A Machine-Learning Approach to Detecting Unknown Bacterial Serovars

Ferit Akova[1], Murat Dundar[1*], V. Jo Davisson[2,3], E. Daniel Hirleman[4], Arun K. Bhunia[5], J. Paul Robinson[3] and Bartek Rajwa[3]

AQ1 [1]*Department• of Computer and Information Science, Indiana University–Purdue University, Indianapolis, IN 46202, USA*

[2]*Department of Medicinal Chemistry and Molecular Pharmacology, Purdue University, W. Lafayette, IN 47907, USA*

[3]*Bindley Bioscience Center, Purdue University, W. Lafayette, IN 47907, USA*

[4]*School of Mechanical Engineering, Purdue University, W. Lafayette, IN 47907, USA*

[5]*Department of Food Science, Purdue University, W. Lafayette, IN 47907, USA*

Received 15 September 2009; revised 13 June 2010; accepted 30 June 2010
DOI:10.1002/sam.10085
Published online in Wiley InterScience (www.interscience.wiley.com).

UNCORRECTED PROOFS

Abstract: Technologies for rapid detection of bacterial pathogens are crucial for securing the food supply. A light-scattering sensor recently developed for real-time identification of multiple colonies has shown great promise for distinguishing bacteria cultures. The classification approach currently used with this system relies on supervised learning. For accurate classification of bacterial pathogens, the training library should be exhaustive, i.e., should consist of samples of all possible pathogens. Yet, the sheer number of existing bacterial serovars and more importantly the effect of their high mutation rate would not allow for a practical and manageable training. In this study, we propose a Bayesian approach to learning with a nonexhaustive training dataset for automated detection of *unmatched* bacterial serovars, i.e., serovars for which no samples exist in the training library. The main contribution of our work is the Wishart conjugate priors defined over class distributions. This allows us to employ the prior information obtained from known classes to make inferences about unknown classes as well. By this means, we identify new classes of informational value and dynamically update the training dataset with these classes to make it increasingly more representative of the sample population. This results in a classifier with improved predictive performance for future samples. We evaluated our approach on a 28-class bacteria dataset and also on the benchmark 26-class letter recognition dataset for further validation. The proposed approach is compared against state-of-the-art involving density-based approaches and support vector domain description, as well as a recently introduced Bayesian approach based on simulated classes. © 2010 Wiley Periodicals, Inc. Statistical Analysis and Data Mining 3: 000–000, 2010

Keywords: nonexhaustive training data; Bayesian classifier; novelty detection; anomaly detection

1. INTRODUCTION

Outbreaks of methicillin-resistant *Staphylococcus aureus* [1], contamination of spinach and ground beef with *Escherichia coli* O157:H7 [2,3], presence of *Salmonella* in peanut butter [4,5], *Listeria monocytogenes* in ready-to-eat meats [6], or *Clostridium botulinum* in canned chili sauce are just a few examples of recent public-health threats. Serious concerns about bioterrorism and the possibility of intentional contamination of food products or agricultural commodities are not limited to bad science-fiction movies anymore [7–10].

Correspondence to: Murat Dundar (dundar@cs.iupui.edu)

© 2010 Wiley Periodicals, Inc.

Traditional bacteria recognition methods based on antibodies or genetic matching remain labor intensive and time consuming, and involve multiple steps. Moreover, samples are usually destroyed by these types of tests and thus are unavailable for further confirmatory assessment.

To perform classification of bacteria in a label-free manner (i.e., without use of biochemical reagents or genetic probes), a prototype system based on optical scattering technology, called BActeria Rapid Detection using Optical scattering Technology (BARDOT) has recently been developed [11]. In this system, bacterial colonies consisting of the progeny of a single parent cell scatter 635-nm laser light to produce unique forward-scatter signatures. Some

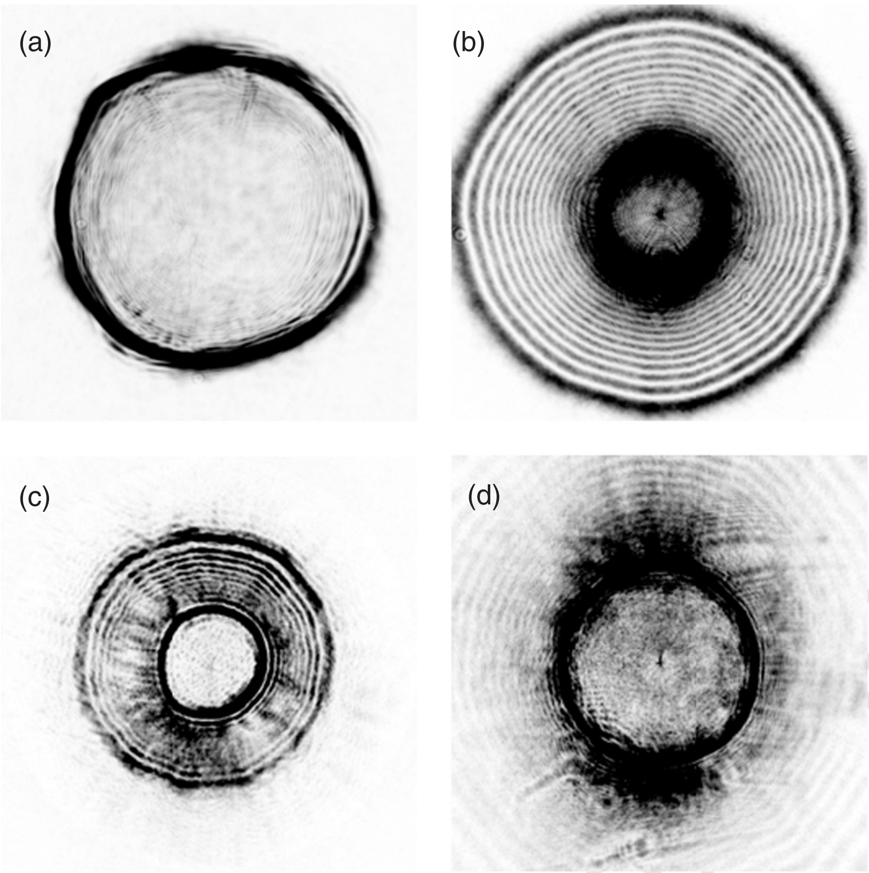


Fig. 1 (a) Sample scatter pattern for Salmonella Typhimurium (Copenhagen). (b) Sample scatter pattern for Vibrio orientalis CECT629. (c) Sample scatter pattern for Listeria seeligeri V45. (d) Sample scatter pattern for *Staphylococcus aureus* S-41.

UNCORRECTED PROOFS

examples of these scatter patterns are shown in Fig. 1. These scatter 'fingerprints,' which carry distinctive characteristics of bacterial phenotypes, are used for the off-line training of a supervised classifier. Subsequently this classifier is employed to identify bacterial colonies obtained from enriched samples submitted for testing. As currently implemented the system shows remarkable accuracy for bacteria belonging to numerous strains of *Listeria, Staphylococcus, Salmonella, Vibrio, and E. coli.*

1.1. Nonexhaustive Training Data

The goal of machine learning is to build robust models that, when deployed in a real-life application, generalize well to as-yet unseen examples of the sample population. Among the many factors that influence the generalizability of a learning algorithm, an exhaustive training dataset is perhaps the most critical. A training dataset is exhaustive if it contains samples from all classes of informational value. When some of the classes are not yet known and hence not represented, the resulting training dataset is *nonexhaustive*. A classifier trained using this dataset will misclassify a sample of a yet unseen class with probability 1, making the associated learning problem ill defined.

Generally, in applications with evolving datasets, the existing set of known classes is by definition nonexhaustive. To relate this to the bacterial detection application considered in this study, for the purpose of training only the most prevalent serovars of bacteria are used, as it is impractical to assume the presence of all bacteria types in the tested samples. This is because the sheer number of serovars would not allow for a practical and manageable training: *Salmonella* alone has over 2400 serovars. Additionally, bacteria are characterized by a high mutation rate, which can influence their pathogenicity, and new emerging pathogens may be rapidly introduced to a geographical area. Therefore, any training dataset for bacteria is inherently nonexhaustive and collecting an exhaustive set is impossible. On the other hand, classifying pathogenic bacteria as nonpathogenic would have unfortunate consequences. Therefore, the current traditional supervised classifier should be supplemented with a new rigorous machine-learning approach capable of addressing the problem of the nonexhaustive nature of available training libraries.

