One-Class Classification with Gaussian Processes

Michael Kemmler*, Erik Rodner*, and Joachim Denzler

Chair for Computer Vision
Friedrich Schiller University of Jena, Germany
{Michael.Kemmler,Erik.Rodner,Joachim.Denzler}@uni-jena.de
http://www.inf-cv.uni-jena.de

Abstract. Detecting instances of unknown categories is an important task for a multitude of problems such as object recognition, event detection, and defect localization. This paper investigates the use of Gaussian process (GP) priors for this area of research. Focusing on the task of one-class classification for visual object recognition, we analyze different measures derived from GP regression and approximate GP classification. Experiments are performed using a large set of categories and different image kernel functions. Our findings show that the well-known Support Vector Data Description is significantly outperformed by at least two GP measures which indicates high potential of Gaussian processes for one-class classification.

1 Introduction

Many applications have to deal with a large set of images from a single class (positive examples) and only few or zero learning examples from a counter class (negative examples). Learning a classifier in such situations is known as one-class classification (OCC), novelty detection, outlier detection, or density estimation. The latter concentrates on estimating a well-defined probability distribution from the data, whereas OCC is more related to binary classification problems without the need for normalization.

Application scenarios often arise due to the difficulty of obtaining training examples for rare cases, such as images of defects in defect localization tasks [1] or image data from non-healthy patients in medical applications [2]. In these cases, one-class classification (OCC) offers to describe the distribution of positive examples and to treat negative examples as outliers which can be detected without explicitly learning their appearance. Another motivation to use OCC is the difficulty of describing a background or counter class. This problem of finding an appropriate unbiased set of representatives exists in the area of object detection or content based image retrieval [3, 4].

Earlier work concentrates on density estimation with parametric generative models such as single normal distributions or Gaussian mixture models [5]. These methods often make assumptions about the nature of the underlying distribution. Nowadays, kernel methods like one-class Support Vector Machines [6]

* The first two authors contributed equally to this paper.

(also called 1-SVM) or the highly related Support Vector Data Description (SVDD [7]), offer to circumvent such assumptions in the original space of feature vectors using the kernel trick. Those methods inherit provable generalization properties from learning theory [6] and can handle even infinite dimensional feature spaces.

In this paper we propose several approaches to OCC with Gaussian process (GP) priors. Machine learning with GP priors allows to formulate kernel based learning in a Bayesian framework [8] and has proven to be competitive with SVM-based classifiers for binary and multi-class categorization of images [9]. Nevertheless, their use for OCC scenarios has mostly been studied in the case of proper density estimation [10], which requires sophisticated MCMC techniques to obtain a properly normalized density. Alternatively, Kim et al. [11] present a clustering approach which indirectly uses a GP prior and the predictive variance.

Our new variants for OCC with GP priors include the idea of [11] as a special case. We further investigate the suitability of approximate inference methods for GP classification, such as Laplace approximation (LA) or expectation propagation (EP) [8]. We additionally show how to apply our GP-based and a common SVM-based method to the task of object recognition by utilizing kernel functions based on pyramid matching [12].

The remainder of the paper is structured as follows. First, we briefly review classification with GP priors. Our approach to one-class classification with GP is presented in Sect. 3. Section 4 describes Support Vector Data Description, a baseline method used for comparison. Kernel-based OCC for visual object recognition, as explained in Sect. 5, is utilized to evaluate and compare all methods and their variants in Sect. 6. A summary of our findings and a discussion of future research directions conclude the paper.

2 Classification via Gaussian Processes

This section gives a brief introduction to GP classification. Since classification is motivated from non-parametric Bayesian regression, we first briefly introduce the regression case with real-valued outputs $y \in \mathbb{R}$, before we discuss approximate methods for GP classification with binary labels $y \in \{-1, 1\}$.

