Automatic Identification of Novel Bacteria using Raman Spectroscopy and Gaussian Processes

Michael Kemmler[a], Erik Rodner[a], Petra Rösch[b], Jürgen Popp[b,c], Joachim Denzler[a]

[a] *Chair for Computer Vision*
Department of Mathematics and Computer Science
Friedrich Schiller University of Jena, Germany
[b] *Institute of Physical Chemistry and Abbe Center of Photonics*
Friedrich Schiller University of Jena, Germany
[c] *Institute of Photonic Technology, Jena, Germany*

Abstract

Raman spectroscopy is successfully used for the reliable classification of complex biological samples. Much effort concentrates on the accurate prediction of known categories for highly relevant tasks in a wide area of applications such as cancer detection and bacteria recognition. However, the resulting recognition systems cannot always be directly used in practice since unseen samples might not belong to classes present in the training set. Our work aims to tackle this problem of novelty detection using a recently proposed approach based on Gaussian processes. By learning novelty scores for a large bacteria Raman dataset comprising 50 different strains, we analyze the behavior of this method on an independent dataset which includes known as well as unknown categories. Our experiment reveals that Gaussian processes can be successfully applied to the task of finding unknown bacterial strains, leading to superior results compared to state-of-the-art methods such as Gaussian mixture models and Support Vector Data Description.

Keywords: Raman spectroscopy, Gaussian processes, novelty detection, one-class classification, bacteria recognition

1. Introduction

Over the last years, the identification of biological material such as microorganisms and tissue samples received a high amount of attention. A large set of techniques [1, 2] was developed to access as much information as possible from samples. Among these, vibrational techniques [3, 4, 5] such as Infrared [3] and Raman spectroscopy [6] showed to be especially suitable for many applications due to their non-destructive and fast methodology [7, 8]. Moreover, it has been shown that very reliable recognition results can be obtained without time-consuming cultivation steps, since it is possible to extract informative spectra from single cells [9, 10, 11, 12].

Due to these advantages, Raman spectroscopy is a widespread tool for a variety of tasks such as food control [8], cancer detection [13, 14, 15] and the identification of microorganisms [9, 11]. Equipped with powerful multivariate data analysis methods, accurate recognition rates can be achieved [11, 12, 16]. Apart from these encouraging results, the applicability of many recently proposed recognition systems to practical scenarios may be questioned. This is mainly due to the fact that experimental settings often implicitly require that test examples belong to (at least) one category already encountered in the training step (i.e. contained in the training database). Particularly for applications such as pathogen identification [17, 18, 19, 20] or general bacteria classification [11, 12, 21], this constraint can usually not be guaranteed a priori. Thus, additional data treatment is needed which justifies the assignment to one of a few known classes. Otherwise the classification decision is erroneous and can lead to a severely misleading analysis. This kind of outlier detection is also important if a classification system learns bacteria categories iteratively.

In the following work, Raman spectra obtained from single cells rather than bulk spectra are used since they enable a fast analysis and fit to the demands of practical applications. This setting requires sophisticated machine learning algorithms due to the higher variability of the data.

Related Work. In the literature, the task of detecting samples from previously unseen categories is known as novelty detection or one-class classification [22]. A popular approach for tackling this problem is to use generative models, where a probability distribution of the training data is modeled. For example Schmid et al. [11] use a generative Gaussian Mixture Model [23, 24] (GMM) for detecting novel bacterial Raman spectra. In contrast, Dundar et al. [25] propose to model each category with a single Gaussian distribution and to simulate a number of unknown categories by sampling from a Gaussian prior estimated from all known classes. This idea is related to the general idea of one-class classification by generating counter-examples [26, 27]. Their approach is applied to the recognition of bacterial cultures using optical-scattering methods [28] and tries to tackle the same basic problem as studied in this paper.

We recently introduced a non–parametric method [12] based on the Gaussian processes (GP) framework [29] which has shown to be applicable for a variety of one-class and novelty detection tasks in the area of visual object recognition [12]. The proposed method does not make any restrictive assumptions about the shape of the class distribution as assumed for Gaussian mixture models. This allows building flexible models for data with large intra-class variation, such as Raman spectra. Our method is highly related to the Support Vector Data Description [30, 31] which proved to be applicable for various one-class classification problems [32, 33, 34].

Our Contribution. This work aims to introduce GP-based novelty detection proposed in [12] for Raman spectroscopic applications. We additionally want to highlight the importance of taking into account previously unseen classes, a scenario prevalent in many real-world applications. By studying a bacteria recognition problem, we investigate the suitability of different novelty detection approaches: GMM and Parzen estimation [23], simulating outlier categories [25], SVDD [31] and GP-based methods [12]. A comparison with established approaches reveals that novelty detection based on GP regression can be successfully applied in our setting.

Outline. The paper is structured as follows. The theory of methods used for generating novelty scores is shortly explained in Sect. 2. The experimental setup as well as implementation details are provided in Sect. 4, while results are discussed in Sect. 5. A summary of our findings and directions for future work conclude the paper.

2. Novelty Detection with Gaussian processes

The following section concentrates on GP-based novelty detection, which were recently proposed [12]. Before we explain the utilization of GP for one-class classification, we will first take a short look on Bayesian and especially GP regression. Details and proofs about those fundamental techniques can be found in the textbook of Rasmussen and Williams [29].

2.1. Gaussian Process Regression

Regression analysis deals with the problem of finding a relationship between some input variables $\mathcal{X}$ and output variables $\mathcal{Y}$, given a finite training sample $(\mathbf{X}, \mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$. A standard assumption in various regression frameworks is that outputs are generated by a function $f : \mathcal{X} \to \mathcal{Y}$ and an additional noise process ε (measurement error, label error etc.), i.e.

$$y(\mathbf{x}) = f(\mathbf{x}) + \varepsilon, \tag{1}$$

where one often constrains f to be member of a special parametric family (e.g. linear functions). Hence, regression boils down to estimating the parameters of the function by minimizing some loss function on the given training data.

In Bayesian regression, a probabilistic point of view is followed. Instead of assuming a specific parametric family for f, it is assumed that the function itself is drawn from a probability distribution. This approach allows introducing and propagating uncertainties with respect to the given training data.

