afterwards aggregated and averaged to obtain more stable prediction for the original test example. The experiments on several data sets with decision trees have shown that the method yields significant improvements in prediction accuracy, which are in some cases comparable to results obtained with bagging. But unlike with bagging, this approach makes use only of one model and delays the generation of multiple predictions until the prediction stage. In this way the

method preserves the interpretability and the computational efficiency of the original model. Later, this approach has been also applied with the artificial neural networks and in the context of learning from data streams [26].

The data perturbation approach has been also used in the unsupervised learning, to obtain clustering reliability estimates and assess the clustering stability [36]. The authors propose an application of the bootstrapping to generate a large number of bootstrap simulated clusterings. By analyzing the appearance frequency of a particular example in each of these simulated clusterings, the authors measure the reliability of the initial clustering. So defined reliability estimate may also hold a potential for correcting the clustering outcome for a given example (i.e. assigning the example to another cluster, to which the example belonged in the majority of the simulated clusterings).

Besides improving the predictive accuracy, the approaches with perturbing learning data were also used to solve the problems with time and memory limitations of storing too large data sets. *Pasting* [15] is such an approach that takes small bites of the data, grows a predictor on each small bite and then pastes these predictors together. The approach gives accuracy comparable to that which could be obtained if all data would be held in core memory which is computationally faster. The latter also enables the method to be applicable to on-line learning.

All mentioned approaches iteratively modify the learning set and have been justified in favorably improving the general hypothesis accuracy score. These results suggest that the inclusion or removal of individual learning example, while observing the final model accuracy, may be utilized as an indicator of model stability, related to that individual example. This idea is illustrated in Fig. 1 and was employed in later work in the field of estimating the reliability of individual predictions, which is summarized in Section 3.

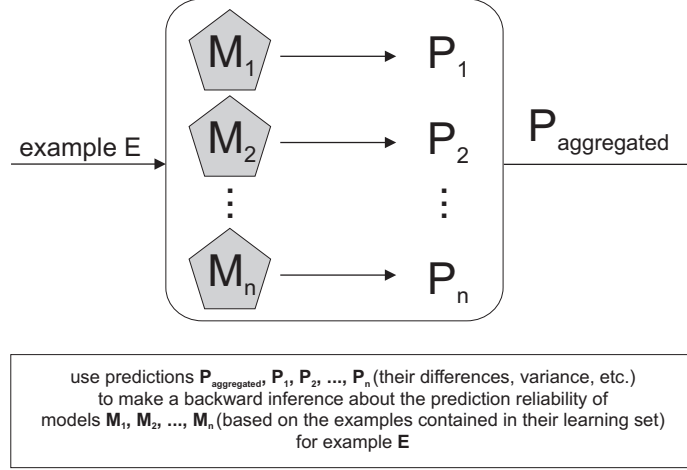
2.2 Usage of Unlabeled Data in Supervised Learning

The core idea behind the learning with unlabeled data is that the additional utilization of unlabeled examples together with the labeled learning examples can significantly improve the accuracy of the predictive model [58]. Since the true labels of unlabeled examples are not known, the employment of such examples does not directly contribute to the knowledge about the dependency between attributes and dependent (predicted) value. Instead, the employment of such examples in the learning process contributes the supplemental information about the true example distribution in the problem space, which facilitates more accurate learning process.

Prior to augmenting models by combining the unlabeled and labeled examples, the unlabeled examples must be assigned a value of the dependent variable. The well-known EM (Expectation - Maximization) algorithm [20, 30] provides a solution to this problem and can be summarized as follows:

1. Build the model using only the labeled data.

Figure 1: Reliability of prediction for example E and a particular model M_i could be quantitatively defined by analyzing the value of prediction P_i with respect to predictions, parameters (contents of the perturbed learning sets) of other models and the stable averaged prediction $P_{aggregated}$.



2. Use the model to estimate the probabilistic density of the possible labelings.
3. Use the model and the distribution information to probabilistically label the unlabeled examples.
4. Using the union of examples with the known labels, and examples with the probabilistically assigned labels, rebuild the model.
5. If a pre-specified termination condition is not met, go to step 2 and include additional data.

Applications of such approaches have shown to provide added value in environments, where large number of unlabeled examples is available, but it is impossible or too costly to label them and use them in the classical supervised learning scenario. Such examples include medical domains, where large amounts of undiagnosed patients' data is available, but no experts to systematically label them; or image recognition problems, where it may be very time consuming to process the graphical data. The augmentation of classifiers in the field of image classification [2] have shown that such approach benefits in greater classification accuracy compared to the traditional approach.

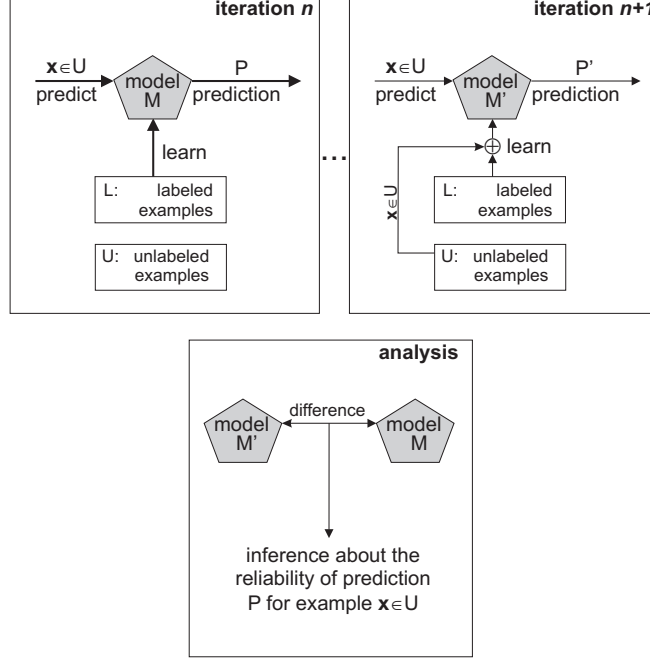
The following research in the same field more focused on the use of unlabeled examples in the context of co-training. Blum and Mitchell [4, 44] showed that the unlabeled examples can supplement the hypothesis, generated on labeled examples, when the problem can be represented into two distinct views. This