2.1 The Regression Case

The regression problem aims at finding a mapping from input space $\mathcal{X}$ to output space $\mathbb{R}$ using labeled training data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathcal{X}^n$, $\mathbf{y} = [y_1, \ldots, y_n]^T \in \mathbb{R}^n$. In the following it is assumed that an output y is generated by a latent function $f : \mathcal{X} \rightarrow \mathbb{R}$ and additive noise ε, i.e. $y = f(\mathbf{x}) + \varepsilon$. Rather than restricting f to a certain parametric family of functions, we only assume that the function is drawn from a specific probability distribution $p(\mathbf{f}|\mathbf{X})$. This allows for a Bayesian treatment of our problem, i.e. we infer the probability of output y_* given a new input x_* and old observations $\mathbf{X}, \mathbf{y}$ by integrating out the corresponding

non-observed function values $f_* = f(\mathbf{x}_*)$ and $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^T$:

$$p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = \int p(f_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)\ p(y_*|f_*)\ df_* \tag{1}$$

$$p(f_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*) = \int p(f_*|\mathbf{X}, \mathbf{f}, \mathbf{x}_*)\ p(\mathbf{f}|\mathbf{X}, \mathbf{y})\ d\mathbf{f} \tag{2}$$

The central assumption in GP regression is a GP prior over latent functions, i.e. all function values are jointly normally distributed:

$$\mathbf{f}|\mathbf{X} \sim \mathcal{N}(m(\mathbf{X}), \kappa(\mathbf{X}, \mathbf{X})) \tag{3}$$

This distribution is solely specified by the mean function $m(\cdot)$ and covariance function $\kappa(\cdot, \cdot)$. If we further assume that the additive noise ε is modelled by a zero-mean Gaussian distribution, i.e. $p(y|f) = \mathcal{N}(f, \sigma_n^2)$, we are able to solve the integrals in closed form. Using a zero-mean GP, the predictive distribution (2) is again Gaussian [8] with moments

$$\mu_* = \mathbf{k}_*^T \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{y} \tag{4}$$

$$\sigma_*^2 = k_{**} - \mathbf{k}_*^T \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{k}_* \tag{5}$$

using abbreviations $\mathbf{K} = \kappa(\mathbf{X}, \mathbf{X})$, $\mathbf{k}_* = \kappa(\mathbf{X}, \mathbf{x}_*)$, and $k_{**} = \kappa(\mathbf{x}_*, \mathbf{x}_*)$. Since we assume i.i.d. Gaussian noise, this also implies that (1) is normally distributed with mean μ_* and variance $\sigma_*^2 + \sigma_n^2$.

2.2 From Regression to Classification

The goal in binary GP classification is to model a function which predicts a confidence for each class $y \in \{-1, 1\}$, given a feature vector $\mathbf{x}$. In order to make probabilistic inference about the output given a training set, we can directly apply the Bayesian formalism from eq. (1) and (2). However, the key problem is that the assumption of Gaussian noise no longer holds, since the output space is discrete. We could either ignore this issue and perform regression on our labels, or we could use a more appropriate likelihood such as the cumulative Gaussian

$$p(y|f) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{yf} \exp\left(-\frac{1}{2}x^2\right)\ dx \tag{6}$$

The disadvantage of the latter procedure is that our predictive distribution (2) is no longer a normal distribution. To overcome this issue, we follow the standard approach to approximate the posterior $p(\mathbf{f}|\mathbf{X}, \mathbf{y})$ with a normal distribution $\hat{p}(\mathbf{f}|\mathbf{X}, \mathbf{y})$. Two well-known approaches, which are also used in this work, are Laplace approximation (LA) and Expectation Propagation (EP). The interested reader is referred to [8].

For the final prediction step, approximations $\hat{p}(\mathbf{f}|\mathbf{X}, \mathbf{y})$ are used to solve (1). Using both Gaussian approximations to the posterior (2) and cumulative Gaussian likelihoods $p(y|f)$, it can be shown that the predictive distribution (1) is also equal to a cumulative Gaussian and can thus be evaluated in closed form [8].

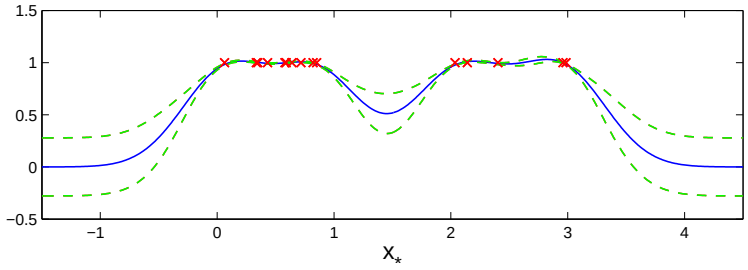

Fig. 1. GP regression using a zero-mean GP prior in an one-dimensional OCC setting. The predictive distribution is visualized via its mean and corresponding confidence interval (scaled variances), where training points are marked as crosses.