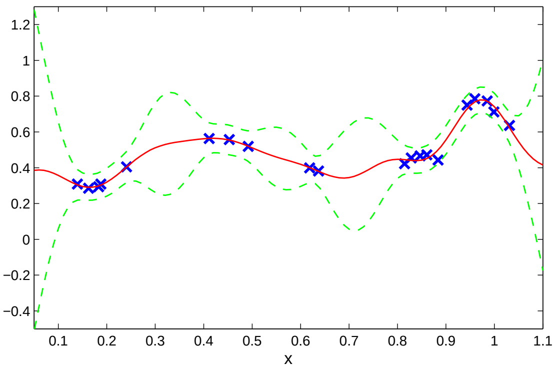

Figure 1: One-dimensional toy example for GP regression with isotropic SE-kernel ($\boldsymbol{\theta} = (1, 0.1)^T$, $\sigma_n^2 = 0.001$). Training points (crosses) are well fitted by the GP mean μ_* (solid line). The 95%-confidence interval $\mu_* \pm 1.96\sigma_*$ (spaced lines) illustrates the uncertainty of the prediction.

GP regression is a special case of Bayesian regression where f is assumed to be distributed according to a GP prior. Such a prior, which can be seen as a normal (i.e. Gaussian) distribution over functions, is solely specified by a mean function $m(\mathbf{x})$ and a positive definite covariance function $\kappa(\mathbf{x}, \mathbf{x}')$. Hence, we assume

$$f(\mathbf{X}) \sim \mathcal{N}(m(\mathbf{X}), \kappa(\mathbf{X}, \mathbf{X})) \tag{2}$$

for any finite collection of random variables $\mathbf{X} \in \mathcal{X}^n$.

The outstanding role of GP regression can be explained by its closed-form prediction for previously unseen inputs $\mathbf{x}_* \in \mathcal{X}$. By assuming Gaussian noise, i.e. $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$ and a zero-mean (m=0) GP prior, it can be shown [35] that outputs $y_* = y(x_*)$ are again normally distributed, i.e. $y_* \sim \mathcal{N}(\mu_*, \sigma_*^2)$:

$$\begin{aligned} \mu_* &= \mathbf{k}_*^T \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{y} && (3) \\ \sigma_*^2 &= k_{**} - \mathbf{k}_*^T \left(\mathbf{K} + \sigma_n^2 \mathbf{I}\right)^{-1} \mathbf{k}_* + \sigma_n^2 \ , && (4) \end{aligned}$$

where μ_* and σ_*^2 denote mean and variance of $p(y_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y})$, respectively, and the shorthands $\mathbf{K} = \kappa(\mathbf{X}, \mathbf{X})$, $\mathbf{k}_* = \kappa(\mathbf{X}, \mathbf{x}_*)$, and $k_{**} = \kappa(\mathbf{x}_*, \mathbf{X})$ are used for the sake of readability. A common covariance function widely used for regression and classification is the isotropic squared-exponential (SE) kernel

$$\kappa_{\mathrm{SE}}(\mathbf{x}, \mathbf{x}') = \theta_1^2 \cdot \exp\left\{ -\frac{1}{2\theta_2^2} \cdot \|\mathbf{x} - \mathbf{x}'\|^2 \right\} \ . \tag{5}$$

An example of GP regression applied to a one-dimensional toy example can be seen in Figure 1. A straightforward way to use GP regression for binary classification is to use binary labels $y \in \{-1, 1\}$, which is also called label regression. Other possibilities are Laplace approximation and Expectation propagation which can cope with non-Gaussian noise [29].

2.2. Extension to Novelty Detection

We recently showed that GP regression can also be used for the general task of novelty detection or one-class classification [12], although labels $\mathbf{y} \in \mathcal{Y}^n$ are usually not associated to data $\mathbf{X} \in \mathcal{X}^n$.

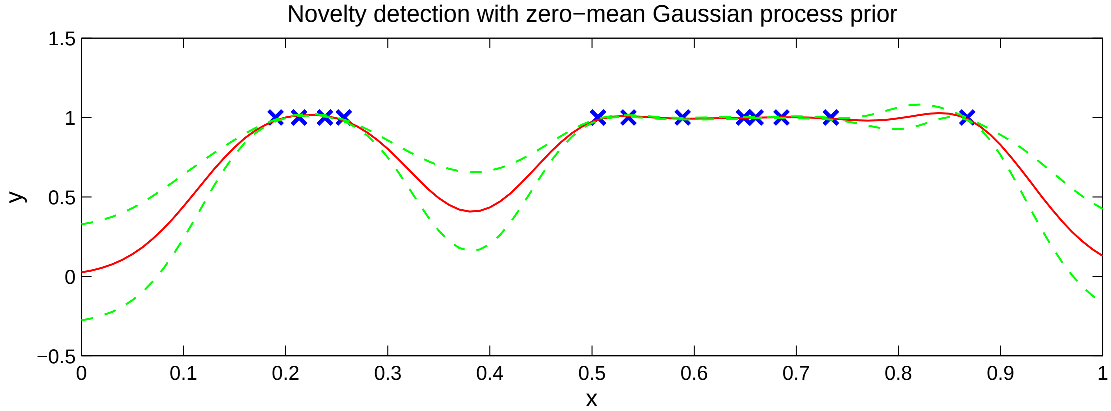

Figure 2: Principles for novelty detection with GP regression. Posterior GP moments for a one-dimensional toy example using isotropic SE-kernel ($\boldsymbol{\theta} = (1, 0.067)^T$, $\sigma_n^2 = 0.01$) are visualized (see Figure 1).