1.2. Proposed Approach and Its Relation to Early Work

One particular area of machine learning that is related to the nonexhaustiveness problem is anomaly detection [12–15]. Both anomaly detection and the problem of nonexhaustive learning aim to detect samples that are not represented in the training data, and in that regard they can be considered similar. However, an anomaly by definition is something peculiar, irregular, abnormal, or difficult to classify. Therefore anomalies can be considered outliers, and as such they could be as different from each other as they are from 'normal cases' [15]. More specifically, anomalies do not necessarily have informational value and it is very difficult if not impractical to model them. In contrast, samples originating from an unknown class have informational value, and just like any class available in the training set they could be modeled, were they known during training.

Another line of work that is related to the current research is developed for 'novelty detection' [16–18]. Unlike anomalies, novelties originate from hidden, missing or not yet known classes and thereby have informational value. Novelty detection is also sometimes referred to in the literature as 'novel class detection.' Most of the early work on novelty detection is developed around one-class classification problems and uses either support estimation [19,20] or density-based models to tackle the nonexhaustive nature of training datasets.

Our earlier work [21] that attempts to discover novelties in the presence of a large number of classes differs from earlier studies by proposing an empirical Bayesian approach to deal with nonexhaustive training datasets. In this method, all classes (known and unknown) are assumed to have Gaussian distributions with a common covariance matrix. A prior is defined over the mean vectors of the classes and its parameters are estimated using the training data acquired from the known classes. A large number of samples are generated from the prior to simulate the space of all classes. A new instance is classified using a maximum likelihood (ML) classifier and is considered a novelty if it is classified into one of the simulated classes. This attempt, although looks promising, has certain limitations. First, the common covariance assumption is quite restrictive. Second, the Gaussian prior defined for the mean vectors requires a very large number of classes to be available in the training dataset, to avoid numerical problems in estimating the parameters of the prior. Third, as the dimensionality increases, the number of simulated classes necessary to achieve higher specificities increases exponentially.

What we present in the current study is a real-time system that works in a multiclass setting, incorporates supervised classification and novelty detection together, and evaluates new samples sequentially. Our approach, which assumes Gaussian distributions for all classes (known and unknown), is based on Bayesian ML detection of novelties using a dynamically updated training dataset. The assumption of Gaussianity implies that the resulting sample covariance matrices are distributed according to a Wishart distribution. Since Wishart and inverted Wishart are conjugate priors, we define an inverted Wishart distribution over the covariance matrices as prior. Under this setting, the posterior distribution given the sample covariance matrices is also an inverted Wishart distribution. Covariance matrices for each class are estimated using the posterior means. Then, a ML classifier is designed using the class data in the training set. When a new sample emerges, class-conditional likelihoods are computed and the sample is classified to the class maximizing the likelihood provided that the maximum value is above a designated threshold. If the likelihood lies below that value, the sample is considered a novelty and a new class is created. The mean vector of this new class is the sample itself, and its covariance matrix is estimated using the posterior mean. Once the parameters are estimated, the existing set of known classes is augmented with this newly created class. In this approach, the parameters of the classes known before training are estimated once in the beginning, whereas those of newly created classes are recursively updated as more samples are assigned to these classes via sequential classification.

The proposed nonexhaustive learning algorithm can be confused with other algorithms developed around the lines of the online/incremental learning concept. The current study deals with the nonexhaustiveness of the classes and proposes an approach to identify novelties before they are incorrectly classified into existing classes. On the other hand, incremental/online learning deals with the issue of learning with the past and present data to improve the classifier performance in general with no specific emphasis on novelty detection. The main conceptual difference between our approach and incremental/online learning is that we consider the initially existing set of classes as definitive. Samples from these classes are obtained and validated using thorough procedures involving manual processing. To avoid updating class parameters with potentially incorrectly classified samples, only newly defined classes for novelties are updated as more samples are classified into these classes. On the other hand, in incremental/online learning, the present data are used to update all class definitions.

The rest of the article is organized as follows. Section 2 presents the technical details of our algorithm. Sectiton 2.1 reviews ML detection and density-based approaches for identifying novelties. Section 2.2 discusses the Gaussianity assumption for class-conditional distributions. Section 2.3 presents Wishart and inverted Wishart conjugate priors for prior modeling and posterior estimation of the covariance matrix. Section 2.4 introduces an algorithm for detecting novelties and discovering new classes. Finally, experimental

UNCORRECTED PROOFS

results are included in Section 3. Therein, we first present results for the bacteria detection problem and then use the benchmark letter recognition dataset for further validation of our approach. The proposed approach is compared against other density-based approaches as well as support vector domain description (SVDD) technique [19] and the simulated Bayesian modeling approach presented by Dundar et al. [21].

2. NOVELTY DETECTION SYSTEM

In this section, we present the details of the proposed approach. Sections 2.1 and 2.2 briefly review ML detection and its implementation with Gaussian class-conditional distributions. Sections 2.3 and 2.4 discuss our contributions to novelty detection.

2.1. Bayesian Maximum Likelihood Detection

Density-based approaches use class-conditional likelihoods of samples to detect novelties. In short, if the maximum of the class-conditional likelihoods is above a designated threshold, then the sample belongs to one of the classes in the training library and is assigned the corresponding class label; otherwise the sample is identified as belonging to an unrepresented class, hence a novelty.

More formally, let Ω, Δ, and Γ denote the set of *all*, *known*, and *unknown* bacteria classes, respectively, with $\Omega = \Delta \cup \Gamma$; A, K, and M are their corresponding cardinalities with $A = K + M$. The decision that minimizes the Bayes risk under the 0/1 loss-function assumption assigns a new sample x^* to the class with the highest posterior probability. More specifically,

$$x^* \in \omega_i^* \;\text{ s.t. }\; p_i^*(\theta_i|x^*) = \max_i \left\{p_i(\theta_i|x^*)\right\}, \tag{1}$$

where $i = \{1, \ldots, A\}$. Here ω_i represents the ith class and θ_i the parameters of its distribution. The classifier obtained by evaluating this decision rule is known as a maximum a posteriori classifier (MAP) [22].

Using Bayes' rule, the above decision rule can be rewritten as follows:

$$x^* \in \omega_i^* \;\text{ s.t. }\; p_i^*(\theta_i|x^*) = \max_i \left\{\frac{f_i(x^*|\theta_i)\pi_i(\theta_i)}{p(x^*)}\right\}, \tag{2}$$

where $f_i(x^*|\theta_i)$ is the *likelihood* of x^*, $\pi(\theta_i)$ is the *prior*, and $p(x^*)$ is the *evidence*. The evidence $p(x^*)$ is the same for all classes, and hence can be removed from the above formulation. When all classes are assumed *a priori* likely, $\pi(\theta_i)$ can be dropped from (2) as well. This leaves us with the ML decision function for classifying x^*:

$$x^* \in \omega_i^* \;\text{ s.t. }\; f_i^*(x^*|\theta_i) = \max_i\{f_i(x^*|\theta_i)\}, \tag{3}$$

where x^* is considered a novelty if $\omega_i^* \in \Gamma$, and a sample of a known class if $\omega_i \in \Delta$.

Since the set of classes is nonexhaustive $f_i(x^*|\theta_i)$ cannot be computed for all classes and as a result the decision function in (3) cannot be evaluated explicitly. We can express (3) in terms of ω_i^* and rewrite it by separating $f_i(x^*|\theta_i)$ of *known* and *unknown* classes as

$$h(x^*) = \begin{cases} x^* \textit{ is known} & \text{if } \psi \geq \gamma, \\ x^* \textit{ is novelty} & \text{if } \psi < \gamma, \end{cases} \tag{4}$$

where $\psi = \max_{\{i:\omega_i \in \Delta\}} \{f_i\ (x^*|\theta_i)\}$ and $\gamma = \max_{\{i:\omega_i \in \Gamma\}} \{f_i\ (z|\theta_i)\}$.

Since no data are available for unknown classes, γ cannot be explicitly estimated. In our experiments, we consider γ as a tuning parameter to optimize sensitivity at a desired specificity or vice versa. In other words, γ is the parameter to adjust for the compromise between sensitivity and specificity of the system.