3 One-Class Classification with Gaussian Process Priors

GP approaches from Sect. 2 are discriminative techniques to tackle the problem of classification [8]. This follows from equation (1) where the conditional density $p(y_*|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$ is modeled. Using discriminative classification techniques directly for one-class classification is a non-trivial task, due to the fact that the density of the input data is not taken into account.

Nevertheless, utilizing a properly chosen Gaussian process prior enables us to derive useful membership scores for OCC. The main idea is to use a mean of the prior with a smaller value than our positive class labels (e.g. $y = 1$), such as a zero mean. This restricts the space of probable latent functions to functions with values gradually decreasing when being far away from observed points. In combination with choosing a smooth covariance function, an important subset of latent functions is obtained which can be employed for OCC (c.f. Fig. 1).

This highlights that the predictive probability $p(y_* = 1|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$ can be utilized, in spite of being a discriminative model. Due to the fact that the predictive probability is solely described by its first and second order moments, it is natural to also investigate the power of predictive mean and variance as alternative membership scores. Their suitability is illustrated in Fig. 1: The mean decreases for inputs distant from the training data and can be directly utilized as an OCC measure. In contrast, the predictive variance σ_*^2 is increasing which suggests that the negative variance value can serve as an alternative criterion for OCC. The latter concept is used in the context of clustering by Kim et al. [11]. Additionally, Kapoor et al. [9] propose the predictive mean divided by the standard devia-

Table 1. Different measures derived from the predictive distribution which are suitable for OCC membership scores.

Mean (M)	$\mu_* = \mathcal{E}(y_*\|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$	Probability (P)	$p(y_* = 1\|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$
neg. Variance (V)	$-\sigma_*^2 = -\mathcal{V}(y_*\|\mathbf{X}, \mathbf{y}, \mathbf{x}_*)$	Heuristic (H)	$\mu_* \cdot \sigma_*^{-1}$

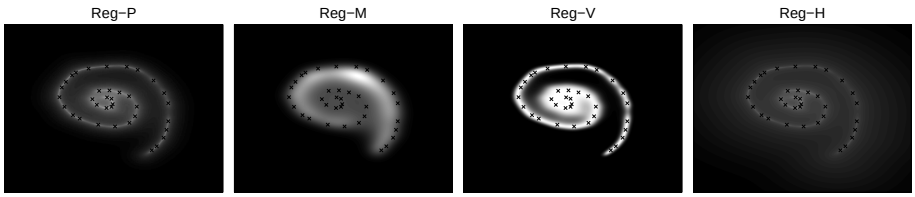

Fig. 2. One-class classification using GP regression (Reg) and measures listed in Table 1. All measures capture the distribution quite well.

tion as a combined measure for describing the uncertainty of the estimation and applied this heuristic successfully in the field of active learning.

All variants, which are summarized in Table 1, are available for GP regression and approximate GP classification with Laplace approximation or Expectation Propagation. The different membership scores produced by the proposed measures are visualized in Fig. 2 using an artificial two-dimensional example.

In the following sections we additionally motivate the use of the mean and variance of GP regression by highlighting the strong relationship to Parzen estimation and normal density distributions respectively.

3.1 Predictive Mean Generalizes Parzen Estimators

There exist various techniques which aim at constructing a proper probability density of observed data. Kernel density estimation, also known as Parzen estimation, is one strategy to achieve this goal. This non-parametric method constructs a probability density by superimposing similarity functions (kernels) on observed data points. Our one-class classification approach using the mean of GP regression has a tight relationship to Parzen estimators. If we assume noise-free observations ($\sigma_n = 0$) and no correlations between training examples ($\mathbf{K} = \mathbf{I}$), the regression mean (4) simplifies to

$$\mu_* = \mathbf{k}_*^T \mathbf{K}^{-1} \mathbf{y} \propto \frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} \left[\mathbf{K}^{-1}\right]_{ij} \cdot \kappa(\mathbf{x}_*, \mathbf{x}_i) = \frac{1}{n} \sum_{i=1}^{n} \kappa(\mathbf{x}_*, \mathbf{x}_i) \tag{7}$$

Please note that by construction $\mathbf{y} = (1, \ldots, 1)^T$. The proposed GP regression mean can hence be seen as an unnormalized Parzen density estimation technique using scaling matrix $\mathbf{K}$, which is implicitly given by covariance function $\kappa(\cdot, \cdot)$.