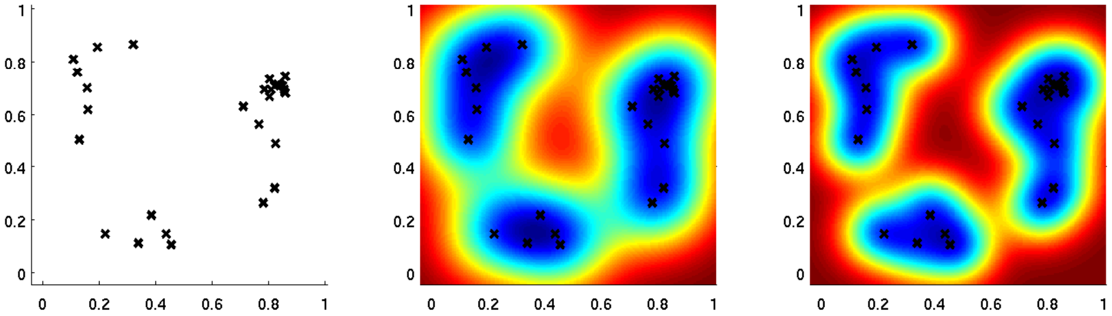

Figure 3: Novelty detection scores in two dimensions. Negative GP posterior mean $-\mu_*$ (middle) and variance σ_*^2 (right) are shown for an artificial training data set (left), where both scores are visualized via a heat map (samples from hot regions are more likely to be novel than those from cold regions).

The approach of [12] proceeds in augmenting all observed data (positive class) with positive class labels $y = 1$ and performing label regression using a zero-mean GP prior[1]. As can be seen in Figure 2 for a one-dimensional toy example, this combination leads to high values of μ_* (close to one) in the vicinity of the training data and low values (close to zero) far away from the observations. Alternatively, the uncertainty in our estimate represented by σ_*^2 increases when being far away from the training samples and nearly vanishes when being close to observations $\mathbf{X}$.

These properties enable the use of GP posterior moments (3) and (4) for novelty detection, where $-\mu_*$ and σ_*^2 can be utilized as scores which measure the novelty of a new sample $\mathbf{x}_*$. The behavior of both scores can be seen in Figure 3 for a two-dimensional toy example, where regions far away from the training data are likely to contain novel samples.

As highlighted in [12], this approach has some shortcomings, since the parameters of the model (i.e. the hyperparameters of the covariance function κ) cannot be estimated automatically. This comes from the fact that only one class is given in the training set and parameter optimization techniques such as maximum marginal likelihood [29] overadapt to the training data (leading to constant functions $y(x_*) = 1$ which are not useful here).

When dealing with Raman spectra datasets which comprise at least two strains (or species or genera), we can circumvent the problem of parameter estimation in the one-class setting by using estimates derived from binary (or multi-class) separation between strains. These estimates can then be used either directly for novelty detection or as a starting point for parameter search techniques such as consistency-based model selection [36].

2.3. Novelty Detection using a One-Vs-All Classifier

It is important to point out that, for the majority of Raman spectroscopic tasks, training datasets comprise labeled spectra from several known classes rather than one single set of spectra labeled as "known". This additional knowledge allows to employ an alternative novelty detection strategy which directly utilizes a multi-class classifier [37, 38].

For this task, we use a GP multi-class classifier which is based on binary one-vs-all tasks. In this approach, each binary task separates one class from all remaining categories and previously unseen spectra from known categories are usually identified with the class which has the highest predictive probability. Since the predictive probabilities of each binary task are

[1] In general, it can be shown that a constant label $y > 0$ in combination with a zero-mean GP prior suffices to yield reasonable estimates, preserving the ordering of scores.

low for spectra of novel categories, the probabilities can be directly seen as a set of one-class classification scores (one score for each known class). Section 4.4 explains how to combine all those scores to a single decision about the novelty of a new example.

3. State-of-the-art Methods for Comparison

3.1. Support Vector Data Description

One straightforward method to estimate a novelty score is to enclose all training data with a sphere. The score for new points is produced by computing the distance to the sphere's center. SVDD [30, 39, 31] follows this approach by finding the minimal enclosing sphere which surrounds the data. Formulated as a quadratic program, it is further possible to allow for noisy measurements by discarding a certain fraction of training points. The amount of neglected training points can be specified by the outlier fraction parameter $\nu \in [0, 1)$ (where $\nu = 0$ is a sloppy abbreviation for the standard formulation that does not take any noisy measurements into account). The fact that all computations involved in the sphere estimation can be expressed as inner products enables the use of arbitrary kernel functions [40] such as (5).

3.2. Gaussian Mixture Models

Novelty detection is tightly coupled with density estimation, since regions with high density (with respect to a given set of strains) are unlikely to contain novel strains and vice versa. Gaussian mixture models [23, 24] (GMM) are popular methods for density estimation which describe the training data as a superposition of scaled normal distributions. The scaling parameters as well as the moments of the normal distributions can be estimated using an iterative algorithm derived from the general Expectation-Maximization [41] principle. The number of clusters k can be determined by cross-validation on the training set. Since GMMs suffer from the curse of dimensionality [23, 24], high-dimensional data often needs to be transformed to a low-dimensional subspace. To obtain a suitable dimensionality of the subspace, cross-validation can again be used. As discussed in Schmid et al. [11], different strategies for novelty detection with GMMs can be followed.

3.3. Parzen Density Estimation

Following the same reasoning as for GMMs, other methods for density estimation can also be utilized for novelty detection. Parzen density estimation [42, 23, 24] is a non-parametric technique (i.e. the number of parameters scales with the number of training examples) which models the density as a superposition of kernel functions. Placing the kernel function $\kappa(\cdot, \cdot)$ over each training example $\mathbf{x}_i$, one arrives at a smoothed histogram $p(\mathbf{x}_*) = n^{-1} \sum_{i=1}^{n} \kappa(\mathbf{x}_i, \mathbf{x}_*)$. If the kernel itself is a density, the Parzen estimate is also a properly normalized probability density. A commonly used kernel is the normal distribution, i.e. $\kappa(\mathbf{x}, \mathbf{x}') = \mathcal{N}(\mathbf{x}'|\mathbf{0}, \mathbf{\Sigma})$, where covariance matrix $\mathbf{\Sigma}$ is the crucial bandwidth parameter of the kernel. The selection of suitable bandwidth parameters is a non-trivial problem, where often cross-validation and ad-hoc strategies are exploited [43]. One well-known ad-hoc method for one-dimensional data is Silverman's rule of thumb [44]. Using assumptions of normally distributed data, approximate estimates $\widetilde{\sigma}_i^2$ to the optimal Gaussian bandwidth parameter $\mathbf{\Sigma} = \mathrm{diag}(\sigma_i^2)$ can be easily computed by $\widetilde{\sigma}_i^2 = 1.06 \cdot \hat{\sigma}_i^2 n^{-\frac{2}{5}}$, where $\hat{\sigma}_i^2$ denotes the unbiased estimate of the data variance in component i of the input vectors.

