Analytically tractable sample-specific confidence measures for semi-supervised learning

Tuo Cui[1], Arne Grumpe[1], Matthias Hillebrand[2], Ulrich Kreßel[2], Franz Kummert[3], Christian Wöhler[1]

[1]Image Analysis Group, TU Dortmund, 44227 Dortmund, Germany
[2]Daimler AG, Group Research and Advanced Engineering, 89081 Ulm, Germany
[3]Applied Informatics, Bielefeld University, 33615 Bielefeld, Germany

1 Introduction

In general, classifiers require a large number of labelled training samples preferably covering a wide range of variation to yield a comprehensive and generalising recognition behaviour. However, manual labelling of huge data sets is costly and time-consuming, which is the main motivation for the development of semi-supervised learning algorithms which autonomously extend the “knowledge” gained by classifiers based on a small amount of initial, manually labelled training samples towards increasingly different representatives of the regarded pattern classes.

Comprehensive overviews of the state of the art in the field of semi-supervised learning are provided e.g. by Zhu and Goldberg [15], Seeger [13], and Chapelle et al. [5]. According to Zhu and Goldberg [15], the approach of “self-training” is characterised by an iterative procedure during which new samples are selected from the large set of available unlabelled samples, and the selected samples are rejected or accepted and autonomously labelled by the classifier. Upon acceptance, the samples are added to the training set along with their autonomously generated labels. At the end of each iteration, re-training of the classifier is performed using the extended training set, and the next training cycle is started.

In this study, we describe a framework for semi-supervised learning of classifiers relying on the concept of self-training. Specifically, we propose a sample selection mechanism based on analytically tractable confidence measures which are inferred for an “ensemble” of two different classifiers, which in this study consists of a polynomial classifier (PC) and on a support vector regression (SVR) algorithm. In this setting of semi-supervised ensemble learning, we compare different confidence measures for unlabelled training samples based on the MNIST handwritten digits data set and a traffic sign data set acquired from a test vehicle.

2 Utilised classification methods

This section provides a brief overview of the two different classification approaches employed in Section 5 in the context of semi-supervised ensemble learning.

2.1 The polynomial classifier

For the PC, a sample is represented as the polynomial structure list $\vec{\Phi}(\vec{x})$ which consists of multiplicative combinations of the elements of the original feature vector $\vec{x}$ which are of an order lower than or equal to the order N of the classifier. In this study, we use a fully quadratic polynomial classifier with $N = 2$, such that $\vec{\Phi}(\vec{x})$ contains all linear and quadratic multiplicative combinations of the original features. The decision function $\vec{d}(\vec{x})$ of the polynomial classifier, which estimates the class membership of the sample, is given by

$$\vec{d}_{\mathrm{PC}}(\vec{x}) = \mathbf{A}^T \vec{\Phi}(\vec{x}) \tag{1}$$

The matrix $\mathbf{A}$ denotes the coefficient matrix which is adjusted during the training process. The elements of the vector $\vec{d}_{\mathrm{PC}}(\vec{x})$ always sum up to 1. A detailed description of the concept of polynomial classifiers is given in [12].

2.2 Support vector regression

We employ the SVR approach rather than the more commonly used support vector machine (SVM) classifier as due to the binary output of the SVM it cannot be integrated into the framework of confidence bands (cf. Section 3.4) in a straightforward manner. The goal of SVR or more precisely ϵ-SVR is to find a function of the input vector $\vec{x}$ that deviates at most by an amount ϵ from the target value $d_{\mathrm{SVR}}^{\mathrm{target}}$. The SVR is described in detail in [14], such that we only repeat the basic equations here.

In the linear case the SVR function is of the form

$$d_{\mathrm{SVR}}(\vec{x}) = \sum_{i=1}^{N_{SV}} \alpha_i \vec{x}_i^T \vec{x} + b, \tag{2}$$

where $\vec{x}_i$ denotes the i-th support vector, N_{SV} is the number of support vectors and α_i is the corresponding weight. This basic approach can be extended to nonlinear functions in a manner similar to the PC by transforming the input vector as well as the support vectors into a higher-dimensional space by a nonlinear mapping $\vec{\Phi}(\vec{x})$ according to

$$d_{\mathrm{SVR}}(\vec{x}) = \sum_{i=1}^{N_{\mathrm{SV}}} \alpha_i \vec{\Phi}(\vec{x}_i)^T \vec{\Phi}(\vec{x}) + b = \sum_{i=1}^{N_{\mathrm{SV}}} \alpha_i \tilde{K}(\vec{x}_i, \vec{x}) + b. \tag{3}$$

The higher-dimensional scalar product can be computed indirectly using the kernel function $\tilde{K}(\vec{x}_i, \vec{x})$. Throughout this work a second-order polynomial kernel of the form

$$\tilde{K}(\vec{x}_i, \vec{x}) = (c\vec{x}_i^T \vec{x} + 1)^2 \tag{4}$$

is used. The parameter c is set to $1/\left\langle \vec{x}_i^T \vec{x}_j \right\rangle_{\mathrm{train}}$, where $\langle \ldots \rangle_{\mathrm{train}}$ denotes the average over the training set. In our pairwise classification scenarios, the target value $d_{\mathrm{SVR}}^{\mathrm{target}}(\vec{x})$ always corresponds to one of the class labels $+1$ and -1. The SVR implementation used in this study is the publicly available LIBSVM toolbox[1] described in [4].

3 Uncertainty of automatic labelling

The confidence of automatically generated labels is one of the main problems that need to be addressed for semi-supervised learning. Adding wrongly labelled data to the training set will result in a poor classification accuracy. This section reviews several methods for measuring the uncertainty.

[1] http://www.csie.ntu.edu.tw/∼cjlin/libsvm

3.1 Empirical uncertainty

The sample based estimation of probability distributions is a common practice in statistics. In terms of classification uncertainty, a sample of the distribution is represented by a classifier. Therefore, the "empirical uncertainty" can be expressed as the standard deviation of class-specific probabilities computed by classifiers trained on different disjoint training sets. Since the amount of training data increases drastically with the number of trained and compared classifiers, this confidence measure is not well suited for semi-supervised learning algorithms which attempt to reduce the required number of manually labelled training samples. Due to the empirical nature of this measure, its estimate of the classification uncertainty approaches the true uncertainty with increasing number of disjoint training sets, such that it will be regarded as a reference value to which all other derived confidence measures will be compared.

3.2 Maximum likelihood

A widely used method that is believed to reflect the classification uncertainty is the maximum class likelihood. In the case of the PC, the output of the polynomial function can be interpreted as a class likelihood because the sum of all outputs equals to one. This is not to be confused with the real probability of the sample belonging to the winning class as the outputs can also be negative or larger than one. The maximum likelihood (ML) is defined by $\max\limits_k d_k$ where d_k is the probability of class k. It is shown in [2] that this method is not well suited for self-learning and leads to a "self-confidence" problem since errors are reinforced. As this method is still widely used it will be evaluated as well.

3.3 RAD criterion

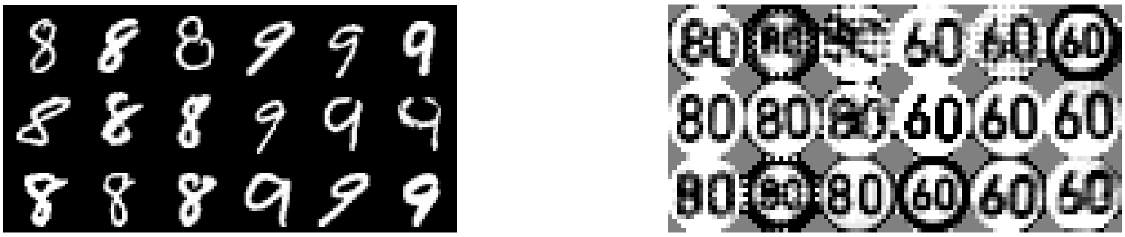
The RAD criterion is another popular measure which is widely believed to be correlated with the uncertainty of the PC. The RAD criterion is defined as the Euclidean distance in the decision space of the PC to the closest target class label. This leads to a correlation with the maximum likelihood method (in a pairwise classification problem, these values are perfectly correlated). The RAD criterion can be computed according to

$$\mathrm{RAD} = \min_k \left\| \vec{d}_{\mathrm{PC}}(\vec{x}) - \vec{d}_{\mathrm{PC}}^{\,\mathrm{target},k}(\vec{x}) \right\|_2 \tag{5}$$

where $\vec{d}_{\mathrm{PC}}(\vec{x})$ denotes the output vector of the PC and $\vec{d}_{\mathrm{PC}}^{\,\mathrm{target},k}(\vec{x})$ the target output vector of class k.