To summarize, if the conditional likelihood of a known class for a sample x^* is less than γ, then x^* is a sample from an unrecognized class; otherwise x^* is a sample from a known class and thus can be assigned a known class label.

2.2. Gaussianity Assumption and Covariance Estimation

The most common and effective way to treat data of unknown nature is to assume Gaussian distributions for all classes, $\omega_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, $\theta_i = \left\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right\}$.

With this assumption in place, Eq. (4) becomes

$$h(x^*) = \begin{cases} x^* \textit{ is known} & \text{if } \min_{\{i:\omega_i \in \Delta\}} g_i(x^*) \leq \gamma, \\ x^* \textit{ is novelty} & \text{if } \min_{\{i:\omega_i \in \Delta\}} g_i(x^*) > \gamma, \end{cases} \tag{5}$$

where $g_i(x^*) = \log(|\boldsymbol{\Sigma}_i|) + (x^* - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(x^* - \boldsymbol{\mu}_i)$ is the negative log-likelihood of class ω_i given x^* and $|\boldsymbol{\Sigma}_i|$ is the determinant of $\boldsymbol{\Sigma}_i$. For $\{i : \omega_i \in \Delta\}$, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ can be estimated from class-conditional data available in the training set.

When dealing with datasets containing limited numbers of training samples and high dimensionality, the covariance estimator plays an important role in the modeling of the class-conditional distributions. The sample covariance can be obtained using the following formula:

$$\boldsymbol{\Sigma}_i = \frac{1}{n_i - 1}\left(X_i - \boldsymbol{\mu}_i \boldsymbol{e}_{n_i}^T\right)\left(X_i - \boldsymbol{\mu}_i \boldsymbol{e}_{n_i}^T\right)^T, \tag{6}$$

UNCORRECTED PROOFS

where n_i is the number of samples in class ω_i, $\boldsymbol{e}_{n_i}$ is a vector of ones of size n_i and $\boldsymbol{\mu}_i$ are the mean vectors estimated as

$$\boldsymbol{\mu}_i = \frac{1}{n_i} X_i \boldsymbol{e}_{n_i} \tag{7}$$

Here for notational simplicity all samples belonging to class ω_i are denoted in the matrix form as $X_i = [x_{i1} \ldots x_{in_i}]$.

When the number of samples available for a given class is less than $d + 1$, where d is the dimensionality, the sample covariance becomes ill conditioned, i.e., the inverse does not exist. In practice, a robust sample covariance requires many more samples than $d + 1$ because the number of parameters to estimate in a covariance matrix increases as the square of the dimensionality. This phenomenon is known as *the curse of dimensionality* [23].

Although the research in covariance estimators using a limited number of samples with high dimensionality has a long history with relatively well-established techniques, two main approaches dominate the field. These are, regularized discriminant analysis (RDA) [24] and empirical Bayes estimators [25]. RDA considers the mixture of sample and pooled covariance and an identity matrix as an estimator, with their weights empirically estimated by cross-validation. On the other hand, the Bayesian approach defines a pair of conjugate prior distributions over the sample and true covariance matrices, and uses the mean of the resulting posterior distribution as an estimator. In RDA, multiple samples from each class are required to estimate the mixing weights by cross-validation, and thus to estimate the covariance matrix, whereas in the Bayesian approach, the covariance estimator is a function of the parameters of the prior distribution, which are estimated using samples of the known classes.

Creating a new class for each detected novelty and defining the class by its mean and covariance matrix form the core component of the proposed approach. The Bayesian approach assumes a common prior for all classes (known and unknown) and estimates the covariance matrix using the posterior mean. In that regard, the use of the Bayesian approach makes intuitive sense in the nonexhaustive setting, mainly because we assume that there is a common pattern among the class distributions of all classes and that it can be captured with known classes only, provided that a sufficiently large number of them are available for training. In the bacterial detection problem, although our training dataset represents only a small portion of a potentially very large number of bacterial serovars, unlike traditional machine learning problems, the number of available classes is still large enough to allow for a reasonably robust estimation of the prior distribution. This facilitates the estimation of the covariance matrices for the new classes, which is especially important when defining a class for the first time using the sample detected as novelty. In the following section, we will present a special family of conjugate priors for covariance estimation under the Bayesian framework.

2.3. Family of Wishart and Inverted-Wishart Conjugate Priors

The assumption of Gaussian samples, i.e., $\omega_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, implies that, the sample covariance matrices S_i, $i = \{1, \ldots, K\}$, where K is the number of known classes, are mutually independent with $f_i S_i \sim W(\boldsymbol{\Sigma}_i, f_i)$. Here $f_i = n_i - 1$ and $W(\boldsymbol{\Sigma}_i, f_i)$ denotes the Wishart distribution with f_i degrees of freedom and a parameter matrix $\boldsymbol{\Sigma}_i$. The inverted Wishart distribution is conjugate to the Wishart distribution and thus provides a convenient prior for $\boldsymbol{\Sigma}_i$.

We assume that $\boldsymbol{\Sigma}_i$ is distributed according to an inverted Wishart distribution with m degrees of freedom as:

$$\boldsymbol{\Sigma}_i \sim W^{-1}((m - d - 1)\Psi, m), \; m > d + 1. \tag{8}$$

The scaling constant $(m - d - 1)$ before Ψ is chosen to satisfy $E\{\boldsymbol{\Sigma}_i\} = \Psi$. Under this setting, the posterior distribution of $\boldsymbol{\Sigma}_i | \{S_1, \ldots, S_K\}$ is obtained as described by Anderson [26]:

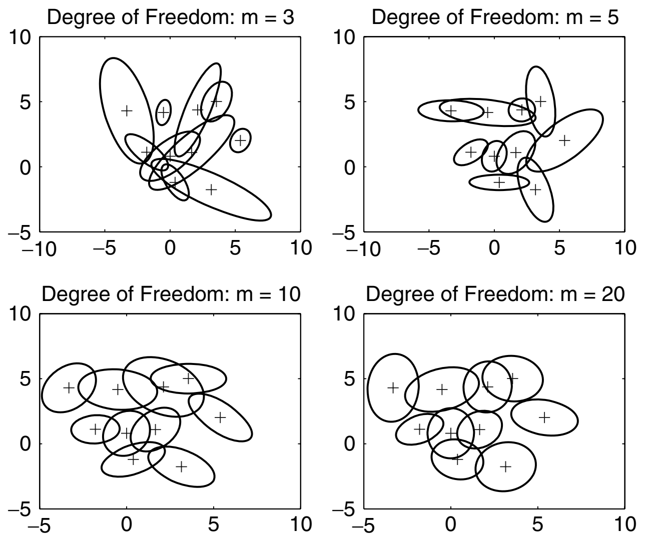
$$\boldsymbol{\Sigma}_i | (S_1, \ldots, S_K) \sim W^{-1} \; (f_i S_i + (m - d - 1)\Psi, \; f_i + m). \tag{9}$$

The mean of this posterior distribution is

$$\widehat{\boldsymbol{\Sigma}}_i(\Psi, m) = \frac{f_i}{f_i + m + d - 1} S_i + \frac{m - d - 1}{f_i + m + d - 1} \Psi. \tag{10}$$

Under squared-error loss, the posterior mean is the Bayes estimator of $\boldsymbol{\Sigma}_i$. The estimator is a weighted average of S_i and Ψ, and it shifts toward S_i for large f_i and approaches Ψ for large m. For a class with just one sample, the estimator returns Ψ, which implies that no matter what the dimensionality is, a nonsingular covariance estimate can be obtained using this estimator, provided that Ψ is nonsingular. The estimator is a function of Ψ and m, which are the parameters of the inverted Wishart prior for $\boldsymbol{\Sigma}_i$. The closed-form estimates for Ψ and m do not exist. Greene and Rayens [25] suggest estimating Ψ by the unbiased and consistent estimate S_p, i.e., the pooled covariance, and maximizing the marginal likelihood of S_i for $m > d + 1$ numerically to estimate m. In this study, we set Ψ to S_p but estimate m to maximize the classification accuracy for the known classes by cross-validating over the

UNCORRECTED PROOFS



Color Figure - Online only

Fig. 2 Simulated classes illustrating the impact of the degree of freedom, m, in the inverted Wishart distribution.

training samples. Here, S_p is the pooled covariance matrix defined by

$$S_p = \frac{f_1 S_1 + f_2 S_2 + \cdots + f_K S_K}{N - K}, \tag{11}$$

where N is the total number of samples available in the training dataset.