3.2 Predictive Variance Models a Gaussian in Feature Space

One approach to describe the data is to estimate a normal distribution in feature space $\mathcal{H}$ induced via a mapping $\Phi : \mathcal{X} \to \mathcal{H}$. It has been shown by Pękalska et al. [13] that computing the variance term in GP regression is equal to the

Mahalanobis distance (to the data mean in feature space) if the regularized ($\sigma_n > 0$) kernel-induced scaling matrix $\boldsymbol{\Sigma} = \widetilde{\Phi}(\mathbf{X})\widetilde{\Phi}(\mathbf{X})^T + \sigma_n^2\mathbf{I}$ is used:

$$\widetilde{\Phi}(\mathbf{x})^T \boldsymbol{\Sigma}^{-1} \widetilde{\Phi}(\mathbf{x}) \;\propto\; \widetilde{\kappa}(\mathbf{x},\mathbf{x}) - \widetilde{\kappa}(\mathbf{x},\mathbf{X}) \left(\widetilde{\kappa}(\mathbf{X},\mathbf{X}) + \sigma_n^2\mathbf{I}\right)^{-1} \widetilde{\kappa}(\mathbf{X},\mathbf{x}) \tag{8}$$

where the tilde indicates operations on zero-mean normalized data, i.e. $\widetilde{\Phi}(\mathbf{x}) = \Phi(\mathbf{x}) - n^{-1}\sum_{i=1}^{n}\Phi(\mathbf{x}_i)$ and $\widetilde{\kappa}(\mathbf{x},\mathbf{x}') = \widetilde{\Phi}(\mathbf{x})^T\widetilde{\Phi}(\mathbf{x}')$. Since the GP variance argument from (5) does not utilize centered kernel matrices, we effectively use the logarithm of the unnormalized zero-mean Gaussian which best describes the data. For the case of constant $\kappa(\mathbf{x},\mathbf{x})$, we are softly "slicing" a hyperellipsoid from the data points which are distributed onto a sphere in feature space.

4 Support Vector Data Description as Baseline Method

In the following section, we present one-class classification with Support Vector Data Description (SVDD) [14] which will be used for comparison due to its common use for OCC applications.

The main idea of this approach is to estimate the smallest hypersphere given by center $\mathbf{c}$ and radius R which encloses our training data $\mathbf{X}$. This problem can be tackled in various ways [15], but it is most conveniently expressed using the quadratic programming framework [14]. In addition to assuming that all training points reside within the hypersphere, one can also account for outliers using a soft variant which allows some points $\mathbf{x}_i$ to be a (squared) distance ξ_i away from the hypersphere. Operating in feature space induced via a mapping $\Phi : \mathcal{X} \rightarrow \mathcal{H}$ to some Hilbert space $\mathcal{H}$, this objective can be solved by the following optimization problem:

$$\min_{\mathbf{c},R,\xi} \; R^2 + \frac{1}{\nu n}\sum_{i\in\mathcal{I}}\xi_i \;\; \text{subject to} \;\; ||\Phi(\mathbf{x}_i) - \mathbf{c}||_2^2 \leq R^2 + \xi_i,\; \xi_i \geq 0 \tag{9}$$

for all $i \in \mathcal{I} := \{1,\ldots,n\}$. The parameter $\nu \in (0,1]$ controls the impact that hypersphere outliers have on the optimization objective, e.g. using small values for ν leads to a strongly regularized objective and thus promotes solutions where few data points lie outside the hypersphere. It should be noted that the Euclidean distance in equation (9) can be generalized to a Bregmanian distances [16].

