Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

A priori reliability prediction with meta-learning based on context information

Jennifer Kreger, Lydia Fischer, Stephan Hasler, Thomas H Weisswange, Ute Bauer-Wersing

2017

Preprint:

This is an accepted article published in International Conference on Artificial Neural Networks (ICANN). The final authenticated version is available online at: https://doi.org/10.1007/978-3-319-68612-7_23

A priori reliability prediction with meta-learning based on context information

Jennifer Kreger[1], Lydia Fischer[2], Stephan Hasler[2], Thomas H. Weisswange[2], Ute Bauer-Wersing[1]

[1] Frankfurt University of Applied Sciences
Nibelungenplatz 1, 60318 Frankfurt am Main, Germany
[2] Honda Research Institute Europe GmbH
Carl-Legien-Str. 30, 63073 Offenbach am Main, Germany

Abstract. Machine learning systems are used in a wide variability of tasks, where reliability is very important. Often from the output of these systems their reliability cannot directly be deduced. We propose an approach to predict the reliability of a machine learning system externally. We tackle this by using an additional machine learning component we call meta-learner. This meta-learner can use the original input as well as supplementary context information for its judgment. With this approach the meta-learner can make a prediction of the performance of the machine learner before this one is actually executed. Based on this prediction unreliable decisions can be rejected and the systems reliability is retained. We show that our method outperforms certainty-based approaches at the example of road terrain detection.

Keywords: reliability prediction, meta-learning, rejection

1 Introduction

Lately machine learning and statistical classification techniques significantly gained in popularity and became an integral part of many daily life applications. Although machine learning algorithms perform well in many situations, sometimes they fail. Of course, wrong outputs have different impacts depending on the intended application, from e. g. user disappointment in the case of recommendation systems or automatic speech recognition systems, up to serious endangerment of life caused by safety critical systems such as self-driving cars or medical diagnosis systems. Hence, productive systems relying on machine learning methods must be able to estimate the reliability of the output in order to differentiate between situations where outputs are safe for further processing and those where failure is likely. Moreover, in real-time safety critical systems this decision has to be made fast.

Therefore research in the field calls more and more attention to reliability estimation of classification methods and efficient rejection strategies [1, 8, 14, 15]

The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-68612-7_23

as an effective way to improve the system's overall performance. Many state-of-the-art rejection approaches are based on certainty measures indicating whether a given data point is classified correctly or not. Uncertain data points are rejected dependent on a predefined threshold, that is obtained by optimizing the so-called error-reject tradeoff [6]. However, to design a proper certainty measure, detailed knowledge of the machine learning algorithm itself is required.

Another approach that aims to improve the performance of machine learners by investigating the system and its output is *meta-learning* [3, 10, 13, 16, 17]. Additional data is gathered that is not considered mandatory to solve the original task. Meta-learning approaches seek to find and use correlations between this additional knowledge and the machine learner's performance. The internal goal is not simply to reject results, but also to feed back information about the learning process into the system. Hence, to enable the system to adapt to the current situation and improve its performance, e. g. by selecting the features most suitable to solve the task under the current circumstances.

Both approaches, certainty-based and meta-learning-based, depend on observations and evaluations of the specific machine learning algorithm and rely on detailed internal knowledge of the machine learner itself. This implies that a model which is able to judge the reliability of one algorithm cannot necessarily be transferred to another machine learning approach. Moreover, methods using the machine learner's outputs or internal state parameters to estimate its performance must definitely execute the machine learning algorithm, even if the result on a data sample turns out as unreliable. This can be time consuming and as a consequence, triggering a defined fall-back mechanism may be late in the case of real-time safety critical applications.

Here, we propose an approach to rapidly predict the reliability of a machine learner's output independent of its underlying algorithm. For this purpose, we adapted the meta-learning approach (cf. Fig. 1). We use only i) the raw input sample and additional data from the environment that is available prior to processing of the machine learner and ii) a performance measure that is derived from the output of the machine learner and the ground truth data associated with the input sample. The machine learner and its output are no direct part of the reject decision, i. e. the prediction phase remains completely unknown to the meta-learner. We test our approach exemplary on two different vision-based segmentation systems, RTDS [9] and Bayesian SegNet [11]. We assume that even with very good models these systems may fail in case of technical issues as strong over-/underexposure or dirt on the camera lense. Also, the systems will probably fail in unknown environments like unpaved roads. For both machine learning applications we demonstrate the suitability of our approach and compare it to a certainty-based rejection method as state-of-the-art alternative. Please note, that our approach is *generic*, i. e. is – in principle – applicable to machine learning systems where ground truth data is available and a suitable performance measure can be deduced. Preliminary results of the study have been published in [12].