Figure 2 illustrates the effect of m on the modeling of the classes. In this example, ten classes are generated. Their mean vectors are chosen randomly from a normal distribution with mean at the origin and covariance matrix equal to 10**I**, where **I** denotes the 2D identity matrix. The covariance matrices of the classes are obtained from an inverted Wishart distribution with the first parameter $\Psi = 0.3\mathbf{I}$, which is designed to yield relatively circular distributions. The parameter m, the degree of freedom, takes the values 3, 5, 10, and 20, respectively, in the four cases shown in Fig. 2. As m increases, initially the classes transform from more elongated distributions to more circular ones but only slight changes in shape and orientation are observed beyond a certain m value.

So far, we have discussed a framework for detecting novelties in real time based on ML evaluation of samples using known classes. Our approach employs a pair of conjugate Wishart priors to estimate the covariance matrices of known classes and detects novelties by thresholding the ML evaluated with known classes. We will refer to this approach as *ML-Wishart*. In traditional novelty detection algorithms, no immediate action is taken for novelties. Once detected, they are left for a follow-up analysis. However, novelties originate from classes of informational value which were not known at the time of training. Pooling novelties showing similar characteristics to individual clusters may potentially recover some of these classes, and as more classes of informational value are introduced, the training dataset becomes more representative. This helps improve the predictive performance of the system not only for detecting novelties but also for classifying future samples of newly discovered classes. Our proposed algorithm, referred as *BayesNoDe*, combines novelty detection with new class discovery and will be presented next.

2.4. Real-time Discovery of New Classes

As formulated in Eq. (5), a new sample $x^* \in \Re^d$ is detected as a novelty if $\min_{\{i:\omega_i \in \Delta\}} g_i(x^*) > \gamma$. In other words, if the negative log-likelihoods of known classes given x^* are all greater than the designated threshold γ, then the sample is considered a novelty.

When a new sample is detected as a novelty, a new class is generated and defined by the parameters, $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *****where $\boldsymbol{\mu}$ is the mean vector of this class and $\boldsymbol{\Sigma}$ is the covariance matrix, both of which are not known. With just one sample, since S is not defined and $f = 0$, the posterior mean in Eq. (10) is equivalent to Ψ and thus the Bayesian estimator for $\boldsymbol{\Sigma}$ becomes $\widehat{\boldsymbol{\Sigma}} = \Psi$. The mean vector, $\boldsymbol{\mu}$ is estimated by $\hat{\boldsymbol{\mu}} = x^*$, i.e. the sample itself, which follows from Eq. (7).

The set of known classes is augmented with this new class. So for the next sample available, the decision function in Eq. (5) is evaluated for classes known initially as well as for the newly created classes. If the sample is detected as a novelty, the above procedure is repeated to generate another class. Otherwise, if the sample is classified into one of the existing classes, then we check for the class that minimizes the negative log-likelihood. If the sample is assigned to a previously discovered class, then we update the class parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ using Eqs. (7) and (10) for that class. Since, there is more than one sample available now, $\hat{\boldsymbol{\Sigma}}$ becomes a mixture of the sample covariance and Ψ. If, on the other hand, the sample is assigned to a class known initially, then no class update is necessary.

Pseudocode for the algorithm capable of detecting novelties and discovering new classes is presented in Algorithm 1.

It is important to note that the Gaussianity assumption made throughout the study is not much of a limitation for either the set of initially known classes or the newly discovered ones, for the following reason. The theory of finite mixture models [27] states that given enough components and under fairly weak assumptions, a mixture model can approximate a given density arbitrarily closely, allowing great flexibility. In other words, even if the initially known classes are not Gaussian, the class-conditional distributions can still be estimated arbitrarily closely, using a mixture of

UNCORRECTED PROOFS

Algorithm 1 BayesNoDe: An Algorithm for Bayesian Novelty Detection and Class Discovery

```
INITIALIZATION
Initialize $\Delta$ with the set of initially known classes
$K \Leftarrow |\Delta|$ {Define $K$ as the number of initially known classes}
for each class $i$ in the set of known classes do
    Estimate $S_i$, $\hat{\boldsymbol{\mu}}_i$, $\hat{\boldsymbol{\Sigma}}_i$
end for
$\Psi \Leftarrow S_p$ {estimate $\Psi$ by the pooled covariance matrix}
$m \Leftarrow mOpt$ {estimate m by cross-validation from a predefined range of m values}
$c \Leftarrow 0$ {initialize the counter for the newly created classes}
ONLINE DETECTION & DISCOVERY
while there exists a new sample $x^*$ do
    for $i$ from 1 to $(K + c)$ do {each $i$ in the current set of known classes}
        Compute $g_i(x^*)$ {compute the negative log-likelihood for class $i$}
    end for
    $j \Leftarrow argmin_i \{g_i(x^*)\}$ {find the class that minimizes the negative log-likelihood for $x^*$}
    if $g_j(x^*) > \gamma$ then
        Increment $c$, generate a new class $\omega_{K+c}$
        Mark $x^*$ as novelty and assign it to the new class $\omega_{K+c}$
        $\hat{\boldsymbol{\mu}}_{K+c} \Leftarrow x^*$ {initialize the mean vector}
        $\hat{\boldsymbol{\Sigma}}_{K+c} \Leftarrow \Psi$ {initialize the covariance matrix with $\Psi$, note $S_i = 0$}
    else {$g_j(x^*)$ is less than the threshold}
        if $j > K$ then {$\omega_j$ is a newly generated class}
            Mark $x^*$ as novelty and assign it to class $\omega_j$
            Update $S_j$, $\hat{\boldsymbol{\mu}}_j$, $\hat{\boldsymbol{\Sigma}}_j$
        else {$j$ is an initially known class}
            Assign $x^*$ to class $\omega_j$, it is not a novelty
            NO update is done to class parameters
        end if
    end if
end while
```

Gaussians. A mixture of Gaussian subclasses can be learned for each class data through a process involving expectation maximization [28] and model selection. Once the Gaussian subcomponents are identified for each class data, the proposed approach can be implemented at the subclass level by considering each subclass as an independent Gaussian class.

Similarly, when discovering new classes, only clusters with Gaussian patterns will be created for novelties. However, true classes with informational value can still be recovered by grouping newly discovered clusters under a higher-level class using domain/expert knowledge.

Next, we present an illustrative example demonstrating the proposed algorithm detecting novelties and creating classes with a 2-D simulated dataset. Similar to our previous example we generate ten classes with their covariance matrices obtained from an inverted Wishart distribution with parameters $\Psi = 0.3\mathbf{I}$ and $m = 10$ and their mean vectors are chosen randomly from a normal distribution with mean at the origin and covariance matrix equal to $10\mathbf{I}$. Here, $\mathbf{I}$ denotes the 2-D identity matrix.

Panel (a) of Fig. 3 shows all ten classes. Known classes are depicted by solid lines, unknown classes by dashed lines. The square sign locates the mean of each class. The ellipses represent the class boundaries as defined by the three standard deviation distance from the class means. A total of 80 samples are generated from the ten classes: 5 from each of the known classes and 20 from each of the unknown classes. Test samples are classified sequentially using the proposed BayesNoDe algorithm. Panels (b)–(d) of Fig. 3 illustrate cases where 16/80, 48/80, and 80/80 samples are classified, respectively. Red solid lines indicate the *estimated* distribution contours for newly discovered classes in each subfigure with the diamond signs locating their estimated means. The blue + signs and red × signs in each subfigure show the samples classified to known and unknown classes, respectively. Panel (e) of Fig. 3 demostrates novelty detection using ML-Wishart, i.e., with a fixed set of classes in the training dataset, and panel (f) of Fig. 3 illustrates the case where no novelty detection is performed at all. In these two figures, the samples marked with red circles indicate samples from the unknown classes misclassified as known. Also in panel (e) of Fig. 3, blue solid lines correspond to $g(z) = \gamma$ as defined in Eq. (5) and indicate the classification boundaries for the unknown samples.