3.4 Confidence bands

In [7] normalised confidence bands are used in the context of semi-supervised learning. Confidence bands are curves enclosing a model function being estimated by a regression analysis. They represent the interval in which the true model is expected to reside with a probability of $1 - \alpha$. Many approaches to the computation of confidence bands exist, including e.g. Monte Carlo techniques [9], bootstrapping methods [6], or analytical approaches [8, 11]. According to Kardaun [8] and Martos et al. [11], the computation of



(a)

(b)

Figure 1: Samples of the (a) MNIST data set and (b) traffic sign data set.

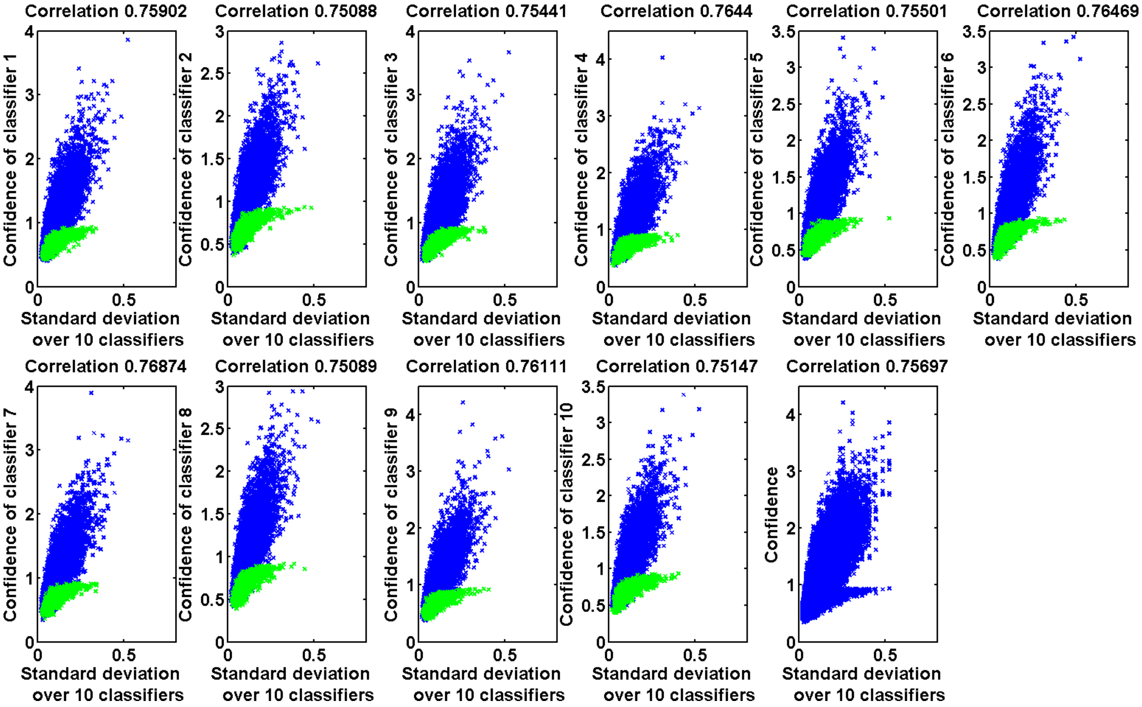


Figure 2: Normalised confidence band value σ_{n} vs. empirical uncertainty for the PC using the MNIST data set. The part of the training data which is known to the respective classifier is marked in green, the unknown part in blue.

confidence bands requires the covariance matrix of the parameters, which are estimated based on the Jacobian matrix $\mathbf{J}$ given by the elements

$$J_{ij} = \frac{\partial r_i}{\partial a_j}, \tag{6}$$

where $r_i = d(\vec{x}_i) - d^{\text{target}}(\vec{x}_i)$ denotes the residual error of the sample $\vec{x}_i$ and a_j the parameters of the classifier function. The matrix $\mathbf{K}$ with the elements

$$K_{ij} = \frac{J_{ij}}{\sigma_i} \tag{7}$$

is the Jacobian matrix divided by the uncertainty σ_i of the sample-specific label $d^{\text{target}}(\vec{x}_i)$. In the context of regression, the value of σ_i corresponds to the measurement error. In our classification setting, it may refer to a finite probability of mislabelling by the human expert or to cases in which a class membership cannot be determined unequivocally due

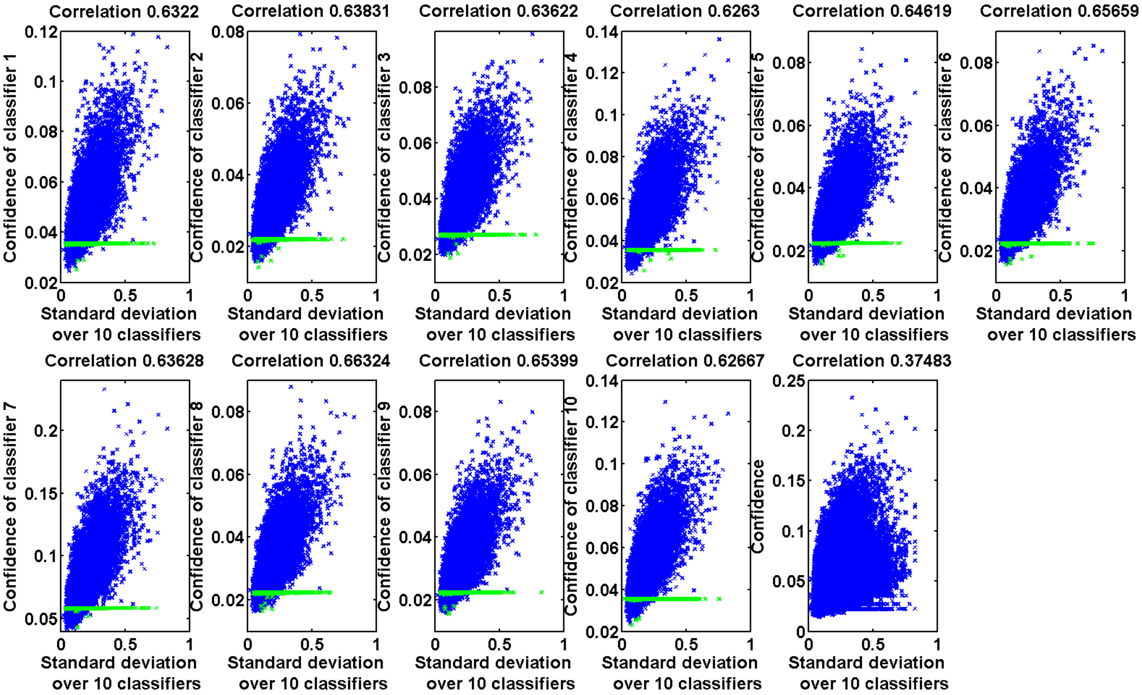


Figure 3: Normalised confidence band value σ_n vs. empirical uncertainty for the SVR with polynomial kernel using the MNIST data set. Colours as in Fig. 2.