means that the description of each learning example can be partitioned in two parts, which are independent and mutually redundant. Their work is applied on the problem of web page classification, where each web page description can be partitioned into the words occurring at that page and the words occurring in hyperlinks that point to that page. Based on the assumption that each of the views is independently sufficient for solving the classification problem by itself, two separate learning sets are formed consisting of the same learning examples, but described with different attribute sets. By building classifiers on each of these two problems, one can use a classifier which was built on one view to classify unlabeled data, presented using the other view. The resulting newly labeled examples can be afterwards included into the original learning set. By using this approach, the learning results have shown promising reduction of classification error.

Similar approaches are used also in application field, other than classification of web pages. De Sa [21] addresses the problem of learning from unlabeled data without experience with previously labeled examples at all. The author applies her approach in the field of computer data classification, in which assigning classes to available examples can be time or expense demanding. As a solution, the author proposes a form of *self-supervised* learning, which partitions the problem into two independent problems in a similar way as in [4]. In the area of processing computer data, such problem representation can be made by separating the image contained in the record from its associated audio signal. The author proposes training of two neural networks on each of the separated data sets and joining the outputs from both networks with the common output layer of neurons. The goal of the proposed self-supervised learning is to minimize the differences between outputs of both networks, making both networks classify two views of the same examples into same two implicit classes. The implicit classes are defined by the codebook vectors, which are at the beginning randomly initialized in the problem space. Each of the examples belongs to the class which is represented by the closest codebook vector. By learning such joint neural network by backpropagating the error of disagreement between two individual networks, the codebook vectors converge to locations which represent new centers of the implicit classes. The results showed that the proposed method does not perform as well as the standard supervised learning approaches, but nevertheless it successfully and innovatively solves the problem of learning from completely unlabeled data.

The requirement that the set of attributes must be separable into two different partitions is overcome in the work of Goldman [32]. The only requirement of the author's co-training strategy is that the used supervised learning algorithm partitions the example space into a set of equivalence classes (e.g. for a decision tree each leaf defines an equivalence class). After building two predictive models using different supervised learning algorithms, the unlabeled examples are added into corresponding equivalence classes of each of the models. Selection of the appropriate equivalence class for a particular unlabeled example and combining of both generated hypotheses is performed by evaluation of confidence intervals. Addition of new examples and building of predictive models is

Figure 2: Reliability of prediction for example E and a particular model M could be quantitatively defined by analyzing the change in model when an unlabeled example E is added to the model’s learning set, yielding model M' .



afterwards repeated in many iterations. The results showed that the described approach favorably influences predictive accuracy, i.e. significantly reduces the prediction error of both predictive models.

The applicable results with employing unlabeled data in the supervised learning indicate that the additional learning examples, which are generated from the same original probabilistic distribution, can be beneficially utilized for improving predictor accuracy. As seen from the variety of approaches, developed in this field, the procedures for including additional learning examples and building new predictive models are performed in turn and in many iterations. This allows the predictive model to gradually increase its accuracy and more accurately label unlabeled examples which are yet to be included in later iterations. Since by each included example the model changes, this conclusion encourages the idea to observe the influence of inclusion of a particular new example in the learning set. Namely, by observing the consequential change in the predictive model, one could try to make inference about the model stability, related to that individual example. This idea is illustrated in Fig. 2 and presents a motivation for research in the field of estimating reliability of the individual predictions, on which we focus later in Section 3.

2.3 Sensitivity Analysis

Methods for measuring the overall accuracy of a particular predictive model or of its individual predictions are founded on the quantitative description of predictive model properties or on the characteristics of the input space. Noise in data and nonuniform distribution of examples represent a challenge for learning algorithms, leading to different prediction accuracies in different parts of the problem space. Apart from distribution of learning examples there are also other causes that influence the accuracy of prediction models: their generalization ability, bias, resistance to noise, avoidance of overfitting, etc. Since these aspects cannot be measured quantitatively, they cannot be used to construct a quantitative measure for evaluation of the accuracy.

