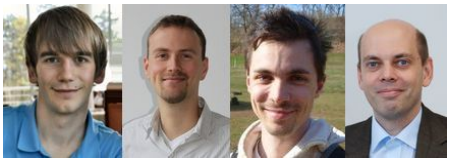
Divergence-Based One-Class Classification Using Gaussian Processes

Paul Bodesheim
paul.bodesheim@uni-jena.de
Erik Rodner
erik.rodner@uni-jena.de
Alexander Freytag
alexander.freytag@uni-jena.de
Joachim Denzler
joachim.denzler@uni-jena.de

Computer Vision Group
Friedrich Schiller University of Jena, Germany
http://www.inf-cv.uni-jena.de



Detecting samples of unknown classes is a key task for active learning [1] and **one-class classification (OCC)** [2]. Starting from a set of only positive training samples, we want to estimate a soft membership score for every new test sample. Applying OCC methods is especially beneficial in situations where either negative data is difficult to model with given samples or where negative samples are hard to obtain.

We present an information theoretic framework for OCC, which allows for deriving several new novelty scores. With these scores, we are able to rank samples according to their novelty and to detect outliers not belonging to a learnt data distribution. **Our new framework sheds light on OCC from a completely different theoretical perspective.** The key idea of our approach is to measure how strongly a new test sample would influence the current model if it was used for training. This is carried out in a probabilistic manner using Jensen-Shannon divergence and the Gaussian process (GP) regression framework. An overview of our presented approach, which is based on mutual information (MI) and divergence measures of information theory, can be seen in Figure 2. Although our formulation is strongly related to active learning [1], we only consider OCC [2] in the paper. In the following, we assume a given training set $\mathbf{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ with labels $\mathbf{y} = \mathbf{1} = (1, \ldots, 1)^T$ and estimate a membership score of a test sample $\mathbf{x}^*$. We score the sample based on the resulting model change after treating it as an additional training sample.

To evaluate the change of the model, we once assume that $\mathbf{x}^*$ belongs to the target class ($y^* = 1$) and once assume the opposite ($y^* = -1$). Since we have no precise knowledge about the correct label of new samples, we model the assumed label $y^* \in \{1, -1\}$ as a random variable. For reasons explained in the paper, we approximate the change of the whole model by relying on a neighborhood of infinite small size and only taking the new sample itself into account. Therefore, we introduce a second random variable $Y^* \in \{1, -1\}$ to evaluate the model on $\mathbf{x}^*$ after the model update, *i.e.*, the variable Y^* is the reclassification result of $\mathbf{x}^*$. The dependencies between the assumed label y^* and the reclassification result Y^* can be measured using the conditional MI:

$$\mathrm{I}(Y^*, y^* \mid \mathbf{D}^*) = \mathrm{H}(Y^* \mid \mathbf{D}^*) - \mathrm{H}(Y^* \mid y^*, \mathbf{D}^*) \tag{1}$$

that depends on the available data $\mathbf{D}^* = (\mathbf{X}, \mathbf{y}, \mathbf{x}^*)$, which contains the training set as well as the new test sample $\mathbf{x}^*$. A low conditional entropy $\mathrm{H}(Y^* \mid y^*, \mathbf{D}^*)$ indicates that the reclassification result Y^* is almost certain given the assumed label y^*. Since the reclassification of $\mathbf{x}^*$ and thus

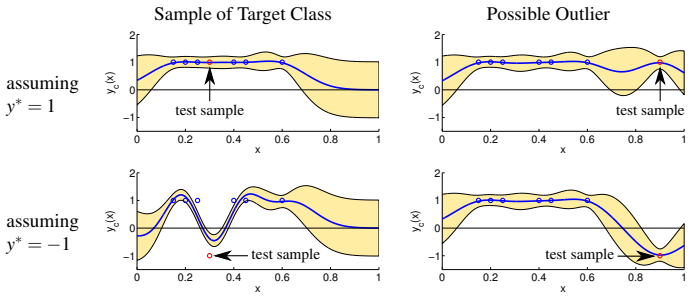


Figure 1: Visualization of our divergence approach. While both label assumptions $y^* \in \{1, -1\}$ of a possible outlier can be verified by reclassification using the model additionally trained with the test sample (blue curve), the assumption $y^* = -1$ will lead to a weak reclassification of a test sample stemming from the target class. Our approach exploits this difference. Classification uncertainty is visualised by shaded areas.