to poor data quality. The matrix $\mathbf{K}$ is then used to estimate the covariance matrix $\mathbf{C}$ according to

$$\mathbf{C} = (\mathbf{K}^T\mathbf{K})^{-1}. \tag{8}$$

To compute the confidence band value for the transformed sample $\vec{w} = \vec{\Phi}(\vec{x}_i)$, we use the Jacobian vector $\vec{g}$ of the model function $d(\vec{w})$ with $g_j(\vec{w}) = \partial d(\vec{w})/\partial a_j$ and the covariance matrix $\mathbf{C}$ to obtain the dimensionless value

$$c_0(\vec{w}) = \vec{g}^T\mathbf{C}\vec{g}. \tag{9}$$

The half width σ_C of the confidence interval then corresponds to

$$\sigma_C(\vec{w}) = \frac{\beta}{2}\sqrt{c_0(\vec{w})\frac{\mathrm{R}}{\nu}}, \tag{10}$$

where $R = \sum_i {r_i}^2$ is the residual sum of squares with r_i being the residual of sample i, and $\nu = N - N_p$ the number of degrees of freedom, where N is the number of samples and N_P the number of free model parameters. The constant β is derived from the inverse cumulative t-student distribution t_cdf^{-1} according to $\beta = t_\mathrm{cdf}^{-1}(1 - \alpha/2, \nu)$ with α as the probability of the desired confidence level for the bands. We set $\alpha = 0.05$ throughout this study. Since it is not obvious how to determine the value of σ_i, the confidence band is normalised to σ_i as proposed in [7], where a uniform value $\sigma_i \equiv \sigma$ is assumed. The correspondingly normalised confidence band

$$\sigma_\mathrm{n}(\vec{w}) = \frac{\sigma_C(\vec{w})}{\sigma} \tag{11}$$

is independent of the value of σ as $c_0(\vec{w}) \propto \sigma^2$ according to Eqs. (7) and (8).

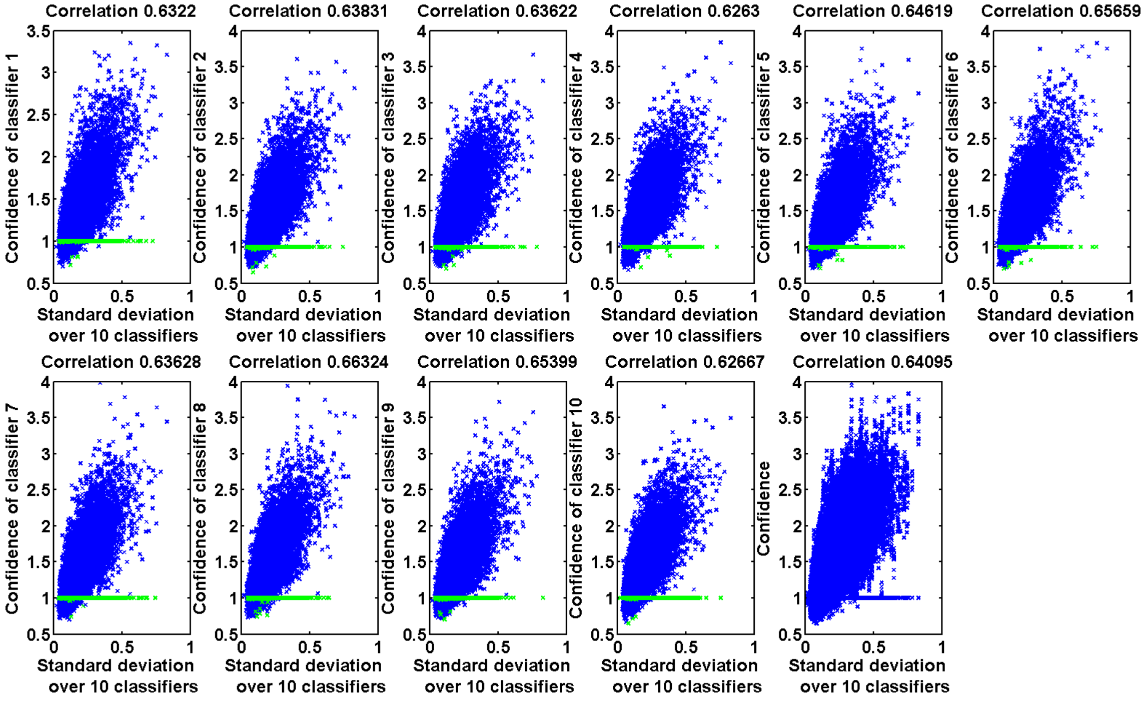


Figure 4: Renormalised confidence band value σ_{B} vs. empirical uncertainty for SVR with polynomial kernel using the MNIST data set. Colours as in Fig. 2.

Due to the dependence of $\mathbf{C}$ on the Jacobian matrix the confidence band can be computed for every classifier which allows to compute the derivative of the residuals with respect to the model parameters. Within this work we specifically regard the PC and the SVR. An extension to other classifiers such as neural networks is straightforward.

In the case of the PC we have $\mathbf{J}_{\mathrm{PC}} = [\vec{\Phi}(\vec{x}_1)\ \vec{\Phi}(\vec{x}_2)\ \ldots\ \vec{\Phi}(\vec{x}_M)]^T$. Its number of free model parameters can be computed according to $N_p^{\mathrm{PC}} = (L-1)M$, where L denotes the number of classes and M the number of elements in the polynomial structure list $\vec{\Phi}(\vec{x})$. This relation follows directly from the fact that all elements of the decision vector of the PC sum up to 1, which imposes a constraint on each row of the parameter matrix $\mathbf{A}$.

For the SVR, the number N_p^{SVR} of free parameters is equal to the number of support vectors, as the output of the SVR is a weighted linear combination of all support vectors and the parameters are the weights. The Jacobian matrix $\mathbf{J}_{\mathrm{SVR}}$ is equal to the kernel matrix $\tilde{\mathbf{K}}$ where the elements $\tilde{K}_{ij} = \tilde{K}(\vec{x}_j, \vec{x}_i)$ are the kernel function evaluated for the j-th support vector and the i-th training sample. Given the corresponding Jacobian matrix and the number of free parameters, respectively, the normalised confidence band can be computed for an arbitrary sample for the PC and the SVR according to Eqs. (8)–(11).

4 Comparison of uncertainty measures

In this section, the previously described uncertainty measures are compared with the empirical uncertainty over different classifiers trained on disjoint training sets.