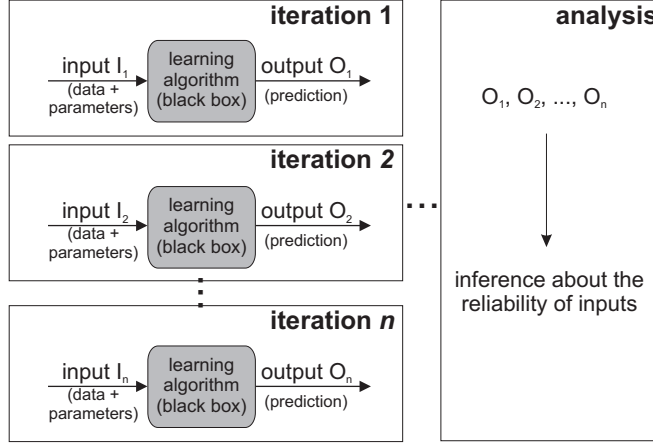
A possible approach to analyze the local particularities in data and predictive model properties is with use of the *sensitivity analysis*. Sensitivity analysis is an approach which is used to study the influence of parameters and model properties to its structure and outputs [13]. It is usually performed as a series of experiments in which the user systematically changes the input parameters and observes the dynamics of changes in outputs. For the usage of this technique, no knowledge of model's mathematical properties is required, hence the model is basically used as a parameter of the method, presenting a black box with inputs and outputs [38], which are the only parameters of interest. This approach has been most widely used in the area of statistics, mathematical programming [52] and natural sciences [51]. The usage of sensitivity analysis with artificial neural networks [33] and with Bayesian networks [37] has shown the potential for application of this approach with the supervised learning algorithms.

In the context of theoretical stability analysis of learning algorithms, the sensitivity analysis has been discussed by Bousquet and Elisseeff [11]. 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 [10] and Elisseeff and Pontil [12] applied these ideas to several learning models and showed how to obtain bounds on their generalization performance.

In a similar way Kearns and Ron [35] 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 [63] or from the way the learning algorithm searches the space.

By proving theoretical usefulness of notion of *stability*, these approaches motivated research about empirical estimation of the individual prediction reliability based on the local stability of the model, as illustrated in Fig. 3. For usage in machine learning, a framework has been proposed [5] for controlled changing of the input (i.e. the learning set) of the learning algorithm, and observing the changes in output (i.e. predictions) of the learning algorithm. This framework defines a systematical approach to modifying the learning set, which

Figure 3: The sensitivity analysis approach considers the predictive algorithm as a black box. By observing the change in system outputs with respect to the controlled change in inputs (examples and model parameters), an inference could be made about the reliability of individual inputs.



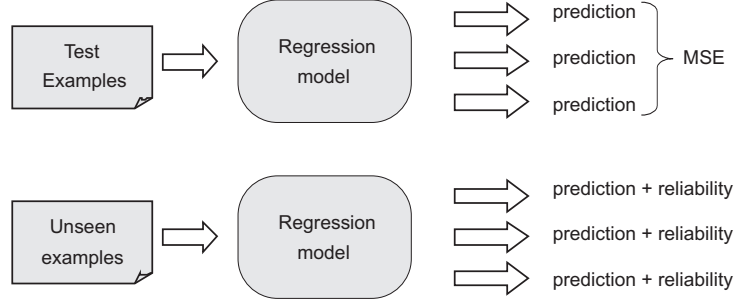
induces a change in the input of the learning algorithm. If this change is small, then the change in output prediction for the modified example is also expected to be small. Since the opposite scenario would indicate the instability in the generated model, the magnitude of output change may therefore be used as a measure of model instability for a modified example. The same reasoning is used also in other model-independent approaches in the field of individual prediction reliability estimation.

3 Reliability Estimation for Individual Examples

Individual reliability estimates enable the user to make a distinction between more and less accurate predictions. The idea and benefits of implementing this challenge in contrast to estimating the average model accuracy (e.g. using MSE) is illustrated in Fig. 4. The reliability estimation of individual examples has also an additional advantage as well. Namely, the calculation of individual predictions' reliability estimates does not require knowing of true label values. In contrast to MSE estimate, which requires a testing data set, the idea of individual predictions' estimates is that they can be calculated for arbitrary unseen examples.

Past research has referred to reliability estimation of individual predictions with different terms and in the different contexts of machine learning. This challenge was most frequently approached in the field of classification. The notion

Figure 4: Reliability estimate for the whole regression model (above) in contrast to reliability estimates for individual predictions (below).



of reliability estimation has most frequently appeared in conjunction with the notion of *transduction* or *transductive reasoning*. The usage of these terms may have different meanings in different environments, but basically *transduction* represents an inference principle that reasons from particular to particular [63] in contrast to inductive learning, which aims at inferring a general rule from particular data. Transductive inference therefore aims at making predictions for unlabeled examples without constructing a general predictive model. Such definition closely relates the transductive reasoning to instance-based learning and case-based reasoning (one of the most well known algorithms in this area is k nearest neighbors). However, the transductive methods represent much wider scale of approaches, since the reasoning can be based also on other criteria (e.g. distribution of examples in input space) and not only using the distance metrics and labels of the nearest neighbors. Transductive methods may also exploit only selected examples of interest and not necessarily the whole input space, which enables them to make other inferences apart from predicting labels. We find inferences about reliability measures of a special interest.

A good criteria to differentiate between various approaches for reliability estimation for individual examples is by determining whether they target a specific predictive model or whether they are model-independent. While the approaches from the first group are less general, they are usually better mathematically or probabilistically founded. 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 focus on influencing the standard parameters, which are available in the supervised learning framework (e.g. the learning set, attributes, etc.). Such approaches are therefore less probabilistically interpretable, allowing such reliability estimates can take values from an arbitrary interval of numbers.

To refer to both groups of estimates, probabilistically interpretable and general, we use a superordinate term *reliability estimate* to name any measure that provides information about trust in individual prediction accuracy. Since the

true error of an unlabeled example is not known, it is also more appropriate to say that we estimate the *prediction reliability* rather than the *prediction error*. This also conforms with the definition of *reliability* which can be defined as *the ability to perform certain tasks conforming to required quality standards* [19]. Namely, the prediction accuracy in regression is considered the required quality standard.