The primal problem can be transferred to the dual optimization problem which is solely expressed in terms of inner products $\kappa(\mathbf{x},\mathbf{x}') = \Phi(\mathbf{x})^T\Phi(\mathbf{x}')$ and thus enables the use of the kernel trick:

$$\max_{\boldsymbol{\alpha}} \; \operatorname{diag}(\mathbf{K})^T\boldsymbol{\alpha} - \boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha} \tag{10}$$

$$\text{subject to} \;\; 0 \leq \alpha_i \leq \frac{1}{\nu n} \;\; \text{for } i \in \mathcal{I} \;\; \text{and} \;\; \sum_{i=1}^{n}\alpha_i = 1 \tag{11}$$

where the center of the ball can be expressed as $\mathbf{c} = \sum_{i=1}^{n}\alpha_i\Phi(\mathbf{x}_i)$ and the squared radius is given by $R^2 = \operatorname{argmax}_i ||\Phi(\mathbf{x}_i) - \mathbf{c}||_2^2$ taking into account all

indices $i \in \mathcal{I}$ satisfying $0 < \alpha_i < \frac{1}{\nu n}$. A newly observed point $\Phi(\mathbf{x}_*)$ thus lies within the hypersphere if $||\Phi(\mathbf{x}_*) - \mathbf{c}||_2^2 = \boldsymbol{\alpha}^T \mathbf{K} \boldsymbol{\alpha} - 2\mathbf{k}_*{}^T \boldsymbol{\alpha} + k_{**} \leq R^2$ holds, using abbreviations from Sect. 2.1.

It should be noted that there exists a tight relationship to Support Vector Machines. It can be shown that for kernels satisfying $\kappa(\mathbf{x}, \mathbf{x}) = const$ for all $\mathbf{x} \in \mathcal{X}$, the SVDD problem is equal to finding the hyperplane which separates the data from the origin with the largest margin (1-SVM [6]).

Other loss functions can also be integrated, e.g. quadratic loss instead of hinge loss which leads to Least-Squares Support Vector Machines (LS-SVM, [17]). It is interesting to note that LS-SVM with a zero bias term is equivalent to using the predictive mean estimated by GP regression. Our extension to one-class classification problems, therefore, directly corresponds to the work of Choi et al. [18] in this special case.

5 One-Class Classification for Visual Object Recognition

Our evaluation of one-class classification with GP priors is based on binary image categorization problems. In the following section we describe how to calculate an image-based kernel function (image kernel) with color features. In addition we also utilize the pyramid of oriented gradients (PHoG) kernel which is based on grayscale images only [19].

Bag-of-Features and Efficient Clustering The Bag-of-Features (BoF) approach [12] has been developed in the last years to one of the state-of-the-art feature extraction methods for image categorization problems [20]. Each image is represented as a set of local features which are clustered during learning. This clustering is utilized to compute histograms which can be used as single global image descriptors.

We extract Opponent-SIFT local features [20] for each image by dense sampling of feature points with a horizontal/vertical pixel spacing of 10 pixels. Clustering is often done by applying k-Means to a small subset of local features. In contrast, it was shown by Moosmann et al. [21] that a Random Forest can be utilized as a very fast clustering technique. The set of local features is iteratively split using mutual information and class labels. Finally, the leafs of the forest represent clusters which can be utilized for BoF. We extend this technique to cluster local features of a single class by selecting a completely random feature and its median value in each inner node to perform randomized clustering.

Spatial Pyramid Matching Given global image descriptors, one could easily apply kernel functions such as radial-basis function (RBF) or polynomial kernels. The disadvantage of this method is that these kernels do not use the position of local features as an additional cue for the presence of an object. The work of Lazebnik et al. [12] shows that incorporating coarse, absolute spatial information with the spatial pyramid matching kernel (S-PMK) is advantageous for image categorization. To measure the similarity between two images with S-PMK, the image is divided recursively into different cells (e.g. by applying a 2×2 grid). As

Table 2. Mean AUC Performance of OCC methods, averaged over all 101 classes. Bold font is used when all remaining measures are significantly outperformed. GP measures significantly superior to SVDD_ν (with optimal ν) are denoted in italic font.

	Reg-P	Reg-M	Reg-V	Reg-H	LA-P	EP-P
PHoG	*0.696*	0.693	0.692	*0.696*	0.684	0.683
Color	*0.761*	0.736	**0.766**	*0.755*	*0.748*	*0.747*
	LA-M	EP-M	LA-V	EP-V	$\text{SVDD}_{0.5}$	$\text{SVDD}_{0.9}$
PHoG	0.684	0.683	0.686	0.685	0.690	0.685
Color	0.745	0.744	*0.758*	*0.757*	0.739	0.746

in [12], BoF histograms are calculated using the local features in each cell and combined with weighted histogram intersection kernels.

