Divergence-Based One-Class Classification Using Gaussian Processes

since 1558

Paul Bodesheim, Erik Rodner, Alexander Freytag, Joachim Denzler

Computer Vision Group, Friedrich Schiller University Jena, Germany

{Paul.Bodesheim,Erik.Rodner,Alexander.Freytag,Joachim.Denzler}@uni-jena.de

http://www.inf-cv.uni-jena.de/

Friedrich Schiller University Jena

Computer Vision Group

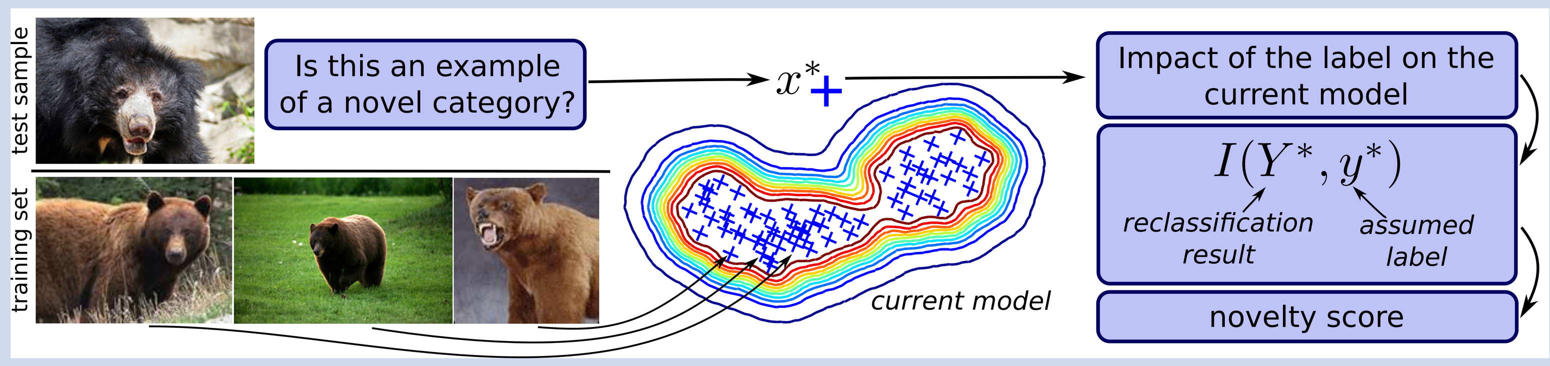
One-class classification (OCC)

- Given: a set of only positive training samples of a single class
- Goal: estimate a soft membership score for a test sample
- Why?: negative data is difficult to model or is hard to obtain

Aim of this work

Shed light on one-class classification from a completely different theoretical perspective

- Measure how strongly a new test sample would influence the current model if it was used for training
- Estimation of model change by comparing reclassification results
- Probabilistic framework based on information theory



Gaussian process regression [RW06]

- Continuous outputs y_c are assumed to be generated according to:

$$y_c(\mathbf{x}) = f(\mathbf{x}) + \varepsilon \qquad (f \text{ ... latent function}, \quad \varepsilon \text{ ... noise term})$$

- Output values of unknown samples $\mathbf{x}^*$ are predicted in a probabilistic fashion by marginalising over latent functions f
- Assumptions:
  1. Latent functions f are drawn from a Gaussian process prior with mean function being zero and covariance function $\kappa(\cdot,\cdot)$
  2. The noise term is normally distributed: $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$.

  $\Rightarrow$ Predictive output value y_c^* for a new sample $\mathbf{x}^*$ given the data $\mathbf{D}^*$ is normally distributed as well: $y_c^* \mid \mathbf{D}^* \sim \mathcal{N}(\mu_*, \sigma_*^2)$

$$\mu_* = \mathbf{k}_*^\mathrm{T} \left(\mathbf{K} + \sigma_n^2\, \mathbf{I}\right)^{-1} \mathbf{1}$$
$$\sigma_*^2 = \mathbf{k}_{**} - \mathbf{k}_*^\mathrm{T} \left(\mathbf{K} + \sigma_n^2\, \mathbf{I}\right)^{-1} \mathbf{k}_* + \sigma_n^2$$