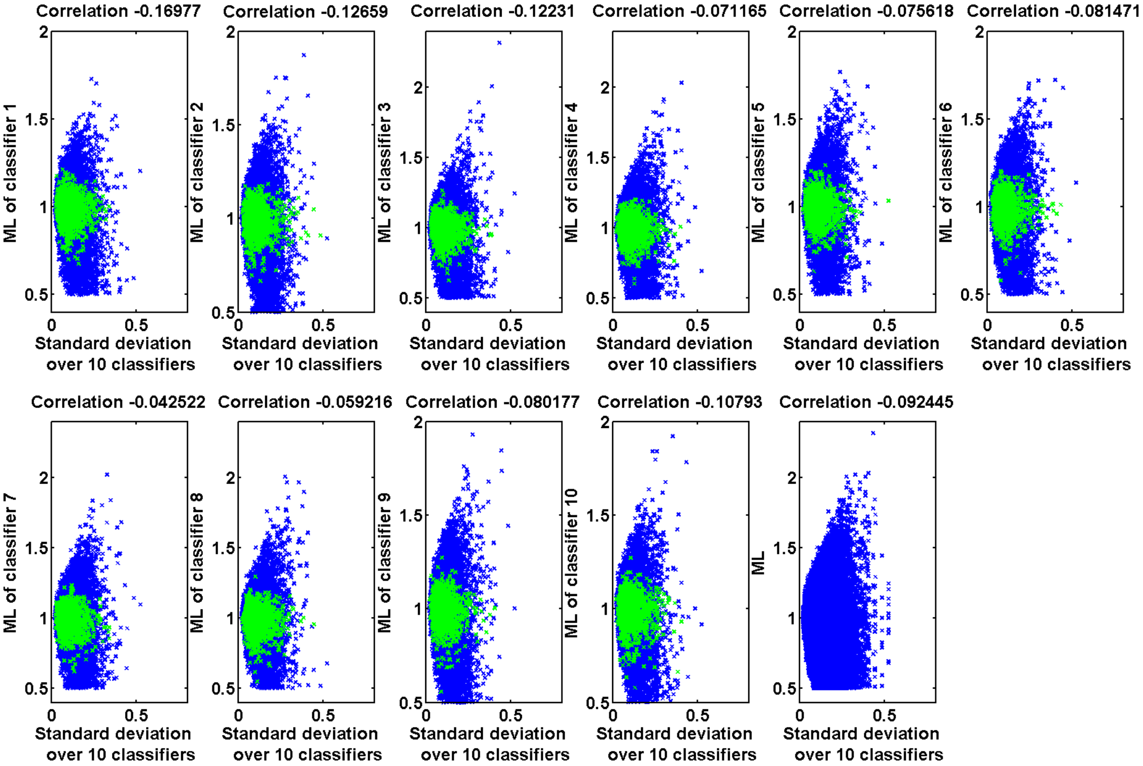


Figure 5: Maximum likelihood vs. empirical uncertainty for the PC using the MNIST data set. Colours as in Fig. 2.

4.1 MNIST data set

In order to determine the empirical uncertainty, we first used the MNIST handwritten digits data set[2] and specifically regarded the pairwise classification problem of separating the classes “8” and “9”. The training set consists of 5851 samples of class “8” and 4949 samples of class “9”, while the test set comprises 974 samples of class “8” and 1009 samples of class “9”. The training data are divided into ten equally sized disjoint training sets. The division is performed classwise and remaining examples are added to the test set. A principal component analysis (PCA) [12] with a reconstruction error of $r^2 = 0.25$ was applied to the data set.

Figs. 2 and 3 show the normalised confidence band values σ_n on the training set plotted against the empirical uncertainty for the PC and the SVR, respectively. Notably, the part of the training data known to the respective classifier (marked in green) is associated with low normalised confidence band values even when the empirical uncertainty is high, since nine of the ten classifiers have not been trained with these data. This implies that the “expert” trained with these data yields considerably lower confidence band values than the other classifiers. The average normalised confidence band value is different for each classifier, resulting in a poor overall correlation as shown in the bottom right plot. This behaviour could be reproduced for all classifiers. The plots for the test set bear no additional information as they look very similar to those obtained for the unused training data displayed in blue and are therefore omitted.

The fact that all known training samples are associated with an almost constant value of

[2] http://yann.lecun.com/exdb/mnist/

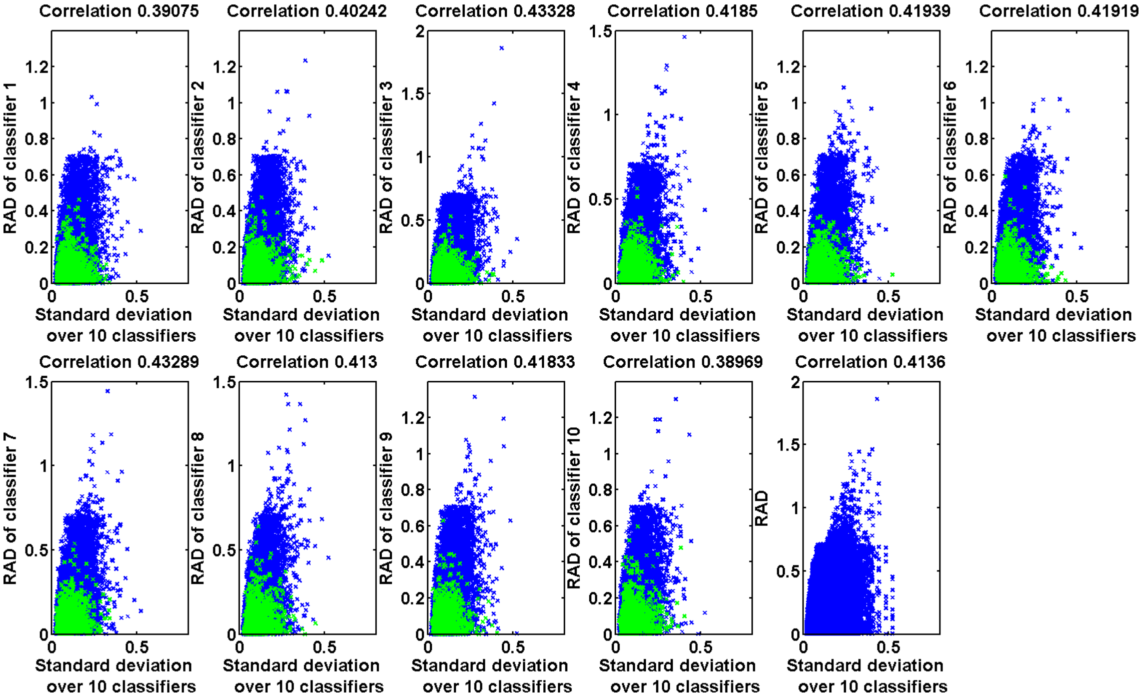


Figure 6: RAD criterion vs. empirical uncertainty for the PC using the MNIST data set. Colours as in Fig. 2.

classifier	training set norm. σ_{n}	training set renorm. σ_{B}	training set accuracy [%]	test set norm. σ_{n}	test set renorm. σ_{B}	test set accuracy [%]
2nd order PC	0.7570	0.7569	98.51	0.8108	0.8107	98.19
SVR pol. kernel	0.3748	0.6409	98.73	0.4281	0.7622	98.66

Table 1: Correlation between confidence band values and empirical uncertainty on the MNIST data set.

the normalised confidence band suggests that it might be advantageous to renormalise all computed confidence band values by the average normalised confidence band value $\langle\sigma_{\mathrm{n}}\rangle_{\mathrm{train}}$ of the training set, leading to the renormalised confidence band values given by

$$\sigma_{\mathrm{B}}(\vec{w}) = \frac{\sigma_{\mathrm{n}}(\vec{w})}{\langle\sigma_{\mathrm{n}}\rangle_{\mathrm{train}}}. \tag{12}$$

The resulting scatter plots are shown in Fig. 4. The non-overlapping areas in the cumulative plot have disappeared and are merged into one single "ray", leading to a drastically increased correlation value with respect to the empirical confidence. The overall correlation values for the renormalised as well as for the normalised confidence band values on both the training and the test set are summarised in Table 1. There appears to be a slight decrease in the correlation coefficient in the case of the PC but the effect is minor.

In contrast to the high correlation rates of the normalised and renormalised confidence band values, the maximum likelihood of the PC as well as the RAD criterion are largely uncorrelated with the empirical uncertainty. The corresponding scatter plots are shown in Figs. 5 and 6. The overall correlation coefficients correspond to -0.0924 for the maximum likelihood of the PC and to 0.4136 for the RAD criterion.