6 Experiments

In this section we analyze the proposed approach and its variants empirically which results in the following main outcomes:

1. OCC with the variance criterion estimated by GP regression (Reg-V) is significantly better than all other methods using the color image kernels and it outperforms SVDD for various values of the outlier ratio ν (Sect. 6.1).
2. Approximate GP classification with LA and EP does not lead to a better OCC performance (Sect. 6.1).
3. The performance of the mean of GP regression (Reg-M) varies dramatically for different categories and can even decrease with an increasing amount of training data (Sect. 6.2).
4. Parameterized image kernels offer additional performance boosts with the disadvantage of additional parameter tuning (Sect. 6.3).
5. OCC with image kernels is able to learn the appearance of specific subcategories (Sect. 6.4).

We perform experiments with all 101 object categories of the Caltech 101 database [22]. As performance measure we use the area under the ROC curve (AUC) which is estimated by 50 random splits in training and testing data. In each case a specific number of images from a selected object category is used for training. Testing data consists of the remaining images from the category and all images of the Caltech background category.

6.1 Evaluation of One-Class Classification Methods

To assess the OCC performance, we use 15 randomly chosen examples for training. Instead of performing a detailed analysis of each single category, we average the AUC over all classes and random repetitions to yield a final performance summary for each OCC method. Based on this performance assessment scheme,

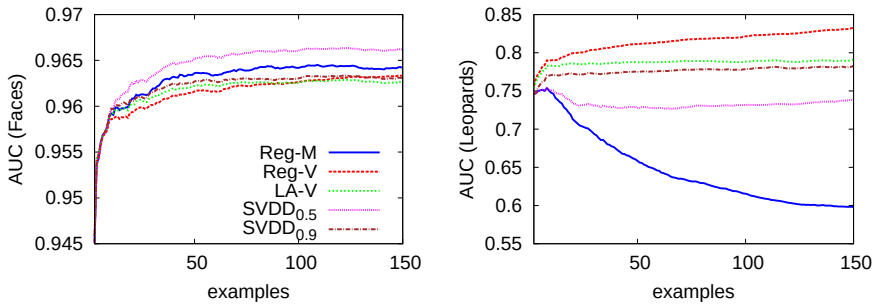

Fig. 3. Results for color feature based image kernels regarding classes *Faces* (Left) and *Leopards* (Right) with varying number of training examples (using same legend).

we compare predictive probability (-P), mean (-M) and variance (-V) of GP regression (Reg) and GP classification using Laplace Approximation (LA) or Expectation Propagation (EP), respectively. We additionally analyze the heuristic $\mu_* \cdot \sigma_*^{-1}$ for GP regression (Reg-H) and compare with SVDD using outlier fraction $\nu \in \{0.1, 0.2, \ldots, 0.9\}$ (SVDD$_\nu$). The results for PHoG and color features are displayed in Table 2, which, for the sake of readability, lists only best performing SVDD measures.

It can be immediately seen that PHoG features are significantly inferior to color features. Therefore, experiments in subsequent sections only deal with color-based image kernels. Although the average performance of all measures are quite similar, SVDD is significantly outperformed for all tested ν (t-test, $p \leq 0.025$) by at least two GP measures. The method of choice for our task is GP regression variance (Reg-V) which significantly outperforms all other methods using color features. Employing PHoG based image kernels, Reg-V also achieves at least comparable performance to SVDD for any tested parameter ν.

Our results also highlight that making inference with cumulative Gaussian likelihoods does not generally improve OCC, since LA and EP are consistently outperformed by GP regression measures Reg-V and Reg-P. Hence, the proposed OCC measures do not benefit from the noise model of (6) (and corresponding approximations) that are more suitable for classification in a theoretical sense.

6.2 Performance with an Increasing Number of Training Examples

To obtain an asymptotic performance behavior of all outlier detection methods, we repeat the experiments of Sect. 6.1 with a densely varying number of training examples. As can be seen in Fig. 3, the performance behavior highly depends on the class. Classifying *Faces*, the performance increases with a higher number of training examples in almost all cases. A totally different behavior, however, is observed for *Leopards*, where the averaged AUCs of Reg-M (and related Reg-P and Reg-H) substantially decrease when more than 8 training examples are used. For small ν, SVDD also exhibits this behavior in our experiments.