- Label regression for OCC [KRD10]: GP-Mean, GP-Var, GP-Pred

Divergence-based one-class classification

- Assumed label of test sample: $y^* \in \{-1, 1\}$
- Reclassification result of test sample: $Y^* \in \{-1, 1\}$
- Influence of test sample on current model via **conditional mutual information**:

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

H ... Shannon entropy, $\mathbf{D}^* = (\mathbf{X}, \mathbf{y}, \mathbf{x}^*)$, $\mathbf{X}$ ... training samples, $\mathbf{y} = \mathbf{1}$ ... labels, $\mathbf{x}^*$ ... test sample

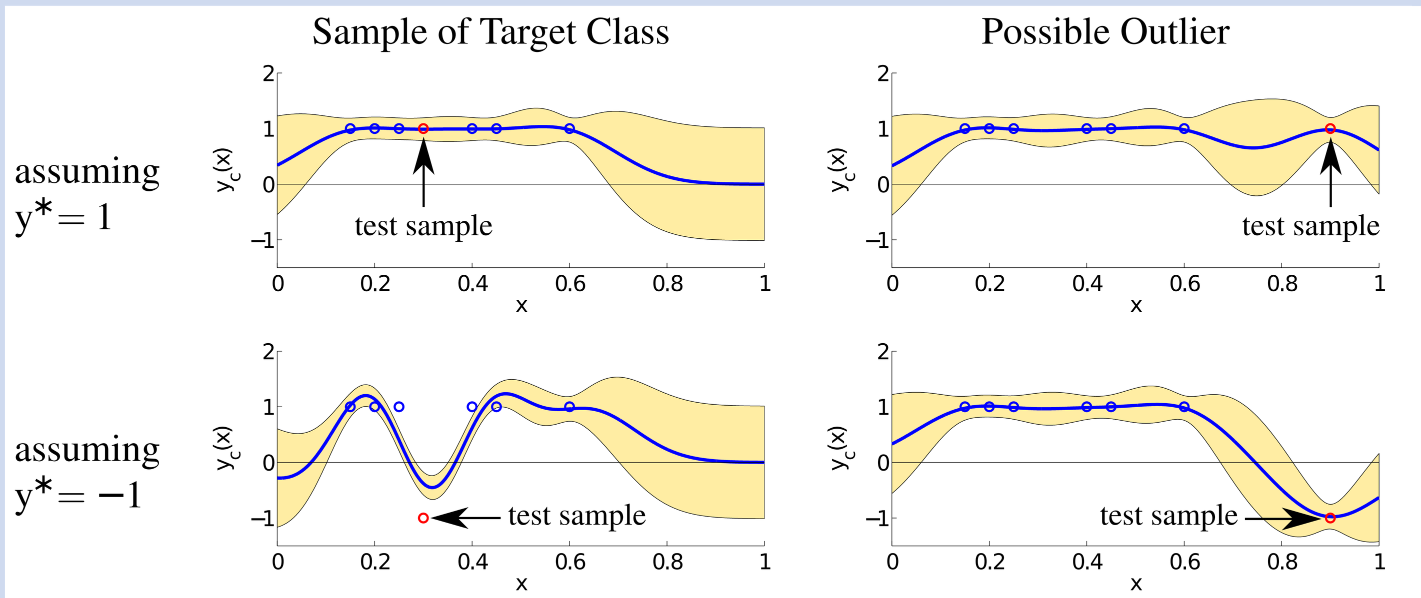
- Conditional mutual information turns out to be equal to the **Jensen-Shannon (JS) divergence**:

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

$\mathrm{D}_{\mathrm{KL}}(\cdot||\cdot)$ ... Kullback-Leibler divergence, $\mathbf{m} = \pi \cdot \mathbf{p}_1 + (1-\pi) \cdot \mathbf{p}_{-1}$ ... mixture of

$\mathbf{p}_1 = \mathrm{p}(Y^* \mid y^* = 1, \mathbf{D}^*)$ and $\mathbf{p}_{-1} = \mathrm{p}(Y^* \mid y^* = -1, \mathbf{D}^*)$ with prior $\pi = \mathrm{p}(y^* = 1 \mid \mathbf{D}^*)$

- Visualisation:



Gaussian process probabilities for divergence-based OCC

- Prior probabilities (new novelty measure **GP-Prob**):