2 Methods

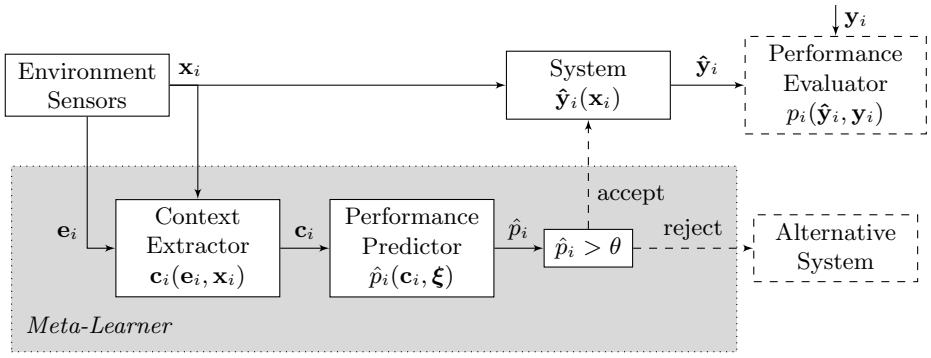

Fig. 1. Description of a general system (top row) enhanced with our *Meta-Learner* approach (gray box).

The structure of our approach is shown in Fig. 1. We assume to have a given *System* that generates output $\hat{\mathbf{y}}_i$ for input $\mathbf{x}_i$. Usually the *System* is a machine learner that was trained on a classification or regression task. Based on ground truth information $\mathbf{y}_i$ a performance measure p_i can be computed offline.

The aim of our *Meta-Learner* extension (Fig. 1: gray box) is to predict the performance of the *System* for the current input $\mathbf{x}_i$ at run time. For this it computes context features $\mathbf{c}_i$ on sensor measurements of the environment that are usually composed of the *System*'s input $\mathbf{x}_i$ and an additional input $\mathbf{e}_i$. These context features are used by the *Performance Predictor* to estimate a scalar value $\hat{p}_i$ that is later compared against a threshold θ to decide if the *System*'s output $\hat{\mathbf{y}}_i$ will be credible or not. In the latter case potentially an *Alternative System* could be used to handle the current situation. This *Alternative System* is usually more costly.

The important aspect of our *Meta-Learner* method is that it treats the *System* as a black box, i. e. it has no knowledge about the *System*'s inner structure and state. Furthermore, at run time the final decision of *accepting* or *rejecting* does not depend on the *System*'s output $\hat{\mathbf{y}}_i$. Therefore the *Meta-Learner* can perform an early reject decision and potentially save computation time.

In general, the *Performance Predictor* is a regression model that estimates the true performance of the *System*. Hence the regression parameters $\boldsymbol{\xi}$ are chosen as to minimize the prediction error between p_i and the estimated *System*'s performance $\hat{p}_i$ for a given training set $\{\mathbf{c}_i, p_i\}_{i=1}^{N}$. Again the *System*'s output $\hat{\mathbf{y}}_i$ is not used directly, but in its *abstract* form as a real-valued performance measure p_i. To compute p_i, ground truth information $\mathbf{y}_i$ is required.

For choosing an optimal threshold θ, the costs for a *reject* decision and a false *accept* decision have to be taken into account. Rejecting a sample should

increase the *Evaluated Performance* for the remaining samples. Note that the costs of rejecting a sample directly depend on the *Alternative System*. An optimal threshold can only be chosen if all costs are known.

3 Application

We apply our meta-learning approach in the context of computer vision and image segmentation, precisely we focus on machine learning methods performing segmentation tasks on complex traffic scenes. In this setting a input data sample $\mathbf{x}_i$ is an RGB image, the output of the image segmentation systems $\hat{\mathbf{y}}_i$ is an array of class labels indicating for each pixel in $\mathbf{x}_i$, whether it belongs to e. g. road terrain or any other defined class that can occur in traffic scenes. Accordingly, we need a measure that takes the performance p_i accumulated over a whole image into account and that is later predicted by our meta-learner. We choose the *Quality* score [9] as performance measure:

$$p_i := p_i(\hat{\mathbf{y}}_i, \mathbf{y}_i) = \frac{TP}{TP + FN + FP} \quad \in [0, 1] \tag{1}$$

where the quantities true positives (TP), false negatives (FN), and false positives (FP) are determined by comparing the system's output $\hat{\mathbf{y}}_i$ to labeled ground truth images $\mathbf{y}_i$ pixel-wise. The *Quality* score does not depend the true negatives (TN) and is therefore less biased by class imbalances. For multiple classes the quality is averaged over all classes.