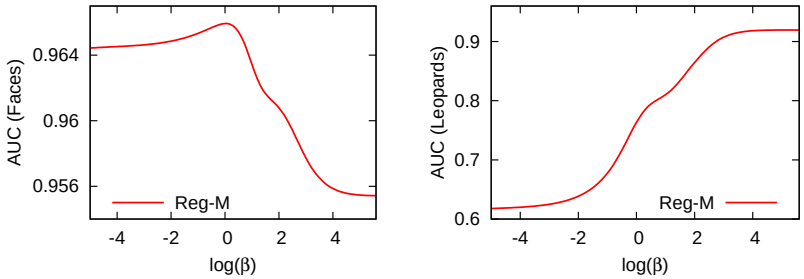

Fig. 4. Influence of an additional smoothness parameter β of a re-parameterized image kernel on the OCC performance for the categories *Faces* (Left) and *Leopards* (Right).

6.3 Influence of an Additional Smoothness Parameter

Estimating the correct smoothness of the predicted distribution is one of the major problems in one-class classification and density estimation. This smoothness is often controlled by a parameterized kernel, such as a RBF kernel. In contrast, our used kernel functions are not parameterized and the decreasing performance of the Reg-M method in the last experiment might be due to this inflexibility.

For further investigation, we parameterize our image kernel function by transforming it into a metric, which is then plugged into a distance substitution kernel [23]: $\kappa_\beta(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\beta\left(\kappa(\boldsymbol{x}, \boldsymbol{x}) - 2\kappa(\boldsymbol{x}, \boldsymbol{x}') + \kappa(\boldsymbol{x}', \boldsymbol{x}')\right)\right)$. We perform experiments with κ_β utilizing 100 training examples and a varying value of β. The results for the categories *Faces* and *Leopards* are plotted in Fig. 4.

Let us first have a look on the right plot and the results for *Leopards*. With a small value of β, the performance is comparable to the unparameterized version (cf. Fig. 3, right side). However, by increasing the parameter we achieve a performance above 0.9 and superior to other methods, such as Reg-V. This behavior differs significantly from the influence of β on the performance of the task *Faces*, which decreases after a small maximum. Right after the displayed points, we ran into severe numerical problems in both settings due to small kernel values below double precision. We expect a similar gain in performance by tuning the scale parameter of the cumulative Gaussian noise model, but we skip this investigation to future research. This analysis shows that introducing an additional smoothness hyperparameter offers a great potential, though efficient optimization using the training set is yet unsolved.

6.4 Qualitative Evaluation

Instead of separating a specific category from the provided background images of the Caltech database, we also apply our OCC methods to the difficult task of estimating a membership score of a specific sub-category. We therefore trained our GP methods and SVDD using 30 images of a type of chair called *windsor chair*, which has a characteristic wooden backrest. The performance is tested

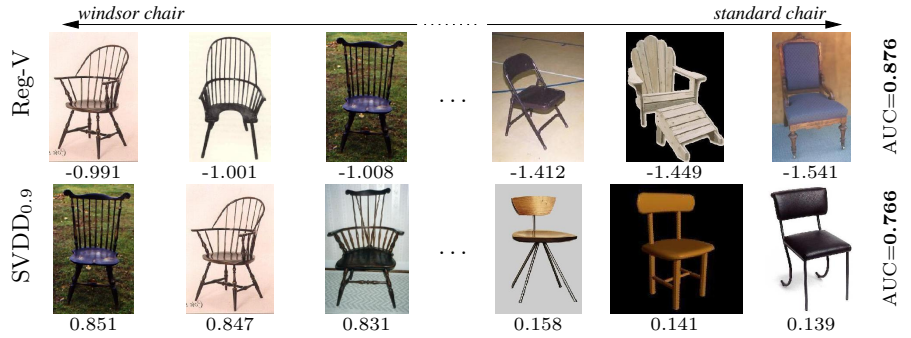

Fig. 5. Results obtained by training different OCC methods for *windsor chair* and separating against *chair*: the three best ranked images (all of them are characteristic windsor chairs) and the three last ranked images with corresponding output values.

on all remaining *windsor chairs* and images of the category *chairs*. Results are illustrated in Fig. 5 with the best and last ranked images. The qualitative results are similar, but the AUC values clearly show that Reg-V is superior.