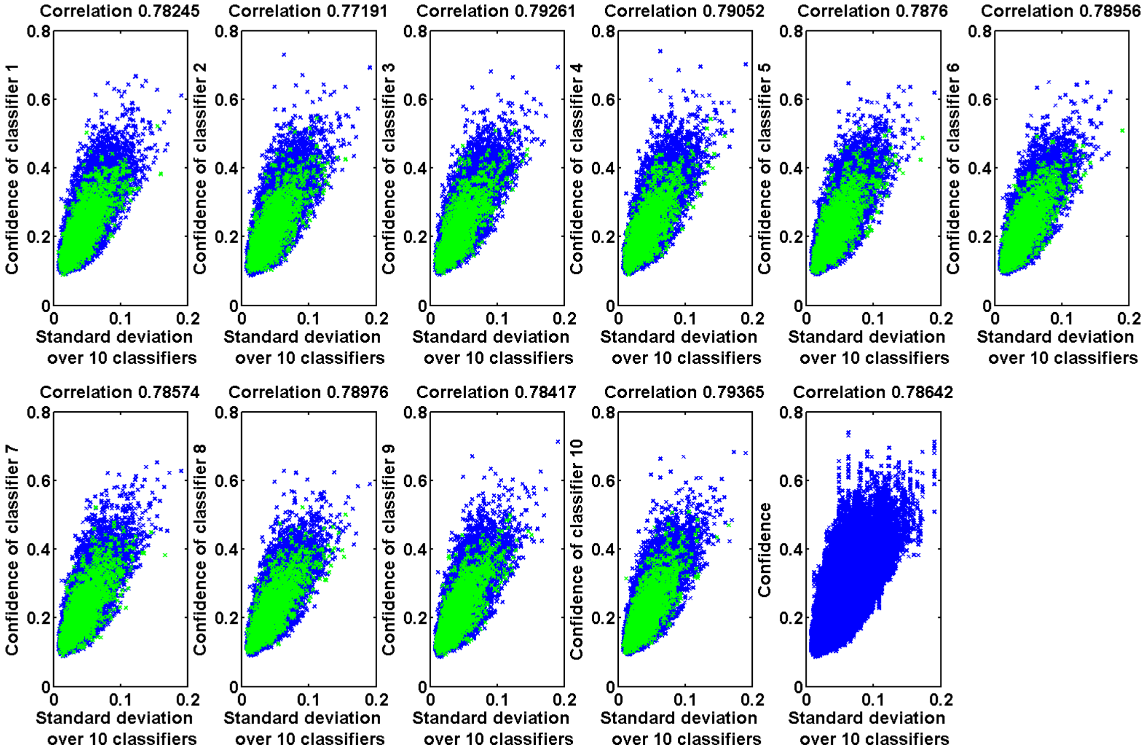


Figure 7: Normalised confidence band value σ_{n} vs. empirical uncertainty for the PC using the traffic sign data set. Colours as in Fig. 2.

4.2 Traffic sign data set

The traffic sign samples were extracted from input images acquired by a camera mounted behind the windscreen of a test vehicle. All circular shapes in the input images, among them the traffic signs, were extracted by a Hough transform based technique. The corresponding regions were scaled to a size of 17×17 pixels and normalised in contrast [10]. In this study, we specifically regard the classification problem of distinguishing speed limit signs of the classes "60 km/h" and "80 km/h". The utilised data set comprises 15000 samples of each class. It is divided into a training set consisting of 70% of the samples and a test set consisting of 30% of the samples. A PCA with a reconstruction error of $r^2 = 0.25$ was applied to the data set.

The results obtained for the traffic sign data set are similar to those presented for the MNIST data set, with the exception of the normalised confidence band values of the PC. The corresponding scatter plot of the normalised confidence band value σ_{n} vs. the empirical uncertainty is shown in Fig. 7, where no systematic difference between the distributions of the known part (green) and the unknown part (blue) of the training data is observed. The overall correlation values are presented in Table 2. For the PC, the difference between the correlation coefficients obtained for σ_{n} and σ_{B} is negligible. However, for the SVR the renormalised confidence band value σ_{B} shows an increased correlation coefficient when compared to the normalised confidence band value σ_{n}, as illustrated by Figs. 8 and 9. Similar to the MNIST data set, the known part of the training data (marked in green) is associated with low confidence band values even when the empirical uncertainty is high. For the maximum likelihood of the PC, the correlation coefficient amounts to -0.2089 and for the RAD criterion to 0.3464, which is fairly similar to the values obtained for the MNIST data set.

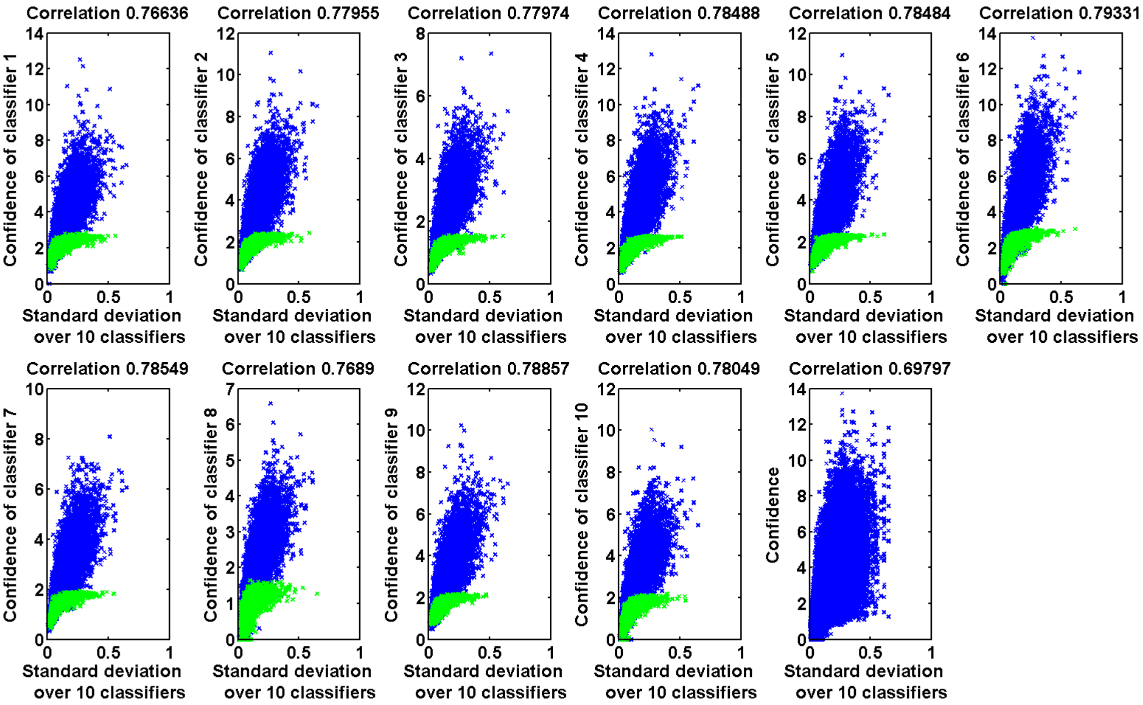


Figure 8: Normalised confidence band value σ_{n} vs. empirical uncertainty for the SVR using the traffic sign data set. Colours as in Fig. 2.

classifier		training set			test set	
	norm. σ_{n}	renorm. σ_{B}	accuracy [%]	norm. σ_{n}	renorm. σ_{B}	accuracy [%]
PC 2nd order	0.7865	0.7864	97.65	0.7802	0.7802	97.56
SVR pol. kernel	0.6980	0.7737	99.68	0.7393	0.8216	99.49

Table 2: Correlation between confidence band values and empirical uncertainty on the traffic sign data set.

5 Semi-supervised learning

In order to create an elementary semi-supervised ensemble learning scenario, a two-classifier setup is chosen which can be seen as a small ensemble. The first classifier is a SVR, which is known to be able to cope with small amounts of training data but tends to be sensitive with respect to wrong training labels. The second classifier is a PC, which may need more data but is less sensitive to wrong training labels. Both classifiers are trained on an initial, manually labelled training set of variable size, comprising a few percent of the available training data. The test set has a fixed size of 30% of the overall size of the data set.