To design a system for our scenarios that is computationally inexpensive and fast, we restrict the extraction of context features $\mathbf{c}_i$ to: 1) methods describing the image $\mathbf{x}_i$ holistically, and 2) methods resulting in low dimensional feature vectors. We therefore choose the global image descriptor Centrist [18] complemented with a color histogram. So far we do not consider further input $\mathbf{e}_i$ from the environment.

The first machine learning system we apply our meta-learner to is the road terrain detection system (RTDS [9]). It differentiates between road and non-road areas (Fig. 2 a–d). RTDS returns a confidence value for each pixel indicating the probability of belonging to road area in a 2-dimensional bird's eye view. Thresholding on the confidence map generates the segmentation result $\hat{\mathbf{y}}_i^{\mathrm{RTDS}}$ for road and non-road area of the corresponding input image $\mathbf{x}_i$. The original data set consists of N_{nat}=204 raw natural images of traffic scenes for which hand-labeled ground-truth is available. Unfortunately, the original data set is biased and contains only few samples where RTDS has poor results. Since it is the overall goal to provide an efficient reject strategy, examples where the machine learning method fails are vital to our approach. Therefore, we artificially enriched the original data set: we added disturbances by applying different sorts of global noise to some of the natural images (γ-correction, Gaussian blur, and salt noise). In summary, we extend the original data set with a set of artificially disturbed images to $N_{\mathrm{nat+art}}$=536 images in total.

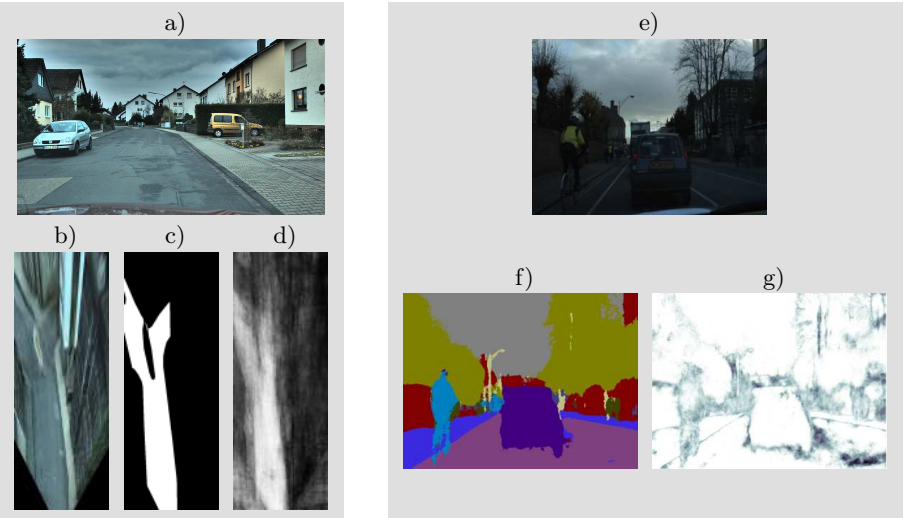

Fig. 2. RTDS input image (a), image transformed into a 2D bird's eye view (b), hand labeled ground truth (c), output confidence matrix (d).
Bayesian SegNet [11] input image (e), output segmentation prediction (f), and certainty output (g, light means high certainty).

The second machine learning system is Bayesian SegNet[3] [11]. It is a segmentation system designed especially for traffic scene analysis and handles 11 related classes. The output $\hat{\mathbf{y}}_i^{\text{SegNet}}$ is a color coded image indicating the predicted classes. Additionally there is a certainty map for each class that are averaged into a single certainty map (Fig. 2 e–g). Bayesian SegNet is trained on the CamVid data [4] with a test set of N_{camvid}=232 images.

In the experiments we compare the rejection strategy based on the quality predicted by our meta-learners against the true quality (ground truth) and a state-of-the-art certainty-based approach [8]. We consider three different regression models for our meta-learner: (i) AdaBoost [7] with decision trees, known to be robust and fast, (ii) support vector machines (SVM [5]), as a more sophisticated model, and (iii) k-nearest neighbor regression (KNN [2]), as a simple model[4].

4 Experiments

To evaluate our meta-learner we plot the mean quality over the rate of rejected samples. Therefor we iteratively increase the rejection threshold θ starting with no rejection ($\theta = 0$, rejection rate $r = 0$) such that one more data sample is rejected. When one data sample is remaining we stop the rejection.