3.4. Simulation of Missing Classes

One alternative approach proposed by Dundar et al. [25] is to explicitly simulate potential classes that are not part of the training set. Following the assumption that all classes can be described by a normal distribution with a pooled covariance function, a fixed number M of classes is created by sampling from a class prior distribution. The parameters of the class prior are simply estimated using the means of all given categories. A new test sample is categorized as new if it can be best explained by one of the simulated classes. Since the same covariance parameter is assumed for all categories, this boils down to a Nearest Neighbor classification using the means of all known and simulated classes. One major drawback of this approach is its random character, since simulated classes are generated according to random draws from the class prior. As mentioned by the authors [25], this also implies that a subspace reduction of the data should be performed in order to avoid sampling in high dimensional spaces.

4. Experiments

4.1. Raman Spectra Datasets

The current study is based on two datasets, measured by a micro-Raman setup (HR LabRam invers, Jobin-Yvon-Horiba, Bensheim, Germany). The spectrometer has an entrance slit of 100 μm, has a focal length of 800 mm, and is equipped with a 300-lines/mm grating. As excitation wavelengths the 532-nm line of a frequency doubled Nd:YAG laser (Coherent Compass, Dieburg, Germany) with a laser power of approx. 2.4 mW incident on the sample were used. The Raman scattered light was detected by a CCD camera operating at 220 K. A Leica PLFluoar 100× objective (NA 0.75) focused the laser light onto the samples ($\approx$0.7 μm focus diameter). The spectrometer was calibrated each day prior to measuring (using titanium dioxide). All cells were recorded from fused silica plates with an integration time of 60 s. The first (large) Raman spectra dataset $\mathcal{D}_L$ contains 5743 spectra from 10 different bacterial species (cf. Tab. 1) and 50 strains. The strains were chosen according to their occurrence in clean-room environments. The microorganisms were purchased from the German Collection of Microorganisms and Cell cultures (DSMZ, Braunschweig, Germany) and from the Institute for Infectious Biology at the University of Würzburg. The employed cultivation media consisted of NA (nutrition agar), S-1-NA (standard 1 nutrition agar), CA (corynebacterium agar) and CASO (trypticase soy yeast extract medium). The microorganisms were cultured under varying conditions with respect to nutrient medium, growing time and temperature. In addition, an independent test set $\mathcal{D}_I$, consisting

Table 1: Large dataset $\mathcal{D}_L$ comprising 50 “known” strains from 10 different species.

genus	species	strain	#examples
Bacillus	megaterium	DSM90	94
	pumilus	DSM27	65
		DSM2893	57
		DSM354	56
		DSM355	78
		DSM361	119
		DSM766	70
		DSM8786	74
	sphaericus	DSM28	71
		DSM396	98
		DSM488	73
	subtilis	DSM10	85
		DSM1087	94
		DSM347	103
		DSM618	85
		DSM6399	80
		DSM6889	80
		DSM9565	88
Escherichia	coli	DSM1058	68
		DSM2769	108
		DSM423	103
		DSM429	90
		DSM498	86
		DSM499	83
		DSM613	94
Micrococcus	luteus	DSM142340	99
		DSM14235	92
		DSM1605	90
		DSM1790	82
		DSM20030	124
		DSM348	82
		DSM46257	92
	lylae	DSM20315	102
		DSM20318	84
Staphylococcus	cohnii	DSM20260	137
		DSM20261	82
		DSM20262	79
		DSM6669	120
		DSM6718	124
		DSM6719	123
	epidermidis	195Isolat	85
		2682Isolat	203
		ATTC12218	139
		ATTC35984	833
		DSM1798	177
		DSM20042	171
		DSM3269	130
		DSM3270	146
	warneri	DSM20036	125
		DSM20316	120
			$\sum = 5743$

of 16 known strains (present in $\mathcal{D}_L$) and 6 unknown strains was recorded. Please note that the independent dataset $\mathcal{D}_I$ is equal to the one used in [11, 45]. Our large dataset $\mathcal{D}_L$ however is more complex, since the original training database in [11] includes only 29 bacterial strains from 9 species.

The composition of the large dataset $\mathcal{D}_L$ and independent dataset $\mathcal{D}_I$ is listed in Table 1 and Table 2, respectively.

4.2. Spectral Pre-processing

Both datasets $\mathcal{D}_L$ and $\mathcal{D}_I$ are pre-processed by performing local quadratic interpolation to obtain Raman intensities on a fixed (integer) wavenumber grid. All Raman signals are then cropped to the integer wavenumber range $\mathcal{I} = [540, 3350]\,\text{cm}^{-1}$ which is covered by all spectra. In order to suppress spike noise introduced by cosmic radiation, a running median filter is employed. For numerical stability, all spectra are further normalized to unit length.

Table 2: Independent dataset $\mathcal{D}_I$ comprising 299 Raman spectra (130 samples from 16 known strains and 169 samples from 6 unknown strains).

genus	species	strain	#examples	known in $\mathcal{D}_L$
Bacillus	sphaericus	DSM28	8	✓
		DSM396	7	✓
	subtilis	DSM347	8	✓
Escherichia	coli	DSM1058	20	✓
		DSM423	7	✓
		DSM426	24	×
		DSM498	7	✓
		DSM5208	26	×
Lactobacillus	acidophilus	DSM9126	25	×
Micrococcus	luteus	DSM3906	45	×
		DSM20030	6	✓
	lylae	DSM20315	5	✓
		DSM20318	5	✓
Staphylococcus	cohnii	DSM20260	7	✓
		DSM6669	8	✓
		DSM6718	5	✓
		DSM6719	5	✓
	epidermidis	195 Isolat	20	✓
		ATTC35984	7	✓
	warneri	DSM20036	5	✓
	hominis	BCD2684	21	×
Streptococcus	thermophilus	DSM20617	28	×
			$\sum = 299$	