The trained classifiers provide labels for the training samples by themselves, where a new label is accepted and added to the training set if (i) both classifiers yield the same label and (ii) the confidence band values of both classifiers are within the range $\sigma_{\mathrm{B,min}} \leq \sigma_{\mathrm{B}} \leq \sigma_{\mathrm{B,max}}$. The acceptable range of the renormalised confidence band value σ_{B} is determined for each classifier using $\sigma_{\mathrm{B,min}} = 0.9 \, \langle \sigma_{\mathrm{B}} \rangle_{\mathrm{train}}$ and $\sigma_{\mathrm{B,max}} = 3.0 \, \langle \sigma_{\mathrm{B}} \rangle_{\mathrm{train}}$. A new training cycle is initiated when 100 new samples have been inserted into the training set. This procedure is repeated until all available samples that match the above criteria have been

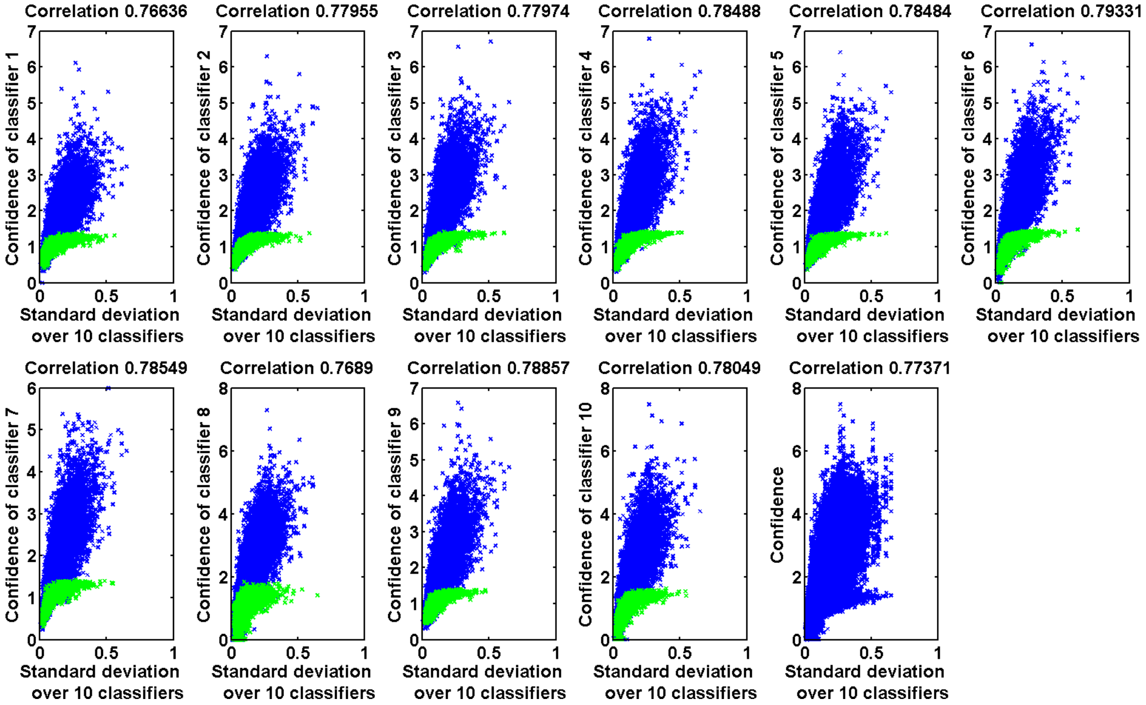


Figure 9: Renormalised confidence band value σ_{B} vs. empirical uncertainty for the SVR with polynomial kernel using the MNIST data set. Colours as in Fig. 2.

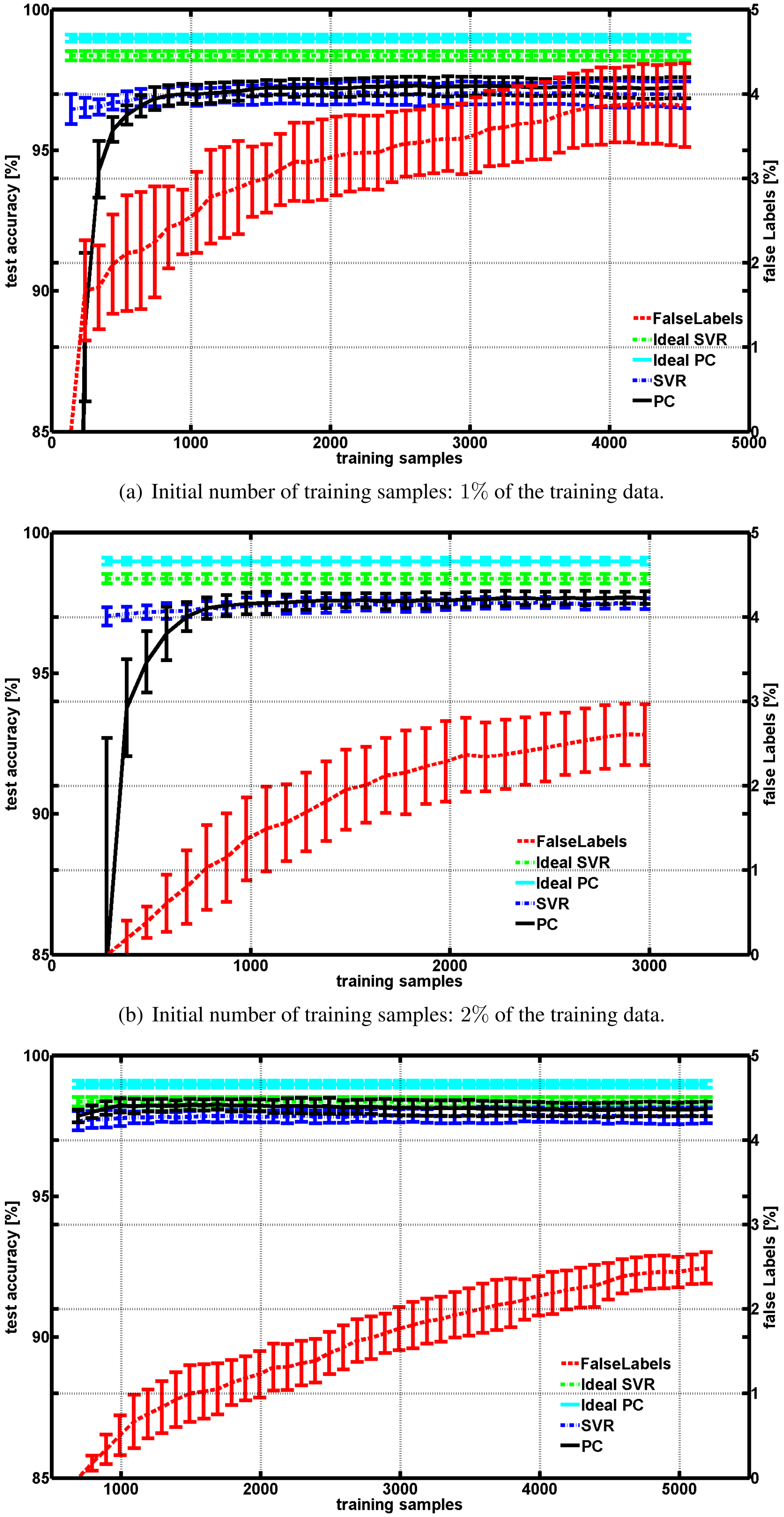
added to the training set with their autonomously generated labels. If no new training samples are found that fulfill the selection criterion, the semi-supervised learning process terminates automatically.

The semi-supervised learning setup is applied to the MNIST data set and the traffic sign data set. PCA-based dimensionality reduction is performed with a reconstruction error of $r^2 = 0.3$ for the MNIST data set and $r^2 = 0.2$ for the traffic sign data set. Both PCAs are computed based on the initial training set, respectively.

5.1 Results obtained for the MNIST data set

For the MNIST data set, the size of the initial, manually labelled training set was set to 1%, 2% and 5% of the overall size of the training data set in order to identify the influence of the size of the initial training set. The corresponding learning curves are displayed in Fig. 10. Due to the random division into training and test data, different results are observed for each individual run of the semi-supervised learning scheme. Hence, the learning curves in Fig. 10 denote the mean and the standard deviation over ten runs, interpolated to equally spaced numbers of training samples using Akima interpolation [1]. All accuracies on the test set are compared with the accuracy of an “ideal” classifier trained on all available manually labelled training samples.