In the following we focus on each group of methods individually.

3.1 Extensions of Existing Predictive Models

To enable the users of classification and regression models to gain more insight into the reliability of individual predictions, various extensions of existing classification and regression models were proposed, enabling them to output predictions, extended with their corresponding reliability estimates.

By using mathematical and statistical foundations to expand the basic formulation of a predictive model, it is feasible to define reliability estimates which are grounded in the probabilistic theory, hence having values with the probabilistic interpretation. Such reliability estimates, called *confidence measures*, have values belonging to the interval $[0, 1]$, where 0 represents the confidence of the most unreliable prediction and 1 the confidence of the most reliable one.

Previous studies have referred to reliability of single predictions with different terms. Gammerman, Vovk and Vapnik [27] and later by Saunders, Gammerman and Vovk [53] propose an extension of the support vector machine (SVM) algorithm for classification and show that their modified SVM successfully produces the reliability estimates and outperforms other predictive algorithms. Besides achieving favorable results, the authors also introduced the notions of *confidence* and *credibility*, denoting a probabilistic reliability estimate and probability for not classifying an example into the second most probable class, respectively. This approach was later applied in selected practical domains, including face recognition problem [41], where confidence and credibility proved to be informative measures of classification reliability. Later on, the work continued with Nourtdinov, et al. [45] demonstrating the use of *confidence* value in the context of ridge regression. Using residuals of learning examples and a *p*-value function the authors improved the basic ridge regression with confidence regions.

Focusing on multilayer perceptrons, Weigend and Nix [64] extended their predictions with adjoined reliability estimates by expanding the original perceptron with an additional output neuron, that was intended to predict the variance in the neighborhood of the input example. The variance was learnt as a part of the backpropagation learning, during which the variances for the learning examples were calculated and presented as the output targets. Despite the big sensitiveness to the local changes, the favorable experimental results showed that the predicted variance estimates can be used as the reliability estimates.

In 1997 Heskes [34] proposed a method for computation of *prediction intervals* for the ensembles of neural networks. Being defined as a degree of agreement between predicted value and example's label value, the prediction interval is therefore an estimate of the individual prediction reliability as well.

The proposed reliability estimates are defined using the variance of outputs of individual neural networks in the ensemble, where each of them was trained and stopped on the bootstrap replicates of the original data set. In 1999, Carney and Cunningham [16] proposed an improvement of the former method by dividing the original ensemble to smaller, equally sized ensembles and averaging the variances among them. The experimental results showed an increase in stability and accuracy of such improved reliability estimates.

The overview shows that the former group of approaches is designed for the use with particular predictive models and due to their specific formalisms they cannot be used with other models. The other branch of the research, on which we focus in the following subsection, therefore explores the approaches which are independent of the predictive model, hence being more general.

3.2 Model-Independent Approaches

Many approaches to model-independent reliability estimation are concerned with local modeling of prediction error based on input space properties and local learning [31, 55, 54, 69, 3]. In this context most frequently the local cross validation is applied to calculate the prediction and the prediction error for the example of interest using a local model in the problem subspace. Using the local leave-one-out procedure, the local errors of the example’s neighbors are acquired, enabling the reliability estimate of the example to be defined as their weighted average. These approaches have indicated good results, but are sensitive to distance metric, noise in the neighborhood and the type of the local predictive model.

As an alternative to local modeling, Tsuda et al. [62] proposed an algorithm to predict the leave-one-out error of a single example for kernel based classifiers (support vector machines and linear programming machines). Their work introduces a meta-level of reasoning and presents a meta-algorithm for predicting the leave-one-out error. The algorithm is based on observation, whether the omission of a particular learning example from the learning set would cause the example to be misclassified in the prediction stage. The simulations showed that the proposed meta-learning approach compares favorably to the conventional theoretical leave-one-out error bounds, owing to which its leave-one-out error estimate can be reliably used for model selection.

As an application of the transductive reasoning principle, Kukar and Kononenko [40] 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. The proposed estimates are based on the change of the posterior class distribution between the *inductive* and *transductive* classifiers. Inductive classifier is the one generated on the original learning set of examples, while the learning set of the transductive classifier includes an additional example, the one for which the classification reliability is being estimated. The usage of term *transductive* in this context is justified by purpose of the transductive model, whose task is to make inference about classification reliability for a given

example of interest and not to perform classification. To quantitatively express the difference between posterior class distributions, the authors propose a set of metrics (variation distance, Bhattacharyy’s distance, harmonic mean, standardized Euclidean distance, cosine of angle between the vectors representing distributions, etc.). These metrics and the products of their various combinations are used to separate sets of correctly and incorrectly classified examples by defining a threshold for each of the used metrics. Since the metrics are not bound to a particular model formalization, the approach can be used with the arbitrary learning algorithm. The approach as such therefore represents a first joint implementation of ideas, coming from the field of perturbing data and the field of using unlabeled examples in supervised learning, as illustrated in Fig. 1 and 2. The results have shown successful separation of correctly and incorrectly classified examples in many testing domains.

Bosnić and Kononenko [9, 6] 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 the prediction error.