4.3. Implementation Details

All experiments concerning novelty detection were done in Matlab using our code for one-class classification with GP regression [12] which can be downloaded at our homepage[2]. For parameter estimation we used the Matlab code[3] of Rasmussen et al. provided alongside their text book [29]. As in [46], the multi-class problem was tackled in one-vs-all fashion using a binary GP classifier with Laplace approximation and cumulative Gaussian likelihood. The parameter of the covariance function were estimated by maximizing marginal likelihood using the Conjugate Gradient optimizer `minimize` with 10 iterations for each binary one-vs-all problem. The additive noise component was set to a small value $\sigma_n^2 = 0.01$ to avoid numerical instabilities. SVDD was realized using Matlab's `quadprog` function. For GP regression and classification, as well as for SVDD, we used the isotropic SE-kernel (5) in all our experiments. Kernel parameters θ_1 and θ_2 are obtained by the one-vs-all classifier described in Sect. 2.3. For SVDD, different values for outlier fraction parameter $\nu \in \{0, \ldots, 0.9\}$ were investigated. For the GMM, we followed the approach of Schmid et al. [11] using Principle Component Analysis (PCA) as subspace reduction method and a full covariance matrix which is pooled over all strains. The number d of PCA components as well as the number k of normal distributions in the model were obtained by 10-fold cross-validation. Maximizing the average recognition rate on a 5×5-grid ($d \in \{10, 20, 30, 50, 80\}$ and $k \in \{5, 10, 20, 30, 50\}$), the optimum for our dataset was found to be $d = 30$ and $k = 30$. The approach of [25] was

[2] available at http://www.inf-cv.uni-jena.de/kemmler

[3] accessible at http://www.gaussianprocess.org/gpml/code/matlab/doc/

Table 3: Results of Novelty Detection using multiple labels (cf. 5.1), where $\mathcal{D}_L$ was utilized for training and $\mathcal{D}_I$ for testing. The average recognition rate (ARR) is also given as an unbiased performance measure of the novelty detection results. Methods used: Gaussian process regression mean and variance (GPR-M, GPR-V) [12], Support Vector Data Description (SVDD) [31], Gaussian Mixture Model (GMM), Parzen Density Estimation (Parzen), one-vs.-all GP classifier (OVA), sampling technique of [25] (MLS).

	GPR-V	GPR-M	SVDD	GMM	Parzen	OVA	MLS [25]
correctly recognized as novel (sensitivity)	169 (100%)	27 (16.0%)	2 (1.2%)	96 (56.8%)	169 (100%)	129 (76.3%)	112.7 (67.0%)
correctly recognized as known (specificity)	89 (68.5%)	125 (96.2%)	130 (100%)	117 (90.0%)	80 (61.5%)	73 (56.2%)	87.76 (67.5%)
average recognition rate	**84.2%**	56.1%	50.3%	73.4%	80.8%	66.2%	67.1%

re-implemented in Matlab and analyzed for a varying number $M \in \{10, 100, 500, 1000, 5000, 10000, 50000\}$ of simulated classes. Because of the randomized nature of the algorithm due to sampling, we always average over 100 runs. As in GMMs, we projected the Raman spectra onto the first $d = 30$ PCA components. The latter step is also done for Parzen density estimation. Using a normal density with diagonal covariance as kernel, Silverman's rule of thumb was used for estimating the bandwidth parameters for each dimension independently.

4.4. *Novelty Decision Process*

If not stated otherwise, novelty detection scores are computed for each single class separately. This is done to allow a fair comparison of all tested methods since the one-vs-all approach (cf. Sect. 2.3) inherently uses this class-centered methodology. Schmid et al. [11] also noted that this approach might be beneficial since treating the whole dataset as one class can be difficult due to high variability of the underlying strains. For estimating the class-specific inlier/outlier boundary, one threshold needs to be determined for each strain. In order to enable a direct comparison to Schmid et al. [11], we follow their strategy which is based on a pre-defined expected outlier ratio and which is similar to the usage of the outlier fraction ν in the SVDD approach (cf. Sect. 3.1): For each class, novelty scores are computed on the training data and the corresponding threshold is set equal to the 5-percentile of those values. This strategy artificially treats 5% of the training data as novelty. In our setting, we hence end up with 50 thresholds, one for each strain. For a test sample all 50 strain-dependent novelty scores are computed. Based on the learned thresholds, a decision is made whether this test sample is novel or not. Only if the test sample is treated as novelty by all strains, e.g. by using our GP novelty scores, it is declared as novel with respect to the whole dataset.

5. Results and Discussion

In the following, we compare the different approaches to novelty detection from Sect. 2 and highlight benefits and shortcomings of the respective methods. We also analyze whether the independent description of classes as proposed in [11] is necessary in our application. We end with a discussion of parameter estimation in non-parametric models for novelty detection.

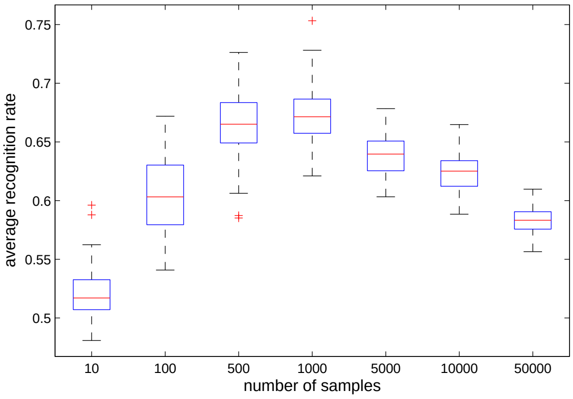

Figure 4: Novelty detection based on Maximum Likelihood with a varying number M of simulated outlier categories (MLS). The number of samples from the class prior is a critical parameter (optimum found at M=1000), where under- or oversampling leads to inferior performance.