[3] Available at: `https://github.com/alexgkendall/caffe-segnet`
[4] We use the implementations of scikit learn available at: `http://scikit-learn.org/stable`

$$r(\theta) = \frac{1}{N} \sum \begin{cases} 1 & \hat{p}_i <= \theta \\ 0 & else \end{cases} \tag{2}$$

After each rejection we compute the mean quality of the machine learner for the remaining data samples:

$$p_{\text{mean}}(\theta) = \frac{1}{N} \sum \begin{cases} p_i & \hat{p}_i > \theta \\ 0 & else \end{cases} \tag{3}$$

Note that a small rejection rate causing a high increase of the mean quality is appreciated. Due to the small amount of available data we use leave-one-out cross validation averaging over 10 repetitions.

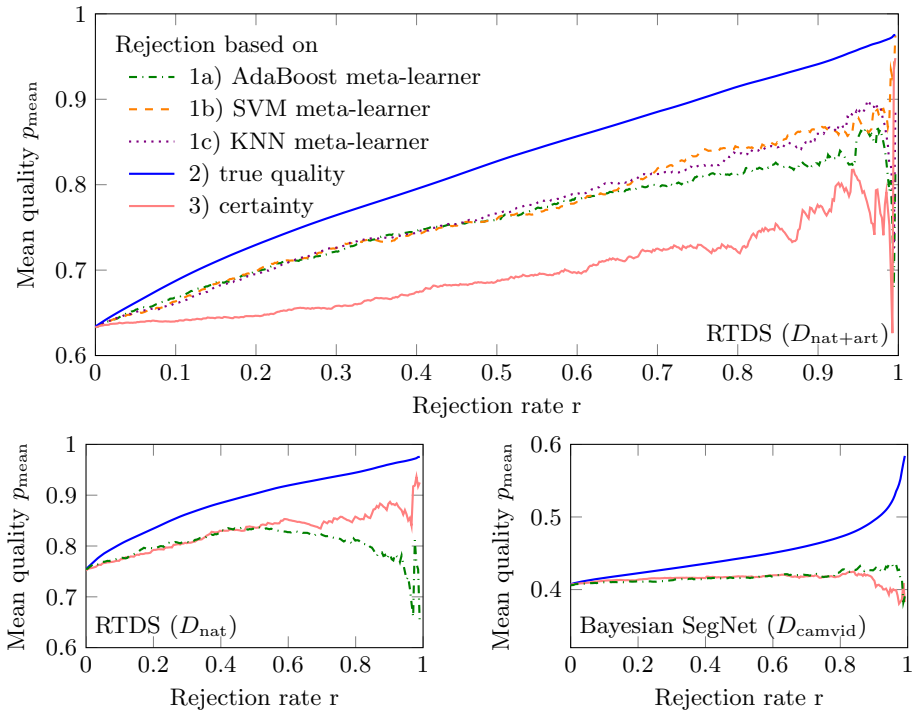

Fig. 3. Comparison of different rejection approaches. The worst sample is rejected dependent on 1) the prediction of the meta-learner, 2) the true quality, 3) the classification certainty. Top: comparison of different regression models – (a) Ada-Boost, (b) SVM, (c) KNN – on RTDS with $D_{\text{nat+art}}$. Bottom left: RTDS with D_{nat}. Bottom right: Bayesian SegNet with D_{camvid}.

In Fig. 3 (top) we compare the rejection performance of different meta-learners using varying regression models (AdaBoost, SVM, KNN) to the ground

truth and a certainty-based approach. For the related experiments we use the data set $D_{\text{nat+art}}$ with artificial and natural images, as well as the obtained RTDS qualities. It can be seen that the meta-learners almost reach the best possible performance obtained with the true output quality for the interesting regime of small rejection rates (<10%). For higher rates the difference increases but it is still outperforming the certainty-based approach. The results also state that there is nearly no difference between the analyzed meta-learners. Therefore we use AdaBoost in the following because it is known to deliver robust and fast results.

The bottom plots show the RTDS and SegNet results only using natural images. In both scenarios our approach provides similar results as the certainty-based approach. For the RTDS both rejection approaches reach a reasonable improvement of the mean quality on the natural images with low rejection rates. The difficulty on the Bayesian SegNet is the homogeneous distribution of the related quality values. Therefore both rejection approaches achieve only slight improvements of the mean output quality, but even the rejection based on the true quality values can barely achieve an increase of the quality.