The extended work of Bosnić and Kononenko [5] standardized the framework for the usage of sensitivity analysis approach to develop the individual predictions’ reliability measures, evaluated the appropriateness of this technique for five regression models, and justified the motivation for this work using the Minimum Description Length principle [42]. Using the previous work in the same field, they supplemented the methodology with ideas from the field of sensitivity analysis, as illustrated in Fig. 3. Given the example and its *initial prediction* for which the reliability was to be estimated, the authors repeatedly modified the model learning set to obtain a set of *sensitivity predictions*. Based on sensitivity predictions they proposed the reliability estimates which evaluated the *local* bias and *local* variance in the problem subspace, leading to information about prediction reliability. The purpose of the controlled local modification of the learning set was to explore the sensitivity of the regression model in a particular part of the problem space. By doing so, the reliability estimates were adapted to the local particularities of data distribution and noise and the sensitivity was thus related to changes of the regression model prediction when the learning set was slightly changed.

The similar ideas are put in practice also in a related field of active learning [17] which is concerned with the optimal choice of which learning examples to include in the learning set to optimize learning. There are many heuristics for choosing the next learning example, based on choosing places where less data is available [65], where the model performs poorly [43], where the confidence in predictions is low [59], where the new example will influence the change in the model most [18] or where we previously found data that resulted in learning [57] (reinforcement learning strategies).

In contrast to the preceding approach for the estimation of classification reliabilities, the reliability estimates for regression predictions were based solely on the outputs of the prediction system and therefore did not require any estimations of distribution functions. The use of such approach was possible due

to the continuous nature of predicted values in regression. Namely, this makes possible to numerically express the difference between two regression predictions, in contrast to classification, where it can only be observed whether the predicted class was the same or different. An open question for the further work remains, whether the sensitivity predictions could also be used to correct the initial predictions of the examples, reducing their error and making them more accurate.

3.3 Performance Comparison of Reliability Estimates

When evaluating reliability estimate’s performance, we are interested mostly in how informative it is about the prediction accuracy or its error. However, with model-dependent approaches which are formalized as model extensions, the estimates are commonly expressed as confidence intervals, providing information that a prediction belongs to an interval with a certain degree of probability. Due to probabilistic founding of these approaches, the information they provide is accurate and does not call for evaluation in contrast to the model-independent reliability estimates.

The model-independent estimates are commonly defined as metrics with an arbitrary interval of target values, which are approach and domain dependent. As such they do not provide an absolute interpretation about the prediction accuracy/error, but allow to relatively determine which predictions are *more* and which are *less* reliable. Besides depending on the approach and the problem domain, the accuracy of model-dependent estimates varies also depending on the used regression model. Comparison of the described key properties of both approach families is given in Table 1.

Table 1: Comparison of the key performance properties of model-dependent and model-independent reliability estimates.

	model-dependent approaches (extensions of existing models)	model-independent approaches (general approaches)
founding	embedded into the model	treating model as a parameter (black box)
estimates’ values	estimates usually probabilistically interpretable	estimate values belong to an arbitrary interval
example of reliability value	prediction for example x_1 is on interval $[14,16]$ with 95% probability	reliability for (\mathbf{x}_1, y_1) is 49 AND reliability for (\mathbf{x}_2, y_2) is 59 \Rightarrow prediction for (\mathbf{x}_2, y_2) is more reliable

In previous work, extensive testing of nine reliability estimates was performed [8], which were based on the various model-independent approaches: sensitivity analysis, measuring variance of bagged predictors, local cross-validation,

density-based estimation, local modeling of the error, and a combination of these approaches. Testing of these estimates was performed in terms of statistical evaluation of their correlation to the prediction error. The obtained results using eight regression models (regression trees, linear regression, neural networks, bagging with the regression trees, support vector machines, locally weighted regression, random forests, and generalized additive model) indicated different performance of the estimates with different regression models. They especially indicated the potential for the usage of:

- sensitivity analysis estimates and local modeling estimates with the regression trees, linear regression, and generalized additive model,
- bagging variance with locally weighted regression,
- local cross-validation estimate with support vector machines, random forests and locally weighted regression,
- combined estimate with the neural networks and bagging with regression trees.

Since the empirical evaluations revealed that the estimates' performance depends on the used regression model and on the particular problem domain, two approaches for the automatic selection of the best performing reliability estimate were proposed: based on the meta-learning and on the internal cross-validation [7]. The testing results of both approaches demonstrated an advantage in performance of dynamically chosen reliability estimates over performance of the individual reliability estimates. In addition, the preliminary testing of the proposed methodology on a medical domain demonstrated the potential for its usage in practice.

4 Conclusion

The paper summarizes approaches which provide motivating ideas for the development of methods for estimation of individual prediction reliability and the field of estimating reliability of individual predictions itself. We have explained the motivation coming from the following research fields:

- **approaches to perturbing data,**
- **usage of unlabeled examples in supervised learning,**
- **sensitivity analysis,**
- **other related research fields:** active learning, transductive reasoning, meta-learning, and reinforcement learning.

Motivation coming from all the above fields indicates the significance of a single example for the evaluation of overall model's performance. As such, it also provides the ideas for evaluation of a single prediction reliability instead of

evaluating the whole model. The above approaches imply that a single prediction reliability can be evaluated by observing changes in the generated model, which occur when a particular example is added or removed from the learning set. In addition, the sensitivity analysis approach offers the general framework which can be used to systematically analyze changes in the generated model, while remaining independent of a particular model.

In the section about the reliability estimation of individual examples we made a differentiation between various approaches by determining whether they target a specific predictive model or whether they are model-independent. In the Appendix we also present a unified excerpt of the related terminology, which often appears in the literature, carrying different meanings.