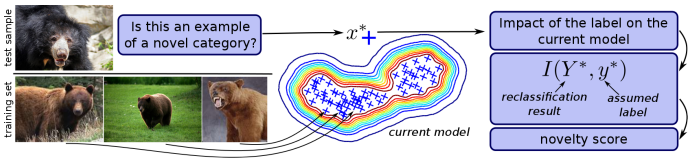


Figure 2: Outline of our approach based on mutual information.

the value of Y^* heavily depends on the training data, one achieves a low conditional entropy if the test sample $\mathbf{x}^*$ is far away from the training samples and the reclassification is mainly influenced by the choice of y^*. Summing up, a low conditional MI is induced by a strong membership to the target class and vice versa. The conditional MI of Eq. 1 turns out to be equal to the Jensen-Shannon (JS) divergence [3]:

$$\begin{aligned} \mathrm{I}(Y^*, y^* \mid \mathbf{D}^*) &= \mathrm{D}_{\mathrm{JS}}^{\pi}(\mathbf{p}_1 || \mathbf{p}_{-1}) && (2) \\ &= \pi \cdot \mathrm{D}_{\mathrm{KL}}(\mathbf{p}_1 || \mathbf{m}) + (1-\pi) \cdot \mathrm{D}_{\mathrm{KL}}(\mathbf{p}_{-1} || \mathbf{m}) \quad , && (3) \end{aligned}$$

where $\mathrm{D}_{\mathrm{KL}}(\cdot||\cdot)$ is the Kullback-Leibler (KL) divergence, $\mathbf{m} = \pi \cdot \mathbf{p}_1 + (1-\pi) \cdot \mathbf{p}_{-1}$ the mixture of the two conditional probability distributions $\mathbf{p}_1 = \mathrm{p}(Y^* \mid y^* = 1, \mathbf{D}^*)$ and $\mathbf{p}_{-1} = \mathrm{p}(Y^* \mid y^* = -1, \mathbf{D}^*)$, and π the prior probability: $\pi = \mathrm{p}(y^* = 1 \mid \mathbf{D}^*)$. Therefore, we propose using the **negative JS divergence as an OCC score**. For the computation, we only need the parameter π as well as conditional probability distributions $\mathbf{p}_1$, and $\mathbf{p}_{-1}$. In this paper, we propose using posterior probabilities of a GP.

In the case of GP regression, continuous outputs y_c are assumed to be generated according to $y_c(\mathbf{x}) = f(\mathbf{x}) + \varepsilon$, where f is a latent function and ε is a noise term. Following a Bayesian framework, output values of unknown samples $\mathbf{x}^*$ are predicted in a probabilistic fashion by marginalising over latent functions f. Using assumptions mentioned in the paper, the predictive output y_c^* for a new sample $\mathbf{x}^*$ given data $\mathbf{D}^*$ is normally distributed as well with moments μ_* and σ_*^2 given in closed form. We compute probabilities $\pi = \mathrm{p}(y^* = 1 \mid \mathbf{D}^*)$ based on these moments via:

$$\pi = \mathrm{p}(y^* = 1 \mid \mathbf{D}^*) = \mathrm{p}(y_c^* > 0 \mid \mathbf{D}^*) = \frac{1}{2} - \frac{1}{2}\,\mathrm{erf}\left(\frac{-\mu_*}{\sqrt{2\sigma_*^2}}\right) \quad , \tag{4}$$

where $\mathrm{erf}(\cdot)$ is the error function. We also need to compute probabilities of the conditional distributions $\mathbf{p}_1$ and $\mathbf{p}_{-1}$ in a similar vein to Eq. (4). The conditional probabilities will arise from a GP model learnt with $N+1$ training samples by treating the current test sample $\mathbf{x}^*$ and its assumed label y^* as training data as well. The assumption $y^* = 1$ is still an OCC setting whereas the assumption $y^* = -1$ leads to a highly imbalanced binary classification scenario. The difference between the behaviour of samples stemming from the target class and outliers is visualised in Figure 1 using a solid line for the predictive mean and shaded areas for the predictive variance of the GP model.

Evaluations on machine learning and image categorization datasets are described in the paper. Our conclusion is that we reach state-of-the-art performance while offering a completely new access to the challenging problem of one-class classification.

[1] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Gaussian processes for object categorization. *International Journal of Computer Vision*, 88(2):169–188, 2010.

[2] Michael Kemmler, Erik Rodner, and Joachim Denzler. One-class classification with gaussian processes. In *ACCV*, pages 489–500, 2010.

[3] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.