5 Conclusion and Future Work

Machine learning methods are used in a wide field of applications wherein reliability is very important. Many of these methods lack information how reliable their output is. Often they are enhanced with a certainty estimation based on in depth knowledge of the machine learner itself. We relax this constraint by estimating the reliability externally only incorporating the raw input data and the machine learner's output performance for training. The meta-learner is independent of the machine learner. The input data can be enriched with information extracted from the environment. For the application solely the input is required. An additional benefit of this approach is that we rapidly determine if the output of the machine learner is reliable without executing it. Hence, the aim of our approach is to reject samples where the machine learner is predicted to fail.

We demonstrate the usefulness of our approach on two exemplary vision-based applications tackling two different traffic scene analysis tasks. Therefore we compare our approach with an optimal baseline and a state-of-the-art certainty-based approach. It turned out that our approach provides almost as good results as the baseline in the relevant regime of small rejection rates and shows a better or similar performance as the certainty-based rejection. Hence, the rejection approach based on the meta-learner's predictions can improve the machine learner's mean performance.

One future research direction could be to incorporate features extracted from the environment. A further interesting approach is to investigate whether one can train the meta-learner to predict the quality of a future data sample. This would be useful in applications dealing with image streams, especially for autonomous cars where a time dependency between data samples is guaranteed. In this case it might be beneficial as well to apply a more sophisticated rejection strategy, e. g.

an adaptive threshold, that adjusts to the current data distribution, assuming an online learning scenario.

References

1. Alvarez, I., Bernard, S., Deffuant, G.: Keep the Decision Tree and Estimate the Class Probabilities Using its Decision Boundary. In: International Joint Conference on Artificial Intelligence. pp. 654–659 (2007)
2. Andoni, A., Indyk, P.: Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In: Annual IEEE Symposium on Foundations of Computer Science. pp. 459–468. IEEE (2006)
3. Brazdil, P., Carrier, C.G., Soares, C., Vilalta, R.: Metalearning: Applications to data mining. Springer Science & Business Media (2008)
4. Brostow, G.J., Fauqueur, J., Cipolla, R.: Semantic object classes in video: A high-definition ground truth database. Pattern Recognition Letters 30(2), 88–97 (2008)
5. Chang, C.C., Lin, C.J.: LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology 2(3), 1–27 (2011)
6. Chow, C.: On optimum recognition error and reject tradeoff. IEEE Transactions on information theory 16(1), 41–46 (1970)
7. Drucker, H.: Improving regressors using boosting techniques. In: International Conference on Machine Learning. vol. 97, pp. 107–115 (1997)
8. Fischer, L., Hammer, B., Wersing, H.: Efficient rejection strategies for prototype-based classification. Neurocomputing 169, 334–342 (2015)
9. Fritsch, J., Kühnl, T., Kummert, F.: Monocular Road Terrain Detection by Combining Visual and Spatial Information. IEEE Transactions on Intelligent Transportation Systems 15(4), 1586–1596 (2014)
10. Giraud-Carrier, C.: Metalearning-a tutorial. In: Tutorial at the International Conference on Machine Learning and Applications. pp. 11–13 (2008)
11. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. arXiv:1511.02680 (2015)
12. Kreger, J., Fischer, L., Hasler, S., Bauer-Wersing, U., Weisswange, T.H.: Quality prediction for a road detection system. In: Machine Learning Reports 04/2016. pp. 93–94 (2016)
13. Lemke, C., Budka, M., Gabrys, B.: Metalearning: a survey of trends and technologies. Artificial Intelligence Review 44(1), 117–130 (2015)
14. Lu, H., Wei, S., Zhou, Z., Miao, Y., Lu, Y.: Regularised extreme learning machine with misclassification cost and rejection cost for gene expression data classification. International journal of data mining and bioinformatics 12(3), 294–312 (2015)
15. Pillai, I., Fumera, G., Roli, F.: Multi-label classification with a reject option. Pattern Recognition 46(8), 2256–2266 (2013)
16. Rossi, A.L.D., Carvalho, A.C., Soares, C.: Meta-learning for periodic algorithm selection in time-changing data. In: IEEE Brazilian Symposium on Neural Networks. pp. 7–12. IEEE (2012)
17. Vilalta, R., Drissi, Y.: A perspective view and survey of meta-learning. Artificial Intelligence Review 18(2), 77–95 (2002)
18. Wu, J., Rehg, J.M.: CENTRIST: A visual descriptor for scene categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(8), 1489–1501 (2011)