As the field individual prediction reliability estimation is relatively new, the following questions arise and call to be addressed in the further work:

1. How to define further and more accurate model-independent reliability estimates?
2. How to make model-independent reliability estimates' values more interpretable (probabilistically?) and comparable among different estimates?
3. How to select the most appropriate reliability estimate for a given problem?
4. What are the possibilities for the application of the described methodology (stock market, medicine, data streams, etc.)?

The latter questions need to be addressed in any practical application of this methodology, as were in the application on a medical prognostics domain [7], which also demonstrates the benefits and potentials of this methodology. 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 goal of the research was to predict the time of possible cancer recurrence after the surgical treatment. The study resulted in complementing the bare recurrence predictions with reliability estimates, helping the doctors with the additional validation of the predictions' accuracies and significantly improving usefulness of the prognostic system.

To conclude, further development of the field of prediction reliability estimation would bring benefit for users of critical decision support systems, where the prediction accuracy implies financial, business, medical and other important consequences. Bringing awareness to the machine learning community about the potentials of this methodology may bring its greater utility as well as the further advances in this field.

A Glossary of Relevant Terms

In the following we present a résumé of the terminology used in the field. The asterisk (*) marks the terms which we define in accordance with the definitions of closely related other terms in the area.

accuracy estimate* One of the aspects of prediction reliability. An estimate which positively correlates with the prediction accuracy or negatively correlates with the prediction error. Estimate, similar to confidence, but more general, since it does not have a probabilistic interpretation (it can take values from an arbitrary interval of real numbers and need not be limited to $[0, 1]$).

confidence Probabilistically expressed accuracy estimate for a given prediction. Value of prediction confidence therefore represents the probability of its accurateness. It is based on an assumed probability distribution and in classification it can be also defined as $1 - p_2$, where p_2 denotes the probability of the second most probable class [53].

confidence interval In statistics, an interval on which an estimate is likely to take values with a given degree of confidence [66]. In regression problems, also the prediction accuracy with respect to the true example value (which may differ from its label due to noise in the data) [34].

credibility A probabilistic measure for estimation of prediction reliability, related to confidence. Instead of estimating accuracy based on assumed probability distribution, it uses its own probabilistic model (e.g. observing the ratio of support vectors among all examples or using the Lagrange coefficients [53]) to perform the reliability estimation.

error estimate* One of the aspects of prediction reliability. An estimate which positively correlates to the prediction error. It does not have a probabilistic interpretation and can therefore take values from an arbitrary interval of real numbers. It may be implemented as inverted accuracy estimate.

prediction interval The output prediction accuracy with respect to the target value (which differs from the true regression value by some noise) [34].

probabilistic error estimate* An error estimate with a probabilistic interpretation, expressed as $1 - \textit{confidence}$.

reliability A general notion in engineering, denoting the ability of a system or a component to perform its required functions under stated conditions for a specified period of time [67]. In machine learning, we can define reliability as any qualitative property or ability of the system which is related to a critical performance indicator (positive or negative) of that system, such as accuracy, inaccuracy, availability, downtime rate, responsiveness, etc.

reliability estimate An estimate for quantitative measuring of reliability (which is in most cases defined qualitatively). According to particular context (see def. of *reliability*), reliability estimate can therefore represent an accuracy estimate, error estimate, availability estimate, etc.

sensitivity A notion similar to stability. Quantitatively expressed dependence between the changes in system parameters and structure, and the critical aspects of the system operation [13].

stability A property of the system not to change its critical performance aspect for more than for a specified threshold, when selected parameters of the system change. Defined in the context of learning hypothesis stability [11].

transduction A wide term, generally denoting *reasoning from particular to particular*. In the context of reliability estimation it represents the basis for many approaches [27, 53, 41]. The transductive reasoning is often used to construct reliability estimates, which measure the probability of how the newly labeled example fits into the distribution of all given examples. The notion of transduction is used also in cases, when the construction of the hypothesis is needed only for examples of interest or for some other intention apart from predicting. Such application may be the estimation of prediction reliability, as in [40, 9].

References

- [1] E. Alpaydin, *Introduction to Machine Learning*, The MIT Press, Cambridge, Massachusetts, 2004.
- [2] S. Baluja, Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data, in: *Proceedings of the 1998 conference on Advances in neural information processing systems II*, M.J. Kearns et al., eds., The MIT Press, 1998, pp. 854–860.
- [3] M. Birattari, H. Bontempi and H. Bersini, Local Learning for Data Analysis, in: *Proceedings of the 8th Belgian-Dutch Conference on Machine Learning*, 1998, pp. 55–61.
- [4] A. Blum and 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.
- [5] Z. Bosnić and I. Kononenko, Estimation of individual prediction reliability using the local sensitivity analysis, *Applied intelligence [Online edition]*, <http://www.springerlink.com/content/e27p2584387532g8/> (2007), 1–17.
- [6] Z. Bosnić and I. Kononenko, Estimation of regressor reliability, *Journal of intelligent systems* **17(1/3)** (2008), 297–311.
- [7] Z. Bosnić and I. Kononenko, *Automatic Selection of Reliability Estimates for Individual Regression Predictions Using Meta-Learning and Internal Cross-Validation*, Technical Report T-02/08 (submitted), available at <http://lkm.fri.uni-lj.si/zoranb/>, University of Ljubljana, 2008.