5.1. *The Multiple-Label Case*

In this section, we assume that all classes are treated separately and hence multiple labels are given. For Gaussian mixture models, Parzen density estimators, SVDD as well as for GP regression, a set of scores and thresholds are computed as outlined in Sect. 4.4. The approach using simulated classes (MLS) from Sect. 3.4 does not require thresholds since outlying classes are directly modeled. All methods are trained on the large Raman spectra dataset $\mathcal{D}_L$ and tested on the independent dataset $\mathcal{D}_I$. The number of correctly recognized novelties and known spectra are presented in Table 3 along with the resulting average recognition rate (with respect to class "novel" and "known").

The results clearly indicate that GP regression is suitable for the task of novel detection in a bacterial context. The variance of the predictive distribution (GPR-V) outperformed all other tested approaches. While all samples from unknown strains are correctly detected, only 41 out of 130 known spectra (31.5%) are erroneously treated as novel which yields an average recognition rate of 84.2%. A comparable performance is obtained using Parzen density estimation with a higher misclassification rate of 39.5% for known spectra. Medium quality performances were obtained using GMMs, the simulated class approach (MLS) and the one-vs-all GP classifier (OVA). Please note that the performance for MLS heavily depends on the number M of samples drawn from the class prior. Moreover, as can be seen in Figure 4, larger values of M do not directly lead

Table 4: Results of Novelty Detection using a single positive class which comprises all training data.

	GPR-V	Parzen	GMM (k=30)	GMM (k=100)	GMM (k=500)	GMM (k=1500)
correctly recognized as novel (sensitivity)	160 (94.7%)	149 (88.2%)	13 (7.7%)	32 (18.9%)	57 (33.7%)	63 (37.3%)
correctly recognized as known (specificity)	101 (77.7%)	107 (82.3%)	128 (98.5%)	129 (99.2%)	128 (98.5%)	128 (98.5%)
average recognition rate	**86.2%**	85.2%	53.1%	59.1%	66.1%	67.9%

to better decision functions. This behavior can be explained by the fact that known classes are eventually suppressed when more and more samples are drawn from the class prior. Since a suitable method for estimating M is not known to us, we henceforth refer to the best performance in order to allow for a fair (or rather optimistic) performance comparison with respect to the remaining methods. The approach based on the mean of the predictive probability from GP regression (GPR-M) as well as SVDD (best performance obtained for $\nu = 0.1$) completely failed in our experiment, leading to recognition results which are close to a random decision (56.1% and 50.3%, respectively).

5.2. *The One-Class Case*

Instead of assuming all class labels are known, we could also treat the training spectra as if they were generated from a single "known" class. This point of view however rules out the comparison to methods which explicitly incorporate the multi-label information such as one-vs-all classifiers and the approach which simulates unknown classes (where the unbiased sample of the class priors' covariance matrix is not defined for one class). It is a priori reasonable to assume that the cancellation of class label information might lead to poor results, since the high variation of the whole data needs to be explained by a single model, e.g. a mixture of Gaussians [11]. This argument however only holds for parametric models, where the information of the data is absorbed into a few parameters. Flexible non-parametric methods such as Parzen density estimation, SVDD and Gaussian processes are not likely to suffer from this problem. These models hence offer the possibility to solve the novelty detection task in only one step requiring only one threshold. This circumvents the merging step where a multitude of class-dependent scores must be taken into account (and where one badly estimated threshold parameter might lead to severe consequences for the whole novelty detection task).

The results of novelty detection using this single-class point of view are displayed in Table 4. For the sake of brevity, SVDD and GPR-M are not listed because of their inferior behavior for the multi-label case. As GP hyperparameters, we use the mean $\overline{\boldsymbol{\theta}}$ of all class-dependent hyperparameters $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{50}\}$ obtained via the OVA method. The results show that treating the dataset as one class does not necessarily imply a drop in novelty detection performance. On the contrary, even a slight increase in average recognition result could be achieved for Parzen density estimation (4.4%) and GPR-V (2%). As in the multi-label case, these two approaches substantially outperform the GMM. The parametric GMM is clearly worse in our experiments, although we increased the number of latent clusters up to $k = 1500$ in order to account for a higher variability of the data.

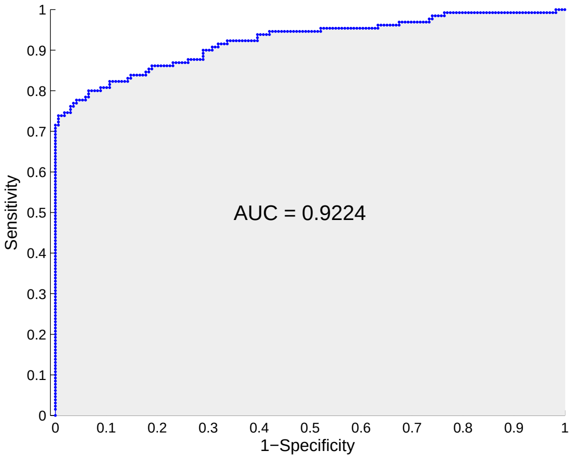

Figure 5: Example ROC curve for GPR-V using $\overline{\boldsymbol{\theta}}$.

5.3. *Bandwidth Parameter Selection*

As mentioned in the introduction of Parzen density estimation, bandwidth parameter selection is a hard problem. In order to estimate hyperparameters of the kernel or even the kernel type itself, labeled training data is often required. This however, is not always available, especially in the one-class classification case. The choice of the kernel is indeed crucial as can be seen in Figure 6, where the GP-based measures GPR-M and GPR-V as well as SVDD (for $\nu = 0.1$) are analyzed with varying length scale parameter θ_2. In order to allow a compact performance comparison, we are resorting to receiver operating characteristic (ROC) curves, which illustrates the sensitivity and (1-)specificity for all possible thresholds. As can be seen in the ROC curve for GPR-V using parameter $\overline{\boldsymbol{\theta}}$ from the previous section, the area under the roc curve (AUC) gives an estimate of the expected performance of the method at hand (cf. Figure 5). Varying the bandwidth parameter $\log \theta_2 \in [-10, 3]$, we can clearly see that the performance changes drastically for both GP sores and SVDD[4]. Moreover, it becomes apparent that GPR-M and SVDD are suitable for novelty detection. This closer analysis reveals that the estimate $\log \overline{\theta}_2 = -1.6145$ which was used in previous settings is far away from the optimal value for GPR-M and SVDD while it leads to successful predictions for GPR-V.