As panels (b)–(d) of Fig. 3 demonstrate, the algorithm gradually recovers the unknown classes as more test samples are introduced, converging to almost ideal distributions after all 80 test samples are classified.

Comparing panels (d) and (e) of Fig. 3 shows the improvement achieved by the BayesNoDe algorithm over the ML-Wishart as a result of the dynamically updated training dataset. When no novelty detection is used, all samples are misclassified as illustrated in panel (f) of Fig. 3.

3. EXPERIMENTS

3.1. Experiment 1: Bacteria Detection

A total of 28 strains (subclasses) from five different bacteria species were considered in this study. The species available are *E. coli, Listeria, Salmonella, Staphylococcus and Vibrio*. Table 1 shows the list of 28 strains from 5 species considered in this study together with the number of samples collected for each one using the BARDOT system described in Section 1. In our experiments, we treated each strain as a separate class and used the number of samples listed in Table 1 from each class for training.

UNCORRECTED PROOFS

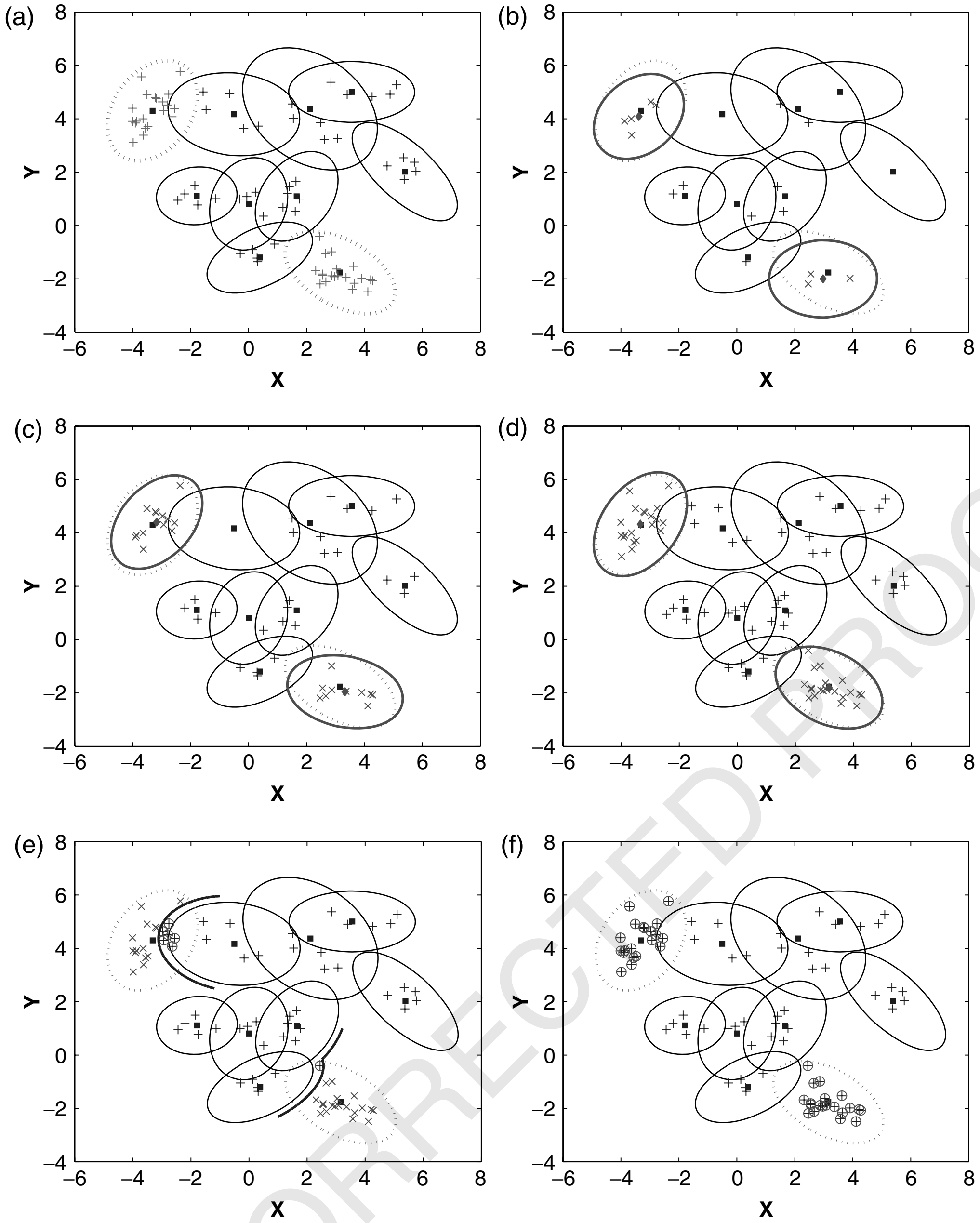


Fig. 3 Illustration of the proposed algorithm with an artificial dataset. Pink dashed lines indicate unknown classes with 20 samples each. Black solid lines indicate known classes with five samples each. Red solid lines indicate newly discovered classes. Mean vectors for the original classes are depicted by the blue squares. Mean vectors for the newly discovered classes are depicted by the red diamonds. Blue + signs, indicate samples from known classes, red × signs indicate samples from unknown classes. Encircled + signs indicate samples from unknown classes incorrectly classified as known. (a) Classes with dashed lines are assumed unknown; (b) 16 out of 80 samples are classified; (c) 48 out of 80 samples are classified; (d) all samples are classified—BayesNoDe; (e) all samples are classified—ML-Wishart; (f) all samples are classified—no novelty detection.

Color Figure - Online only

UNCORRECTED PROOFS

3.1.1. Feature selection

Scatter patterns of the bacteria are characterized by a total of 50 features involving moment invariants and Haralick texture descriptors. Details of the feature extraction process are available in Ref. [29].

3.1.2. Classifier design

The classification methods considered are the SVDD [19], which is the benchmark technique for detecting anomalies and novelties, ML using common covariance (ML-Common), ML using common covariance with

Table 1. The 28 subclasses from five species (classes) considered in this study.

Species	ID	Strain (Subclass)	# Samples
E. Coli sp.	1	O25:K98:NM ETEC	67
	2	O78:H11 ETEC	58
	3	O157:H7 01	64
	4	O157:H7 6458	87
	5	O157:H7 G5295	68
	6	K12 ATCC 29425	65
Listeria spp.	7	*L. innocua* F4248	59
	8	*L. ivanovii* 19119	81
	9	*L. monocytogenes* 19118 (4e)	94
	10	*L. monocytogenes* 7644 (1/2c)	91
	11	*L. monocytogenes* V7 (1/2a)	98
	12	*L. welshimeri* 35897	47
Salmonella spp.	13	*S. Typhimurium* (Copenhagen)	95
	14	*S. Enteritidis* 13096	89
	15	*S. Enteritidis* PT28	90
	16	*S. Tennessee* 825-94	78
Staphylococcus spp.	17	*S. aureus* 13301	46
	18	*S. aureus* PS103	50
	19	*S. aureus* S-41	67
	20	*S. epidermidis* PS302	31
	21	*S. epidermidis* 35547	45
	22	*S. hyicus* T6346	69
Vibrio spp.	23	*V. alginolyticus* CECT521	88
	24	*V. campbellii* CECT523	71
	25	*V. cincinnatiensis* CECT4216	89
	26	*V. hollisae* CECT5069	79
	27	*V. orientalis* CECT629	96
	28	*V. parahaemolyticus* CECT511	92
		Total	2054

Notes: The last column lists the number of samples collected for each strain using the Bardot system.

simulated subclass generation (MLS) [21], ML with the covariance matrix estimated by the posterior mean of the inverted-Wishart distribution (ML-Wishart), and the BayesNoDe algorithm. The ML classifier using sample covariance is not considered here, because sample covariances were ill conditioned for most classes.