- [8] Z. Bosnić and I. Kononenko, *Towards Reliable Reliability Estimates for Individual Regression Predictions*, Technical Report T-01/08 (submitted), available at <http://lkm.fri.uni-lj.si/zoranb/>, University of Ljubljana, 2008.
- [9] Z. Bosnić, I. Kononenko, M. Robnik-Šikonja and M. Kukar, Evaluation of prediction reliability in regression using the transduction principle, in: *Proceedings of Eurocon 2003*, B. Zajc and M. Tkalčič, eds., 2003, pp. 99–103.
- [10] O. Bousquet and A. Elisseeff, Algorithmic Stability and Generalization Performance, in: *Neural Information Processing Systems*, 2001, pp. 196–202.
- [11] O. Bousquet and A. Elisseeff, Stability and generalization, *Journal of Machine Learning Research* **2** (2002), 499–526.
- [12] O. Bousquet and M. Pontil, Leave-one-out error and stability of learning algorithms with applications, in: *Advances in Learning Theory: Methods, Models and Applications*, J.A.K. Suykens et al., eds., IOS Press, 2003.
- [13] L. Breierova and M. Choudhari, *An Introduction to Sensitivity Analysis*, prepared for the MIT System Dynamics in Education Project, MIT, 1996.
- [14] L. Breiman, Bagging Predictors, *Machine Learning* **24** (1996), 123–140.
- [15] L. Breiman, *Pasting bites together for prediction in large data sets and on-line*, Technical Report, University of California, 1997.
- [16] J. Carney and P. Cunningham, Confidence and prediction intervals for neural network ensembles, in: *Proceedings of The International Joint Conference on Neural Networks*, Washington, USA, 1999, pp. 1215–1218.
- [17] D.A. Cohn, Z. Ghahramani and M. I. Jordan, Active Learning with Statistical Models, *Advances in Neural Information Processing Systems* **7** (1995), 705–712.
- [18] D.A. Cohn, L. Atlas and R. Ladner, Training Connectionist Networks with Queries and Selective Sampling, *Advances in Neural Information Processing Systems* **2** (1990), 566–573.
- [19] M.J. Crowder, A.C. Kimber, R.L. Smith and T. J. Sweeting, Statistical concepts in reliability, in: *Statistical Analysis of Reliability Data*, Chapman & Hall, London, UK, 1991, pp. 1–11.
- [20] A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* **39**(1) (1977), 1–38.

- [21] V. de Sa, Learning Classification with Unlabeled Data, in: *Proceedings of Neural Information Processing Systems*, J.D. Cowan et al., eds., Morgan Kaufmann Publishers, San Francisco, CA, 1993, pp. 112–119.
- [22] H. Drucker, Improving regressors using boosting techniques, in: *Machine Learning: Proceedings of the Fourteenth International Conference*, 1997, pp. 107–115.
- [23] A. Dudek, Cluster Quality Indexes for Symbolic Classification An Examination, in: *Advances in Data Analysis*, Springer Berlin Heidelberg, 2007, pp. 31–38.
- [24] G. Elidan, M. Ninio, N. Friedman and D. Shuurmans, Data Perturbation for Escaping Local Maxima in Learning, in: *Proceedings AAAI/IAAI*, 2002, pp. 132–139.
- [25] Y. Freund and 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.
- [26] J. Gama and P.P. Rodrigues, Stream-Based Electricity Load Forecast, in: *Proceedings of PKDD 2007, Lecture Notes in Artificial Intelligence* **4702**, J.N. Kok et al., eds., Warsaw, Poland, 2007, pp. 446–453.
- [27] A. Gammerman, V. Vovk and V. Vapnik, Learning by Transduction, in: *Proceedings of the 14 th Conference on Uncertainty in Artificial Intelligence*, Madison, Wisconsin, 1998, pp. 148–155.
- [28] A. Gelman and J. Hill, *Data Analysis Using Regression and Multi-level/hierarchical Models*, Cambridge University Press, 2006.
- [29] P. Geurts, Dual perturb and combine algorithm, in: *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*, 2001, pp. 196–201.
- [30] Z. Ghahramani and M.I. Jordan, Supervised learning from incomplete data via an EM approach, *Advances in Neural Information Processing Systems* **6** (1994), pp. 120–127.
- [31] G. Giacinto and F. Roli, Dynamic classifier selection based on multiple classifier behaviour, *Pattern Recognition* **34**(9) (2001), 1879–1881.
- [32] S. Goldman and Y. Zhou, Enhancing Supervised Learning with Unlabeled Data, in: *Proceedings of 17th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA, 2000, pp. 327–334.
- [33] H. Hashem, Sensitivity Analysis for Feedforward Artificial Neural Networks with Differentiable Activation Functions, in: *Proceedings of 1992 International Joint Conference on Neural Networks IJCNN92* **I**, 1992, pp. 419–424.