The initial accuracy increases with the number of samples in the initial training set. In the course of the semi-supervised learning process, the recognition accuracies of the two classifiers increase at the beginning but may decrease again if too many false labels are added to the training set. The ideal accuracy of 98.35% of the SVR and 98.98% of the PC has not been reached. It is noteworthy that by reducing the initial training set size to 2% of the overall size of the training data set, the SVR still yields good accuracy values

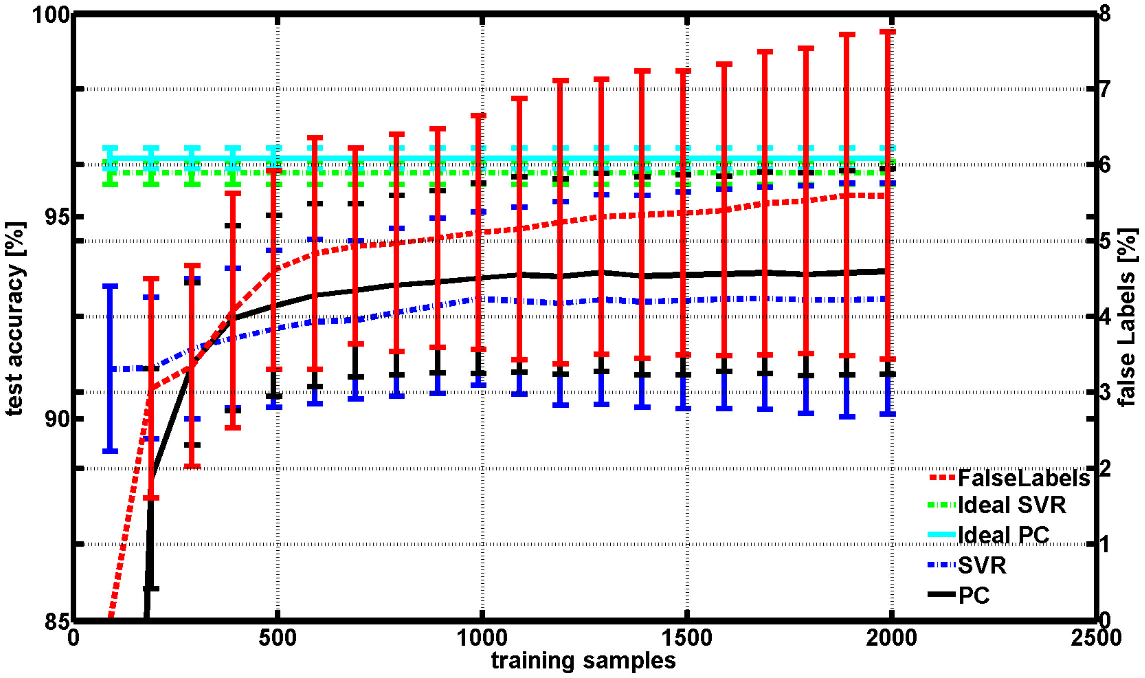


(a) Initial number of training samples: 1% of the training data.

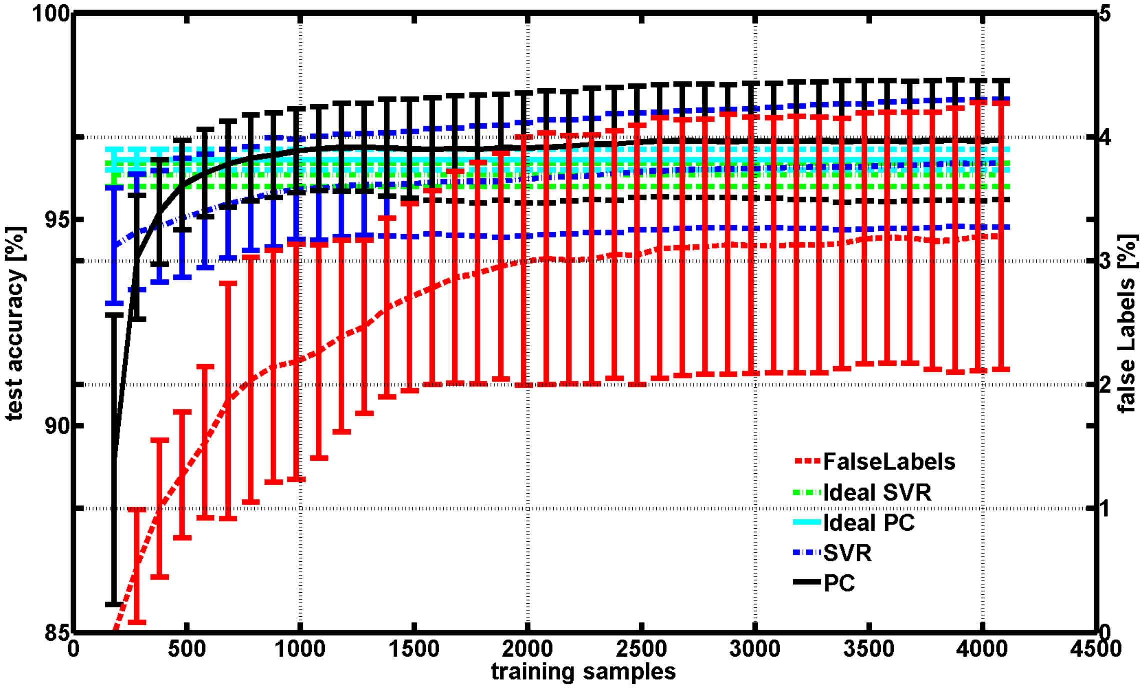
(b) Initial number of training samples: 2% of the training data.

(c) Initial number of training samples: 5% of the training data.

Figure 10: Semi-supervised learning results for the MNIST data set for different amounts of initial, manually labelled training data.



(a) Initial number of training samples: 0.3% of the training data.



(b) Initial number of training samples: 0.6% of the training data.

Figure 11: Semi-supervised learning results for the traffic sign data set for different amounts of initial, manually labelled training data.

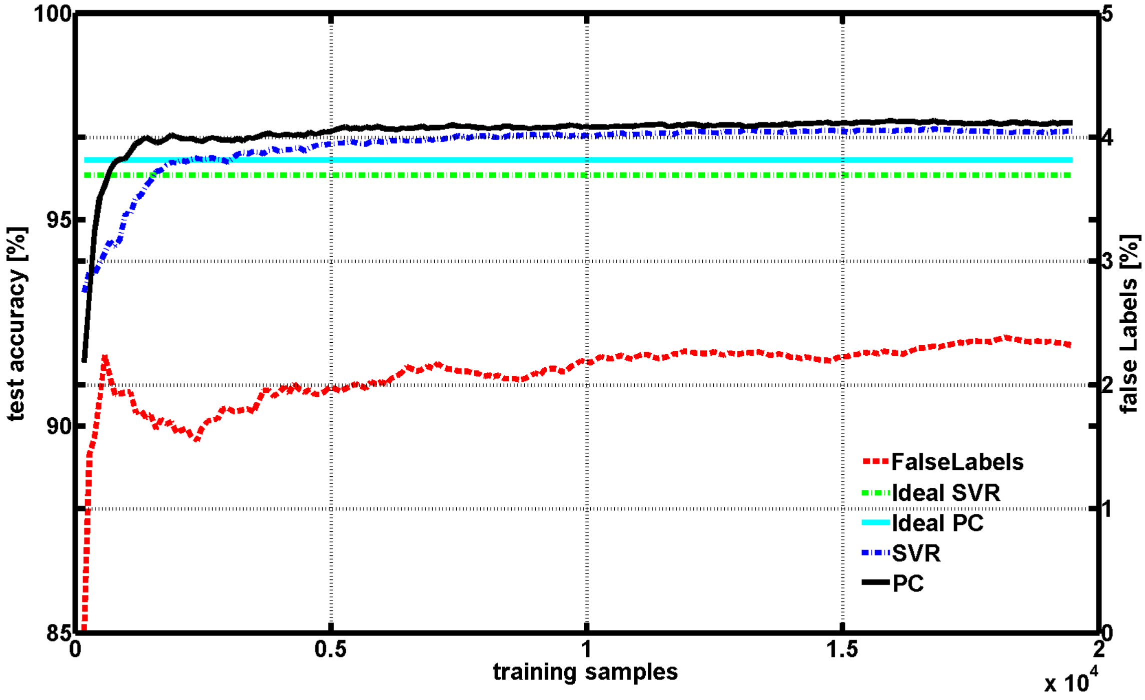


Figure 12: Semi-supervised learning results (single run) for the traffic sign data set. No confidence bands were taken into account ($\sigma_{\mathrm{B,min}} = 0$ and $\sigma_{\mathrm{B,max}} \to \infty$). The initial number of training samples amounts to 0.6% of the training data.