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

- Conditional probabilities from reclassification:

$$\mathrm{p}(Y^* = 1 \mid y^*, \mathbf{D}^*) = \mathrm{p}\left(y_c^* > 0 \mid y^*, \mathbf{D}^*\right)$$
$$\mathrm{p}(Y^* = -1 \mid y^*, \mathbf{D}^*) = 1 - \mathrm{p}\left(y_c^* > 0 \mid y^*, \mathbf{D}^*\right)$$

- Balancing in reclassification via different noise levels is optional

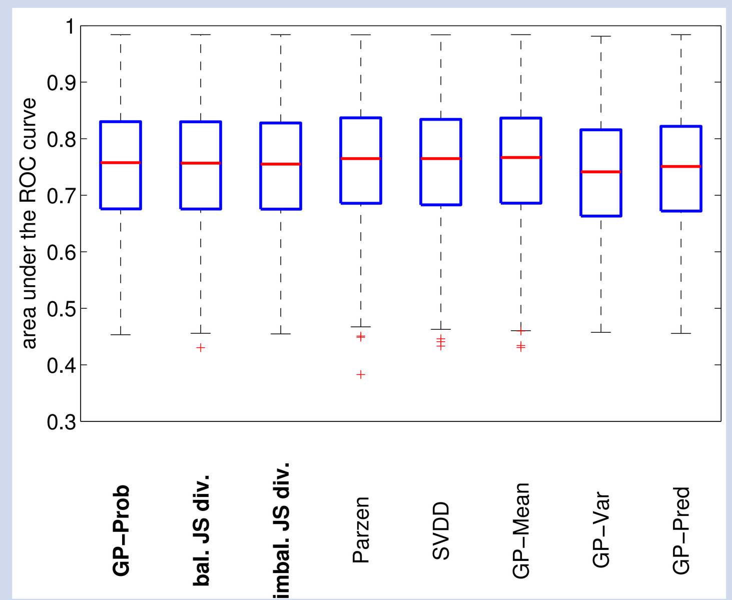
  $\Rightarrow$ **balanced or imbalanced JS divergence**

Experimental Results

- UCI datasets (best three results of each task are underlined)

OCC method	Median AUC of target class *Iris-Versicolour*	*Iris-Virginica*	*Sonar-Rocks*	*Sonar-Mines*
GP-Prob	0.981	0.966	0.625	0.772
bal. JS div.	0.981	0.967	0.618	0.768
imbal. JS div.	0.981	0.968	0.624	0.773
Parzen [Bis06]	0.973	0.960	0.602	0.771
SVDD [TD04]	0.986	0.971	0.609	0.761
GP-Mean [KRD10]	0.983	0.974	0.613	0.756
GP-Var [KRD10]	0.979	0.964	0.608	0.770
GP-Pred [KRD10]	0.980	0.968	0.618	0.776

- ImageNet (Visual Object Recognition)



OCC method	Median AUC (Std. dev.)	
GP-Prob	0.758	(±0.103)
bal. JS div.	0.757	(±0.103)
imbal. JS div.	0.755	(±0.103)
Parzen [Bis06]	0.765	(±0.105)
SVDD [TD04]	0.765	(±0.103)
GP-Mean [KRD10]	0.767	(±0.103)
GP-Var [KRD10]	0.741	(±0.104)
GP-Pred [KRD10]	0.751	(±0.103)

Conclusions

- New one-class classification framework based on information theory
- Gaussian process probabilities are suitable for this framework
- Results comparable to state-of-the-art

References

Bishop, Christopher M.: *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2006

Filippone, Maurizio ; Sanguinetti, Guido: Information theoretic novelty detection. In: *Pattern Recognition* 43 (2010), Nr. 3, S. 805–814

Kemmler, Michael ; Rodner, Erik ; Denzler, Joachim: One-Class Classification with Gaussian Processes. In: *ACCV*, 2010, S. 489–500

Rasmussen, Carl E. ; Williams, Christopher K. I.: *Gaussian Processes for Machine Learning*. The MIT Press, 2006

Tax, David M. J. ; Duin, Robert P. W.: Support Vector Data Description. In: *Machine Learning* 54 (2004), Nr. 1, S. 45–66