7 Conclusions and Further Work

We present an approach for one-class classification (OCC) with Gaussian process (GP) priors and studied the suitability of different measures, such as mean and variance of the predictive distribution. The GP framework allows to use different approximation methods to handle the underlying classification problem such as regression, Laplace approximation and Expectation Propagation. All aspects are compared against the popular Support Vector Data Description (SVDD) approach of Tax and Duin [14] in a visual object recognition application using state-of-the-art image-kernels [12, 19]. It turns out that using the predictive variance of GP regression is the method of choice and, compared to SVDD, achieves significantly better results for object recognition tasks with color features. Using the estimated mean frequently leads to good results, but highly depends on the given classification task.

We also show that the introduction of an additional smoothness parameter can lead to a large performance gain. Thus, estimating correct hyperparameters of the re-parameterized kernel function would be an interesting topic for future research. Another idea worth investigating is the use of specifically tailored noise models for OCC. In general, our new approach to OCC is not restricted to object recognition and its application to tasks such as defect localization would be interesting.

References

1. Hodge, V., Austin, J.: A survey of outlier detection methodologies. Artificial Intelligence Review **22** (2004) 85–126

2. Tarassenko, L., Hayton, P., Cerneaz, N., Brady, M.: Novelty detection for the identification of masses in mammograms. In: Fourth International Conference on Artificial Neural Networks. (1995) 442–447
3. Chen, Y., Zhou, X., Huang, T.S.: One-class svm for learning in image retrieval. In: Proceedings of the IEEE Conference on Image Processing. (2001)
4. Lai, C., Tax, D.M.J., Duin, R.P.W., Pękalska, E., Paclík, P.: On combining one-class classifiers for image database retrieval. In: Proceedings of the Third International Workshop on Multiple Classifier Systems. (2002) 212–221
5. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). 1 edn. Springer (2007)
6. Schölkopf, B., Platt, J.C., Shawe-Taylor, J.C., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Comput. **13** (2001) 1443–1471
7. Tax, D.M.J., Duin, R.P.W.: Support vector data description. Mach. Learn. **54** (2004) 45–66
8. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press (2005)
9. Kapoor, A., Grauman, K., Urtasun, R., Darrell, T.: Gaussian processes for object categorization. International Journal of Computer Vision **88** (2010) 169–188
10. Adams, R.P., Murray, I., MacKay, D.: The gaussian process density sampler. In: NIPS. (2009)
11. Kim, H.C., Lee, J.: Pseudo-density estimation for clustering with gaussian processes. In: Advances in Neural Networks - ISNN. Volume 3971. (2006) 1238–1243
12. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the 2006 IEEE Conf. on Computer Vision and Pattern Recognition. (2006) 2169–2178
13. Pękalska, E., Haasdonk, B.: Kernel discriminant analysis for positive definite and indefinite kernels. IEEE Trans. Pattern Anal. Mach. Intell. **31** (2009) 1017–1032
14. Tax, D.M.J., Duin, R.P.W.: Support vector data description. Mach. Learn. **54** (2004) 45–66
15. Yildirim, E.A.: Two algorithms for the minimum enclosing ball problem. SIAM Journal on Optimization **19** (2008) 1368–1391
16. Crammer, K., Singer, Y.: Learning algorithms for enclosing points in bregmanian spheres. In: Learning Theory and Kernel Machines, Springer (2003) 388–402
17. Suykens, J.A.K., Gestel, T.V., Brabanter, J.D., Moor, B.D., Vandewalle, J.: Least Squares Support Vector Machines. World Scientific Pub. Co. (2002)
18. Choi, Y.S.: Least squares one-class support vector machine. Pattern Recogn. Lett. **30** (2009) 1236–1240
19. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: Proceedings of the 6th ACM international conference on Image and video retrieval. (2007) 401–408
20. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. (2010)
21. Moosmann, F., Triggs, B., Jurie, F.: Fast discriminative visual codebooks using randomized clustering forests. In: NIPS. (2006) 985–992
22. Fei-Fei, L., Fergus, R., Perona, P.: One-shot learning of object categories. IEEE Transactions on Pattern Analysis and Machine Intelligence **28** (2006) 594–611
23. Vedaldi, A., Soatto, S.: Relaxed matching kernels for object recognition. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition. (2008)