while the PC accuracy suffers from severe overfitting. By adding new labels the SVR ensures that most labels are correct, such that the PC generalises fairly well. In almost all runs the PC surpasses the SVR after a few samples are collected. Furthermore, the computational complexity of the PC remains the same with increasing size of the training set while the SVR becomes computationally increasingly complex due to the growing number of support vectors. The same behaviour is observed when the size of the initial training set is further decreased to 1% of the overall size of the training data set.

5.2 Results obtained for the traffic sign data set

Similar results are obtained for the traffic sign data set (cf. Fig. 11). Again, the learning curves denote the mean and the standard deviation over ten runs, respectively. If the size of the initial training set is too small, the gain in accuracy is also small while a large initial training set size yields no increase in performance during the semi-supervised learning. In Fig. 11b, the performance of the PC exceeds that of the SVR already after a small number of autonomously labelled samples are added to the training set. Notably, the accuracies of both classifiers averaged over the ten runs become very similar to the corresponding ideal accuracies of 96.08% for the SVR and 96.44% for the PC. For six of the ten runs, the accuracy obtained by semi-supervised learning exceeds the ideal accuracy of the PC (which is higher than that of the SVR), where the maximum achieved gain of the PC amounts to 1.98% accuracy. This behaviour is probably due to the presence of poorly imaged samples in the data set that affect the classification accuracy. Apparently, the semi-supervised learning algorithm achieves to avoid a selection of these poorly imaged training samples.

For comparison, we have regarded a semi-supervised learning scenario without taking into account the renormalised confidence bands, i.e. with $\sigma_{\mathrm{B,min}} = 0$ and $\sigma_{\mathrm{B,max}} \to \infty$. Hence, an unlabelled training sample is accepted if both classifiers yield the same class assignment. This approach is somewhat similar to the well-known co-training method [3]. However, rather than using two different feature representations of the data ("views") as in [3], we employ two different classifier architectures instead. The results for a single run performed for the traffic sign data set are shown in Fig. 12. The achieved recognition accuracy is comparable to the result obtained based on renormalised confidence bands with the same number of initial manually labelled training samples (cf. Fig. 11b). However, the semi-supervised learning process in which the renormalised confidence band values are taken into account is autonomously terminated much earlier. As a consequence, the training set for the scenario without confidence bands (cf. Fig. 12) becomes very large, such that the number of support vectors of the SVR (and therefore its computational complexity) is more than an order of magnitude higher than for the scenario of Fig. 11b in which renormalised confidence band values are used. Neglecting confidence band information results in an extended "saturation period" (cf. Fig. 12), where early stopping would only be possible using a manually labelled validation set.

Hence, the confidence band values provide a favourable early-stopping criterion for the semi-supervised learning process without the need for a manually labelled validation set. This property is of high importance in the examined scenario since in this setting manually labelled samples are regarded as highly expensive.

6 Summary and conclusion

In this study, we have described a sample selection mechanism based on analytically tractable confidence measures which are inferred for an "ensemble" of two different classifiers (PC and SVR) in the context of semi-supervised learning. We have compared different confidence measures for unlabelled training samples based on the MNIST handwritten digits data set and a traffic sign data set, where we have introduced a renormalised confidence band measure as a criterion for accepting or rejecting the autonomously generated labels for new samples.

As a general result, our experimental evaluation of semi-supervised learning has shown that for a small initial number of training samples, the computationally highly complex and memory-demanding SVR approach is able to increase autonomously the initially very poor recognition performance of the computationally much more efficient PC to a level that comes close to the "ideal" performance resulting from a training process in which all available training samples are used with their correct labels. For the traffic sign data set, the semi-supervised learning procedure may even result in classifier accuracies which exceed the ideal accuracies of the SVR and the PC, which indicates that the proposed algorithm selects those samples which improve the recognition behaviour in an advantageous manner. In particular, it is demonstrated that the confidence band values provide a favourable early-stopping criterion for the semi-supervised learning process without the need for a manually labelled validation set.

Future work will address semi-supervised ensemble learning based on a more "multi-faceted" classifier ensemble containing a larger number of different classifier architectures and a variety of feature representations of the data.

References

[1] Akima, H.: *A New Method of Interpolation and Smooth Curve Fitting Based on Local Procedures*. Journal of the ACM (JACM) 17(4): 589-602. 1970.

[2] Begleiter, R.; El-Yaniv, R.; Pechyony, D.: *Repairing Self-Confident Active-Transductive Learners Using Systematic Exploration*. Pattern Recognition Letters 29(9): 1245-1251. 2008.

[3] Blum, A.; Mitchell, T.: Combining labeled and unlabeled data with co-training. In: *Proc. Workshop on Computational Learning Theory*, pp. 92-100. 1998.

[4] Chang, C.-Cc; Lin, C.-J.: *LIBSVM: A library for support vector machines*. ACM Transactions on Intelligent Systems and Technology 2(3): 27:1-27:27. 2011.

[5] Chapelle, O.; Schölkopf, B.; Zien, A. (Hg.): *Semi-Supervised Learning (Adaptive Computation and Machine Learning)*. The MIT Press. 2006.

[6] Härdle, W. K.; Ritov, Y.; Song, S.: Partial Linear Quantile Regression and Bootstrap Confidence Bands. SFB 649 Discussion Papers SFB649DP2010-002, Sonderforschungsbereich 649, Humboldt University, Berlin, Germany. 2010.

[7] Hillebrand, M.; Wöhler, C.; Krüger, L.; Kreßel, U.; Kummert, F.: Self-learning with confidence bands. In: *Proc. Workshop Computational Intelligence*, pp. 302-313. 2010.

[8] Kardaun, O. J. W. F.: *Classical Methods of Statistics*. Springer. 2005.

[9] Kendall, W. S.; Marin, J.-M.; Robert, C. P.: Confidence bands for Brownian motion and applications to Monte Carlo simulation. *Statistics and Computing* 17 (2007) 1, S. 1–10.

[10] Kreßel, U.; Lindner, F.; Wöhler, C.; Linz, A.: Hypothesis verification based on classification at unequal error rates. In: *Proc. Int. Conf. on Artificial Neural Networks*, pp. 874-879, Edinburgh. 1999.

[11] Martos, A.; Krüger, L.; Wöhler, C.: Towards Real Time Camera Self Calibration: Significance and Active Selection. In: *3DPVT '10: Proc. of the 4th Int. Symp. on 3D Data Processing, Visualization and Transmission, Paris, France, 2010.*

[12] Schürmann, J.: *Pattern classification: a unified view of statistical and neural approaches*. John Wiley & Sons, Inc. 1996.

[13] Seeger, M.: Learning with Labeled and Unlabeled Data. Technical report. 2001.

[14] Smola, A. J.; Schölkopf, B.: *A Tutorial on Support Vector Regression*. Statistics and Computing 14(3): 199-222. 2004.

[15] Zhu, X.; Goldberg, A.: *Introduction to Semi-Supervised Learning.* Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool Publishers. 2009.