As explained in Sections 2.1 and 2.2, the general idea of ML classifiers is based on the ML decision function in Eq. (3) and works according to the formulation in Eq. (4). ML-Wishart and ML-Common are the special cases of ML. They differ in estimating the covariance matrices of the training classes. Corresponding mean vectors, $\boldsymbol{\mu}_i$, are all calculated by (7). More specifically, ML-Common implements (5), where $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ for all i, and $\boldsymbol{\Sigma}$ represents the common covariance matrix estimated by the average of the sample covariances. As described in Ref. [21], MLS extends ML-Common by simulating the space of all classes. This approach assumes a Gaussian prior for the mean vectors, and its parameters are estimated using the estimates of the mean vectors for each class. Lastly, for the proposed ML-Wishart and BayesNoDe, the covariance matrices are estimated for each class using the posterior mean defined in Eq. (10). The parameters m and Ψ are estimated as described in Section 2.3.

As for the SVDD algorithm, optimization involves two sets of parameters. These are C, the cost of leaving a training sample outside the support, and σ, the width of the Gaussian radial basis function (RBF) kernel. These parameters are estimated by 10-fold cross-validation at the class level. When optimizing parameters for a given class, the training samples of the given class are considered positive and the samples of remaining classes are considered negative. At each fold of the cross-validation algorithm, SVDD is trained using positive samples only but tested on both positive and negative samples. The parameter set (C_*, σ_*) that optimizes the area under the receiver operating characteristic (ROC) curve is chosen as the optimum set for the given class. This process is repeated for all classes.

3.1.3. Classifier validation and evaluation

Since the training dataset is nonexhaustive, the goal is to design a classifier that accurately detects samples of known classes as known and those of unknown classes as novelty. In this framework, classifiers can be more properly evaluated using ROC curves. Here sensitivity is defined as the number of samples from known classes classified as known divided by the total number of samples from known classes. Specificity is defined as the number of samples from unknown classes detected as novelty, divided by the total number of samples from unknown classes. Multiple sensitivity and specificity values are obtained for each classifier to plot the ROC curves. For the ML-based approaches, different operating points are obtained by varying the threshold γ in Eq. (5). For SVDD, the distances from the center of each class is normalized by the radius of the corresponding sphere. For a new sample, the minimum of the normalized class distances is computed and thresholded to obtain different operating points.

To evaluate the classifiers the 2054 samples are randomly split into two sets, as train and test, with 80% of the samples going into the training set and the remaining 20% into the test. Stratified sampling is used to make sure that each subclass is represented in both sets. This process is repeated

UNCORRECTED PROOFS

ten times to obtain ten different pairs of train-test sets. Then, one subclass from each of the five bacteria species is randomly selected, so a total of 5 subclasses out of the 28 available are identified. All samples of these five classes are removed from the training datasets making these classes unknown. The classifiers are trained with the resulting nonexhaustive training sets and tested on the corresponding test sets. For each data split, the area under the ROC curve, i.e., Az value is computed. The Az values averaged over the ten different train-test splits are recorded along with the standard deviation.

3.1.4. Results and analysis

To account for the possible bias introduced by the set of removed classes the above process is repeated 20 times each time removing a randomly selected set of five classes from the training set. Each such repetition involves running the same experiment with a different nonexhaustive subset of the original data. Az values achieved for each classifier are included in Table 2 for all 20 experiments. As described earlier, these values are the average of the ten runs each executed with a different train-test split and the values in parantheses indicate standard deviations. The mean Az values across all 20 runs are listed in Table 3. These results clearly favor the proposed *BayesNoDe* algorithm, which generated the best area under the curve (AUC) in all 20 repetitions. Standard deviations indicate that the differences are statistically significant in most of the 20 experiments. The BayesNoDe algorithm is an extension of the ML-Wishart algorithm, both of which are proposed in this study. ML-Wishart ranked second, but the results indicate that creating new classes and augmenting the set of known classes with these new classes makes a considerable impact on the prediction accuracy of the classifier and gives the BayesNoDe algorithm a significant advantage over the ML-Wishart. SVDD ranked third along with ML-Common and MLS.

Next, we picked four sample cases out of the 20 using the overall Az values achieved by the classifiers as the selection criteria. Largest Az value among all 20 repetitions is recorded in repetition 10 (Fig. 4, panel (a)). Repetitions 13 and 16 represent two average cases (Fig. 4, panels (b) and (c)). Repetition 20 is included to show results for a relatively poor case (Fig. 4, panel (d)). The ROC curves corresponding to the proposed BayesNoDe algorithm dominate the other curves in all cases. We also analyzed the classification accuracy of the known samples and observed that the known samples are assigned to classes with over 95% accuracy across all operating points for all four cases considered here. These results indicate that the proposed approach not only performs well in identifying samples of the unknown classes as novelties but yields promising results in classifying samples of the known classes as well.

3.2. Experiment 2: Letter Recognition

To show that improvements achieved by the proposed BayesNoDe algorithm is not specific to the Bacterial detection application that motivated this research, we used the benchmark letter recognition dataset from the UCI repository [30] for further validation of the proposed approach for novelty detection. This dataset is mainly selected for its large number of classes. The dataset contains 20 000 samples for 26 classes, one for each letter of the alphabet. Each sample is characterized using 16 features.

This dataset is different than the bacteria detection dataset in that, it is not susceptible to the curse of dimensionality as much. There is an average of 770 samples for each class as opposed to an average of 80 samples for each bacteria subclasses. The dimensionality of the data ($d = 16$) is also much lower than the 50 features used in the bacteria detection dataset.

We followed an experiment design similar to the bacteria detection experiment. The 20 000 samples are randomly splitted into train and test sets, with 80% of the samples going into the training set and the remaining 20% in the test. Stratified sampling is used to make sure each class is represented in both the training and the test sets. This process is repeated five times to obtain five different pairs of train-test sets. Then, five classes are randomly selected and their samples are removed from the training datasets. The classifiers are trained with the resulting nonexhaustive training sets, and tested on the corresponding test sets. For each case, Az value is computed. The Az values averaged over the five different train-test splits are recorded along with the standard deviation.

3.2.1. Classifier design

The same set of classifiers considered in Experiment 3.1 are also considered here. SVDD and MLS are trained the same way as in Experiment 3.1. For the ML-based classifiers, since classes contain a relatively larger number of samples, a single Gaussian might not fit class data well. In this case, as discussed in Section 2.4, the actual class distributions can be modeled more effectively using a mixture of Gaussians. We fit up to five components for each class distribution using standard expectation maximization algorithm [28] with the optimum number of components selected using the Bayesian Information Criterion [31]. Once mixture models are obtained, each subclass is considered as an independent class and all ML-based classifiers are run with the new set of known classes. On the average for each class data mixture fitting returned three subclasses.

3.2.2. Results and analysis

The experiment is repeated twice each time removing a randomly selected set of five classes from the training set.

UNCORRECTED PROOFS

Table 2. *Az* values averaged over ten iterations for all 20 experiments run with the bacteria dataset.

Repetition #	1	2	3	4	5	6	7	8	9	10
BayesNoDe	0.97 (0.01)	0.92 (0.01)	0.98 (0.01)	0.92 (0.01)	0.93 (0.01)	0.95 (0.01)	0.98 (0.01)	0.96 (0.01)	0.95 (0.01)	0.99 (0.01)
ML-Wishart	0.95 (0.01)	0.88 (0.01)	0.96 (0.00)	0.90 (0.01)	0.89 (0.01)	0.92 (0.01)	0.95 (0.01)	0.94 (0.01)	0.94 (0.01)	0.98 (0.01)
ML-Common	0.88 (0.01)	0.71 (0.01)	0.90 (0.01)	0.82 (0.01)	0.79 (0.01)	0.83 (0.01)	0.83 (0.01)	0.87 (0.01)	0.89 (0.01)	0.94 (0.00)
MLS	0.87 (0.01)	0.80 (0.01)	0.82 (0.01)	0.81 (0.01)	0.80 (0.01)	0.84 (0.01)	0.92 (0.01)	0.86 (0.01)	0.78 (0.01)	0.85 (0.01)
SVDD	0.87 (0.01)	0.77 (0.02)	0.90 (0.02)	0.81 (0.03)	0.76 (0.03)	0.81 (0.02)	0.86 (0.02)	0.84 (0.02)	0.86 (0.02)	0.89 (0.01)