[4] For efficiency, a fast implementation from the LIBSVM [47] toolbox was used to solve this larger problem

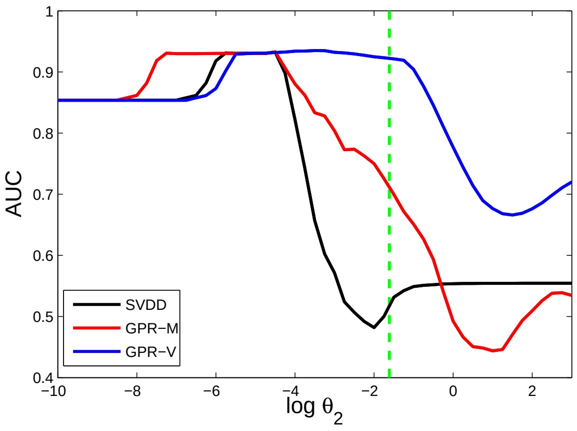

Figure 6: Dependence of hyperparameter selection upon novelty detection performance for SVDD and GP regression scores. The estimated parameter $\bar{\theta}_2$ is marked with a spaced vertical line.

6. Conclusions and Future Work

In this work, we proposed different strategies to tackle the problem of novelty detection for Raman spectroscopy. Using novelty scores derived from state-of-the-art techniques such as Gaussian mixture models, Support Vector Data Description and Gaussian processes, we discussed the power and pitfalls of all methods. For a highly variable bacterial Raman spectra dataset, Gaussian process based scores [12] showed superior performance to all other tested methods.

This paper highlights the difficult task of novelty detection for many real-world applications. Categorization without outlier detection is especially in chemometrics a risky procedure, e.g. if samples are contaminated or the training set is not well selected. Integration of the GP-based scores presented in this paper can be easily realized and leads to a more robust analysis.

Since the problem of novelty detection is far from being solved, much work still needs to be done such as finding reasonable values for kernel hyperparameters for non-parametric techniques and useful decision thresholds. Current kernel functions, such as the isotropic squared-exponential kernel, treat the components of the Raman spectra independently, which seems to be unreasonable with regard to the fact that the spectra correspond to one-dimensional functions. Therefore, another open question is whether we can develop kernel functions specially designed for Raman spectroscopy. The study in this paper only focused on detecting novelties with respect to the strain level. In practical applications it might be beneficial to determine different levels of granularity, e.g. novelty with respect to strains, species and genera. Our novelty detection using Gaussian processes can also be applied to those settings where hierarchical information is used to gain a better understanding concerning the type of novelty detected.

Acknowledgments

We want to thank Mario Krause and Michaela Harz from the Institute of Physical Chemistry at the FSU Jena for capturing the bacterial datasets. This work was partially funded by the TMBWK ProExzellenz project ”MikroPlex“ (PE113-1).

References

[1] G. Barrow, R. Feltham, Cowan and Steel's Manual for the Identification of Medical Bacteria, Cambridge University Press, 3 edn., 1993.

[2] S. Al-Khaldi, M. Mossoba, Gene and Bacterial Identification Using High Throughput Technologies: Genomics, Proteomics, and Phenomics., Nutrition 20 (2004) 32–38.

[3] J. Chalmers, P. Griffiths (Eds.), Handbook of Vibrational Spectroscopy, vol. 1-5, Wiley, 2002.

[4] F. Siebert, P. Hildebrandt, Vibrational Spectroscopy in Life Science, Wiley, 2007.

[5] M. Diem, P. Griffiths, J. Chalmers (Eds.), Vibrational Spectroscopy for Medical Diagnosis, Wiley, 2008.

[6] J. Ferraro, K. Nakamoto, C. Brown, Introductory Raman Spectroscopy (Second Edition), Elsevier, 2003.

[7] P. Buijtels, H. Willemse-Erix, P. Petit, H. Endtz, G. Puppels, H. Verbrugh, A. van Belkum, D. van Soolingen, K. Maquelin, Rapid identification of mycobacteria by Raman spectroscopy, J. Clin. Microbiol. 46 (3) (2008) 961–965.

[8] D. I. Ellis, R. Goodacre, Rapid and quantitative detection of the microbial spoilage of muscle foods: current status and future trends, Trends Food Sci. Tech. 12 (11) (2001) 414 – 424.

[9] P. Rösch, M. Harz, K.-D. Peschke, O. Ronneberger, H. Burkhardt, H.-W. Motzkus, M. Lankers, S. Hofer, H. Thiele, J. Popp, Chemotaxonomic Identification of Single Bacteria by Micro-Raman Spectroscopy: Application to Clean-Room-Relevant Biological Contaminations, Appl. Environ. Microb. 71 (2005) 1626–1637.

[10] M. Harz, P. Rösch, J. Popp, Vibrational spectroscopy – A powerful tool for the rapid identification of microbial cells at the single-cell level, Cytometry 75A (2009) 104–113.

[11] U. Schmid, P. Rösch, M. Krause, M. Harz, J. Popp, K. Baumann, Gaussian mixture discriminant analysis for the single-cell differentiation of bacteria using micro-Raman spectroscopy, Chemometr. Intell. Lab. 96 (2) (2009) 159 – 171.

[12] M. Kemmler, E. Rodner, J. Denzler, One-Class Classification with Gaussian Processes, in: Proc. ACCV, 489–500, 2010.

[13] A. Nijssen, K. Maquelin, L. F. Santos, P. J. Caspers, T. C. Bakker Schut, J. C. den Hollander, M. H. A. Neumann, G. J. Puppels, Discriminating basal cell carcinoma from perilesional skin using high wave-number Raman spectroscopy, J. Biomed. Opt. 12 (3) (2007) 034004.

[14] R. M. Jarvis, R. Goodacre, Ultra-violet resonance Raman spectroscopy for the rapid discrimination of urinary tract infection bacteria, FEMS Microbiol. Lett. 232 (2) (2004) 127–132.