- [34] T. Heskes, Practical Confidence and Prediction Intervals, *Advances in Neural Information Processing Systems* **9** (1997), 176–182.
- [35] M.J. Kearns and D. Ron, Algorithmic Stability and Sanity-Check Bounds for Leave-one-Out Cross-Validation, in: *Computational Learning Theory*, 1997, pp. 152–162.
- [36] M. Kerr and G. Churchill, Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments, *Proc Natl Acad Sci USA* **96** (2001), 8961–8965.
- [37] U. Kjaerulff and L.C. van der Gaag, *Making sensitivity analysis computationally efficient*, submitted to UAI 2000, 2000.
- [38] J. Kleijnen, *Experimental Designs for Sensitivity Analysis of Simulation Models*, tutorial at the Eurosim 2001 Conference, 2001.
- [39] I. Kononenko and M. Kukar, *Machine Learning and Data Mining: Introduction to Principles and Algorithms*, Horwood Publishing Limited, UK, 2007.
- [40] M. Kukar and I. Kononenko, Reliable Classifications with Machine Learning, in: *Proceedings of Machine Learning: ECML-2002*, Springer Verlag, Helsinki, Finland, 2002, pp. 219–231.
- [41] F. Li and H. Wechsler, Open Set Face Recognition Using Transduction, *IEEE transactions on pattern analysis and machine intelligence* **27(11)** (2005), pp. 1686–1697.
- [42] M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications*, Springer-Verlag, New York, 1993.
- [43] A. Linden and F. Weber, Implementing inner drive by competence reflection, in: *Proceedings of the 2nd International Conference on Simulation of Adaptive Behavior*, Hawaii, 1992, pp. 321–326.
- [44] T. Mitchell, The role of unlabelled data in supervised learning, in: *Proceedings of the 6th International Colloquium of Cognitive Science*, San Sebastian, Spain, 1999.
- [45] I. Nouretdinov, T. Melliush and V. Vovk, Ridge Regression Confidence Machine, in: *Proceedings of the 18th International Conf. on Machine Learning*, Morgan Kaufmann, San Francisco, CA, 2001, pp. 385–392.
- [46] K. Osei-Bryson, Assessing Cluster Quality Using Multiple Measures - A Decision Tree Based Approach, in: *The Next Wave in Computing, Optimization, and Decision Technologies*, Springer US, 2005, pp. 371–384.

- [47] K. Patan and T. Parisini, Stochastic learning methods- for dynamic neural networks: Simulated and real-data comparisons, in: *Proceedings of the American Control Conference*, Inst. of Control & Comput. Eng., Poland, 2002, pp. 2577–2582.
- [48] G. Ridgeway, *The State of Boosting*, Technical Report, University of Washington Seattle, 1998.
- [49] G. Ridgeway, D. Madigan and T. Richardson, Boosting methodology for regression problems, in: *Proc. Artificial Intelligence and Statistics*, 1999, pp. 152–161.
- [50] R. Rosipal, M. Girolami and L. Trejo, *On Kernel Principal Component Regression with Covariance Inflation Criterion for Model Selection*, Technical Report, University of Paisley, 2000.
- [51] A. Saltelli, M. Ratto, S. Tarantola and F. Campolongo, Sensitivity Analysis for Chemical Models, *Chemical Reviews* **105**(7) (2005), 2811–2828.
- [52] A. Saltelli, S. Tarantola, F. Campolongo and M. Ratto, *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*, John Wiley & Sons Ltd, England, 2003.
- [53] C. Saunders, A. Gammerman and V. Vovk, Transduction with Confidence and Credibility, in: *Proceedings of IJCAI* **2**, 1999, pp. 722–726.
- [54] S. Schaal and C.G. Atkeson, Assessing the Quality of Learned Local Models, *Advances in Neural Information Processing Systems* **6**, (1994), 160–167.
- [55] S. Schaal and C.G. Atkeson, Constructive Incremental Learning from Only Local Information, *Neural Computation* **10**(8) (1998), 2047–2084.
- [56] R.E. Schapire, A Brief Introduction to Boosting, in: *Proceedings of IJCAI*, 1999, pp. 1401–1406.
- [57] J. Schmidhuber and J. Storck, *Reinforcement driven information acquisition in nondeterministic environments*, Technical Report, Technische Universität München, 1993.
- [58] M. Seeger, *Learning with labeled and unlabeled data*, Technical Report, <http://www.dai.ed.ac.uk/~seeger/papers.html>, 2000.
- [59] S.B. Thrun and K. Möller, Active Exploration in Dynamic Environments, *Advances in Neural Information Processing Systems* **4** (1992), 531–538.
- [60] R. Tibshirani and K. Knight, The covariance inflation criterion for adaptive model selection, *Journal of the Royal Statistical Society B* **61** (1999), 529–546.

- [61] R. Tibshirani and K. Knight, Model Search and Inference by Bootstrap Bumping, *Journal of Computational and Graphical Statistics* **8** (1999), 671–686.
- [62] K. Tsuda, G. Rätsch, S. Mika and K. Müller, Learning to Predict the Leave-One-Out Error of Kernel Based Classifiers, in: *Lecture Notes in Computer Science* **2130**, 2001, pp. 331.
- [63] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Verlag, 1995.
- [64] A. Weigend and 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.
- [65] S.D. Whitehead, A Complexity Analysis of Cooperative Mechanisms in Reinforcement Learning, in: *Proceedings of AAAI*, 1991, pp. 607–613.
- [66] Wikipedia, the free encyclopedia, *Confidence Interval*, http://en.wikipedia.org/wiki/Confidence_%5Cinterval, august 2007.
- [67] Wikipedia, the free encyclopedia, *Reliability*, <http://en.wikipedia.org/wiki/Reliability>, august 2007.
- [68] D.H. Wolpert, Stacked Generalization, *Neural Networks* **5** (1992), 241–259.
- [69] K. Woods, W.P. Kegelmeyer and K. Bowyer, Combination of Multiple Classifiers Using Local Accuracy Estimates, *IEEE Transactions on PAMI* **19**(4) (1997), 405–410.