Repetition #	11	12	13	14	15	16	17	18	19	20
BayesNoDe	0.91 (0.01)	0.98 (0.01)	0.97 (0.01)	0.93 (0.01)	0.89 (0.01)	0.95 (0.01)	0.95 (0.01)	0.82 (0.01)	0.92 (0.01)	0.88 (0.01)
ML-Wishart	0.87 (0.01)	0.96 (0.01)	0.94 (0.01)	0.88 (0.01)	0.85 (0.01)	0.92 (0.01)	0.91 (0.01)	0.79 (0.01)	0.87 (0.01)	0.85 (0.01)
ML-Common	0.80 (0.01)	0.90 (0.01)	0.88 (0.01)	0.76 (0.01)	0.78 (0.03)	0.83 (0.01)	0.81 (0.01)	0.72 (0.01)	0.77 (0.01)	0.81 (0.02)
MLS	0.78 (0.01)	0.82 (0.01)	0.87 (0.01)	0.81 (0.01)	0.86 (0.01)	0.85 (0.01)	0.84 (0.01)	0.80 (0.01)	0.74 (0.01)	0.84 (0.01)
SVDD	0.83 (0.01)	0.90 (0.01)	0.82 (0.06)	0.76 (0.03)	0.77 (0.03)	0.83 (0.01)	0.81 (0.01)	0.73 (0.03)	0.81 (0.02)	0.80 (0.02)

Notes: A set of five subclasses is randomly selected and considered unknown during each of the 20 experiments. BayesNoDe results in the best AUC values for all 20 experiments. Values in parentheses indicate standard deviations.

Table 3. Average *Az* values over 20 experiments.

Methods	Avg. AUC
BayesNoDe	0.94 (0.05)
ML-Wishart	0.91 (0.06)
ML-Common	0.83 (0.04)
MLS	0.83 (0.04)
SVDD	0.82 (0.05)

The ROC curves are plotted in panels (a) and (b) of Fig. 5. For this experiment SVDD seems to model the data well and becomes the sole competitor to BayesNoDe and ML-Wishart. ML-Wishart performs slightly better than SVDD. The detection accuracy of BayesNoDe is almost perfect and as the error bars indicate the improvements achieved over other methods are statistically significant.

4. CONCLUSION

The current research is mainly motivated by the impracticality of the exhaustiveness assumption in defining the number of classes in a training dataset. In this study, we propose a novelty detection scheme, which makes two distinct contributions: novelty detection and modeling. Evaluated samples are identified either as *novelty* or classified into one of the known classes.

The proposed technique is based on the Bayesian modeling of the distribution of the classes via a pair of conjugate Wishart priors. The resultant posterior distribution is used to obtain robust estimates of the covariance matrices of the class-conditional distributions for known as well as newly created classes with limited number of samples. Novelties are detected by evaluating the ML with known classes. Samples are labeled as novelty or known based on whether the ML is smaller or larger than a predefined threshold. Effective modeling of the prior distribution of the classes in this approach requires a relatively large number of known classes. Our research is motivated by a biodetection application. We have performed experiments with a 28-class bacteria dataset and presented results favoring the proposed algorithm over the state-of-the-art for novelty detection. Additional experiments are performed with a 26-class benchmark dataset to further validate the proposed approach and show that improvements are not application specific.

Future research will focus on modeling of the known classes by nonparametric Bayesian approaches involving Gaussian processes, which we believe will allow for more robust modeling of the classes and will improve

UNCORRECTED PROOFS

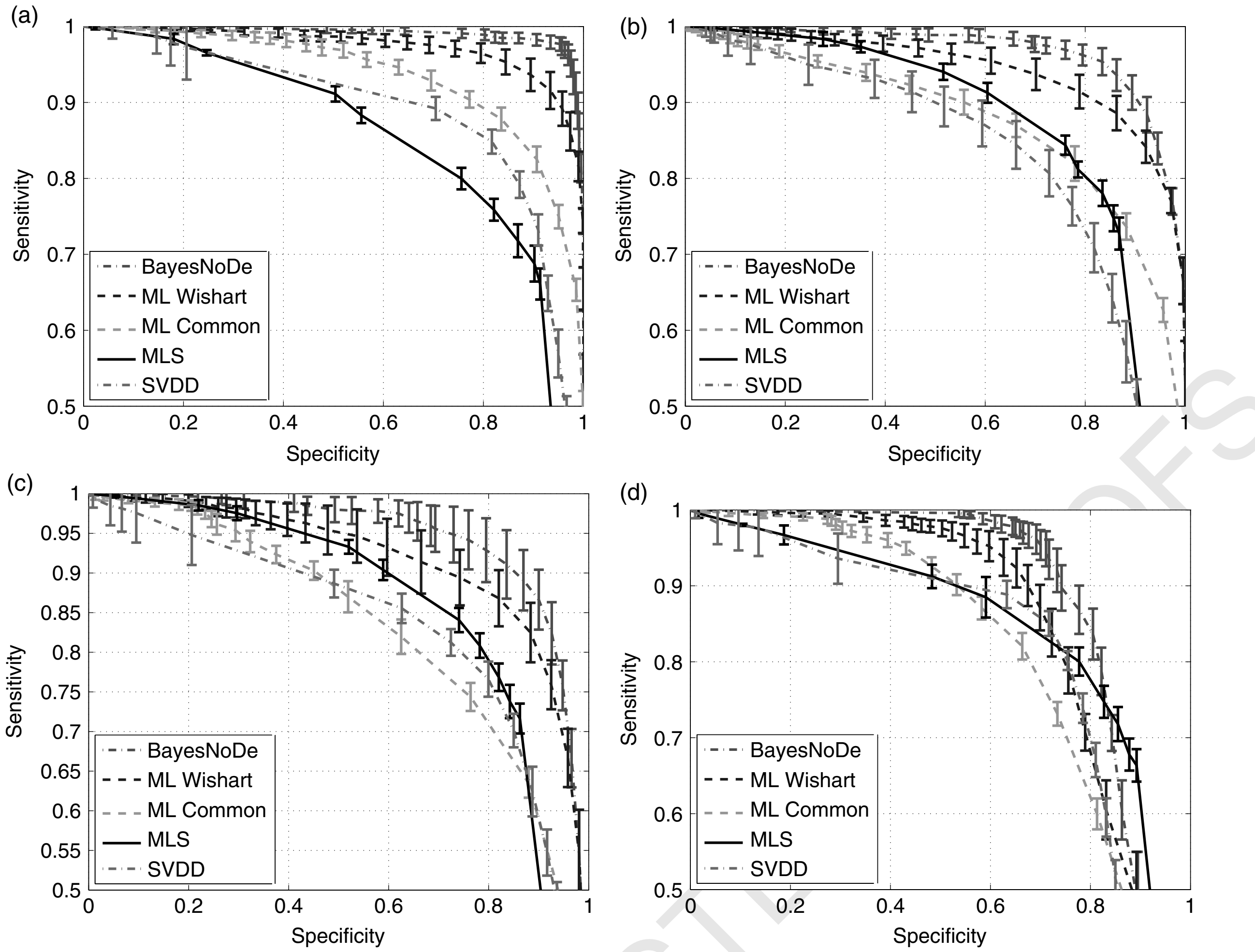


Fig. 4 (a) Repetition 10. Removed subclass ids are: 6, 12, 15, 18, and 27. (b) Repetition 13. Removed subclass ids are: 1, 11, 16, 22, and 23. (c) Repetition 16. Removed subclass ids are: 2, 8, 16, 21, and 28. (d) Repetition 20. Removed subclass ids are: 5, 8, 15, 22, and 26.