[15] A. S. Haka, Z. Volynskaya, J. A. Gardecki, J. Nazemi, R. Shenk, N. Wang, R. R. Dasari, M. Fitzmaurice, M. S. Feld, Diagnosing breast cancer using Raman spectroscopy: prospective analysis, J. Biomed. Opt. 14 (5) (2009) 054023.

[16] U. Neugebauer, T. Bocklitz, J. Clement, C. Krafft, J. Popp, Towards detection and identification of circulating tumour cells using Raman spectroscopy., Analyst 135 (2010) 3178–3182.

[17] K. Maquelin, C. Kirschner, L. P. Choo-Smith, N. van den Braak, H. P. Endtz, D. Naumann, G. J. Puppels, Identification of medically relevant microorganisms by vibrational spectroscopy, J. Microbiol. Meth. 51 (3) (2002) 255 – 271.

[18] M. Harz, S. Stöckel, V. Ciobotă, D. Cialla, P. Rösch, J. Popp, Applications of Raman Spectroscopy to Virology and Microbial Analysis, in: P. Matousek, M. D. Morris (Eds.), Emerging Raman Applications and Techniques in Biomedical and Pharmaceutical Fields, Biological and Medical Physics, Biomedical Engineering, Springer Berlin Heidelberg, 439–463, 2010.

[19] M. G. Forero, G. Cristobal, M. Desco, Automatic identification of Mycobacterium tuberculosis by Gaussian mixture models, J. Microsc. 223 (2006) 120132.

[20] M. Krause, B. Radt, P. Rösch, J. Popp, The identification of single living bacteria by means of fluorescence staining and Raman spectroscopy, J. Raman Spectrosc. 38 (2007) 369–372.

[21] M. Krause, P. Rösch, B. Radt, J. Popp, Localizing and identifying living bacteria in an abiotic environment by a combination of Raman and fluorescence microscopy, Anal. Chem. 80 (2008) 8568–8575.

[22] D. M. J. Tax, One-class classification, Ph.D. thesis, Delft University of Technology, 2001.

[23] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, Wiley, New York, 2. edn., 2001.

[24] C. M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer, 1 edn., 2007.

[25] M. M. Dundar, E. D. Hirleman, A. K. Bhunia, J. P. Robinson, B. Rajwa, Learning with a non-exhaustive training dataset: a case study: detection of bacteria cultures using optical-scattering technology, in: Proc. ACM SIGKDD, 279–288, 2009.

[26] A. Banhalmi, A. Kocsor, R. Busa-Fekete, Counter-Example Generation-Based One-Class Classification, in: J. Kok, J. Koronacki, R. Mantaras, S. Matwin, D. Mladenic, A. Skowron (Eds.), Proc. ECML, vol. 4701, 543–550, 2007.

[27] D. M. J. Tax, R. P. W. Duin, Uniform object generation for optimizing one-class classifiers, J. Mach. Learn. Res. 2 (2002) 155–173, ISSN 1532-4435.

[28] F. Akova, M. Dundar, V. J. Davisson, E. D. Hirleman, A. K. Bhunia, J. P. Robinson, B. Rajwa, A machine-learning approach to detecting unknown bacterial serovars, Stat. Anal. Data Min. 3 (2010) 289–301.

[29] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning), The MIT Press, 2005.

[30] D. M. Tax, R. P. Duin, Support Vector Domain Description, Pattern Recogn. Lett. 20 (11-13) (1999) 1191 – 1199.

[31] D. M. J. Tax, R. P. W. Duin, Support Vector Data Description, Mach. Learn. 54 (2004) 45–66.

[32] Y. Chen, X. Zhou, T. S. Huang, One-Class SVM for Learning in Image Retrieval, in: Proc. ICIP, 2001.

[33] W. De Ruig, J. Weseman, A new approach to confirmation by infrared spectroscopy, J. Chemometr. 4 (1990) 61–77.

[34] J. Yang, N. Zhong, P. Liang, J. Wang, Y. Yao, S. Lu, Brain activation detection by neighborhood one-class SVM, Cogn. Sys. Res. 11 (1) (2010) 16 – 24.

[35] R. von Mises, Mathematical theory of probability and statistics, Academic Press, 1964.

[36] D. M. Tax, K.-R. Müller, A Consistency-Based Model Selection for One-Class Classification, in: Proc. ICPR, vol. 3, 363–366, 2004.

[37] W. Schumacher, M. Kühnert, P. Rösch, J. Popp, Identification and classification of organic and inorganic components of particulate matter via Raman spectroscopy and chemometric approaches, J. Raman Spectrosc. doi:10.1002/jrs.2702.

[38] S. Stöckel, W. Schumacher, S. Meisel, M. Elschner, P. Rösch, J. Popp, Raman Spectroscopy-Compatible Inactivation Method for Pathogenic Endospores, Appl. Environ. Microbiol. 76 (2010) 28952907.

[39] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, R. C. Williamson, Estimating the Support of a High-Dimensional Distribution, Neural Comput. 13 (2001) 1443–1471.

[40] B. Schölkopf, A. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge, MA, USA, ISBN 0262194759, 2001.

[41] A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm., J. Roy. Stat. Soc. Met. B 39 (1) (1977) 1–38.

[42] E. Parzen, On estimation of a probability density function and mode, Ann. Math. Stat. 33 (1962) 1065–1076.

[43] W. Härdle, M. M., S. Sperlich, A. Werwatz, Nonparametric and Semiparametric Models, Springer, 2004.

[44] B. Silverman, Density Estimation for Statistics and Data Analysis, Chapman & Hall/CRC, 1986.

[45] P. Rösch, M. Harz, K.-D. Peschke, O. Ronneberger, H. Burkhardt, A. Schüle, G. Schmautz, M. Lankers, S. Hofer, H. Thiele, H.-W. Motzkus, J. Popp, Online Monitoring and Identification of Bioaerosols, Anal. Chem. 78 (2006) 2163–2170.

[46] M. Kemmler, J. Denzler, P. Rösch, J. Popp, Classification of Microorganisms via Raman Spectroscopy Using Gaussian Processes, in: Proc. DAGM, 81–90, 2010.

[47] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, code available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm, 2001.