Color Figure - Online only

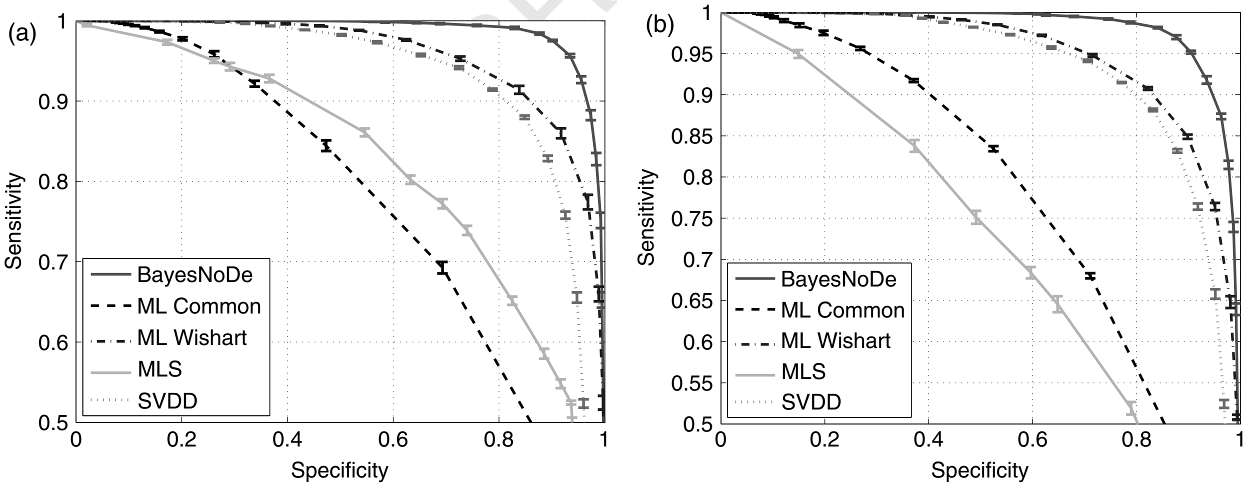


Fig. 5 (a) Removed subclass ids are: 7, 9, 12, 14, and 24. (b) Removed subclass ids are: 2, 9, 11, 12, and 22.

Color Figure - Online only

UNCORRECTED PROOFS

prediction performance of the proposed novelty detection algorithm.

ACKNOWLEDGMENT

The project was supported by Grant Number 5R21AI085 531-02 from the National Institute of Allergy and Infectious Diseases (NIAID). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Allergy and Infectious Diseases or the National Institutes of Health.

A part of this research was supported through a cooperative agreement with the Agricultural Research Service of the US Department of Agriculture project number 1935-42000-035 and the Center for Food Safety and Engineering at Purdue University.

REFERENCES

[1] E. Klein, D. L. Smith, and R. Laxminarayan, Hospitalizations and deaths caused by methicillin-resistant Staphylococcus aureus, United States, 1999–2005, Emerg Infect Dis 13(12) (2007), 1840–1846. PMID: 18258033 [Online]. http://www.ncbi.nlm.nih.gov/pubmed/18258033.

[2] J. Heaton, and K. Jones, Microbial contamination of fruit and vegetables and the behaviour of enteropathogens in the phyllosphere: a review, J Appl Microbiol 104(3) (2008), 613–626 [Online]. http://dx.doi.org/10.1111/j.1365-2672.2007.03587.x.

[3] M. T. Jay, M. Cooley, D. Carychao, et al., Escherichia coli O157:H7 in feral swine near spinach fields and cattle, central california coast, Emerg Infect Dis 13(12) (2007), 1908–1911. PMID: 18258044 [Online]. http://www.ncbi.nlm.nih.gov/pubmed/18258044.

[4] P. Gerner-Smidt, and J. M. Whichard, Foodborne disease trends and reports, Foodborne Pathogens Dis 6(1) (2009), 1–5. PMID: 19182964 [Online]. http://www.ncbi.nlm.nih.gov/pubmed/19182964.

[5] CDC, Multistate outbreak of Salmonella infections associated with peanut butter and peanut butter–containing products–United States, 2008–2009, Morb Mortal Wkly Rep 58(4) (2009), 85–90 [Online]. http://www.cdc.gov/mmwR/preview/mmwrhtml/mm5804a4.htm.

[6] B. Swaminathan, and P. Gerner-Smidt, "The epidemiology of human listeriosis," Microbes and Infection vol. 9 (2007), [Online]. Available: (10), 1236–1243. Aug. http://www.sciencedirect.com/science/article/B6VPN-4NNN0J0-7/2/3dc162f9%e384834c47a0998aa083741d.

[7] F. Ligler, C. Taitt, L. Shriver-Lake, K. Sapsford, Y. Shubin, and J. Golden, Array biosensor for detection of toxins, Anal Bioanal Chem 377(3) (2003), 469–477 [Online]. http://dx.doi.org/10.1007/s00216-003-1992-0.

[8] D. V. Lim, J. M. Simpson, E. A. Kearns, and M. F. Kramer, Current and developing technologies for monitoring agents of bioterrorism and biowarfare, Clin Microbiol Rev 18(4) (2005), 583–607.

[9] L. Manning, R. Baines, and S. Chadd, Deliberate contamination of the food supply chain, Brit Food J 107(4) (2005), 225–245 [Online]. http://www.emeraldinsight.com/10.1108/00070700510589512.

[10] D. A. Relman, E. Choffnes, and S. M. Lemon, In search of biosecurity, Science 311(5769) (2006), 1835 [Online]. http://www.sciencemag.org.

[11] B. Rajwa, M. Venkatapathi, K. Ragheb, P. P. Banada, E. D. Hirleman, T. Lary, and J. P. Robinson, Automated classification of bacterial particles in flow by multiangle scatter measurement and support vector machine classifier, Cytometry. Part A: J Int Soc Anal Cytol 73(4) (2008), 369–379. PMID: 18163466 [Online]. http://www.ncbi.nlm.nih.gov/pubmed/18163466.

[12] T. Lane, and C. E. Brodley, Temporal sequence learning and data reduction for anomaly detection, ACM Trans Inform Syst Security 2 (1998), 150–158.

[13] E. Eskin, A. Arnold, M. Prerau, L. Portnoy, and S. Stolfo, A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data, In Applications of Data Mining in Computer Security, D. Barbará, and S. Jajodia, eds. Kluwer,● 2002. AQ2

[14] D. Pelleg, and A. Moore, Active learning for anomaly and rare-category detection, Advances in Neural Information Processing Systems, Vol. 18, MIT Press, December ● 2004, 1073–1080. AQ3

[15] J. Theiler, and D. M. Cai, Resampling approach for anomaly detection in multispectral images, In Proc. SPIE, 2003. 230–240.

[16] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, Support vector method for novelty detection, Adv Neural Inform Process Syst 12(●) (2000). AQ4

[17] J. Muoz-Mar, L. Bruzzone, and G. Camps-Valls, A support vector domain description approach to supervised classification of remote sensing images, IEEE Trans Geosci Remote Sens 45(8) (2008), 2683–2692.

[18] E. J. Spinosa, and A. C. Carvalho, Support vector machines for novel class detection in bioinformatics, Genet Mol Res 4(3) (2005), 608–615.

[19] D. M. J. Tax, and R. P. W. Duin, Support vector domain description, Pattern Recogn Lett 20(11–13) (1999), 1191–1199.

[20] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson, Estimating the support of a high-dimensional distribution, Neural Comput 13(7) (2001), 1443–1471.

[21] M. Dundar, D. Hirleman, A. K. Bhunia, J. P. Robinson, and B. Rajwa, Learning with a nonexhaustive training dataset. A case study: detection of bacteria cultures using optical-scattering technology, In Proceedings of the Fifteenth Annual SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 2009.

[22] K. Fukunaga, Introduction to Statistical Pattern Recognition, San Diego, CA, Academic Press, 1990.

[23] R. E. Bellman, Dynamic Programming, Princeton University Press,● 1957. AQ5

[24] J. H. Friedman, Regularized discriminant analysis, J Am Stat Assoc 84(405) (1989), 165–175.

[25] T. Greene, and W. Rayens, Partially pooled covariance matrix estimation in discriminant analysis, Commun Stat Theory Meth 18(10) (1989), 3679–3702.

[26] T. W. Anderson, An Introduction to Multivariate Statistical Analysis (3rd ed.), Wiley-Interscience,● 2003. AQ6

[27] G. McLachlan, and D. Peel, Finite Mixture Models, New York, NY, John Wiley & Sons, 2000.

UNCORRECTED PROOFS

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

[29] P. P. Banada, K. Huff, E. Bae, et al., Label-free detection of multiple bacterial pathogens using light-scattering sensor," Biosens Bioelectron 24(6) (2009), 1685–92. PMID: 18945607 [Online]. http://www.ncbi.nlm.nih.gov/pubmed/18945607.

[30] P. W. Frey, and D. J. Slate, Letter recognition using holland-style adaptive classifiers, Mach Learn 6(2●) (1991). AQ7

[31] R. E. Kass, and A. E. Raftery, Bayes factors, J Am Stat Assoc 90(430) (1995), 773–795.

UNCORRECTED PROOFS