Combining Dissimilarity-Based One-Class Classifiers

Elżbieta Pękalska, Marina Skurichina, and Robert P.W. Duin

ICT Group, Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology, The Netherlands
{e.pekalska,m.skurichina,r.p.w.duin}@ewi.tudelft.nl

Abstract. We address a one-class classification (OCC) problem aiming at detection of objects that come from a pre-defined target class. Since the non-target class is ill-defined, an effective set of features discriminating between the targets and non-targets is hard to obtain. Alternatively, when raw data are available, dissimilarity representations describing an object by its dissimilarities to a set of target examples can be used.
A complex problem can be approached by fusing information from a number of such dissimilarity representations. Therefore, we study both the combined dissimilarity representations (on which a single OCC is trained) as well as fixed and trained combiners applied to the outputs of the base OCCs, trained on each representation separately. An experiment focusing on the detection of diseased mucosa in oral cavity is conducted for this purpose. Our results show that both approaches allow for a significant improvement in performance over the best results achieved by the OCCs trained on single representations, however, concerning the computational cost, the use of combined representations might be more advantageous.

1 Introduction

Novelty detection problems arise in applications, where anomalies or outliers should be recognized. Given training examples, the goal is to describe the so-called target class such that resembling objects are accepted as targets and outliers (non-targets) are rejected. Such a detection has to be performed in an unknown or ill-defined context of alternative phenomena. Examples refer to health diagnostics, machine condition monitoring, industrial inspection or face detection. The target class is assumed to be well sampled and well defined. The alternative non-target (outlier) set is usually ill-defined: it is badly sampled (even not present at all) with unknown and hard to predict priors. If available, such non-targets might be structured in ways not represented in the training set. For such types of problems *one-class classifiers* (OCCs) may be very suitable [15, 10], as they are domain or boundary descriptors.

Since the non-target class is ill-defined, in complex problems, an effective set of features discrimination between targets and non-targets cannot be easily found. Hence, it seems appropriate to build a representation on the raw data.

F. Roli, J. Kittler, and T. Windeatt (Eds.): MCS 2004, LNCS 3077, pp. 122–133, 2004.

The dissimilarity representation, describing objects by their dissimilarities to the target examples, may be effective for such problems since it naturally protects the target class against unseen novel examples. Therefore, we will study dissimilarity representations to train one-class classifiers. Optimal representations and dissimilarity measures cannot be found if one of the classes is missing or badly sampled. On the other hand, when one analyzes a particular phenomenon, the model knowledge can be captured by various dissimilarity representations describing different problem characteristics. In this way, a problem is tackled from a wider perspective: each additional representation may incorporate useful information. Combining OCCs becomes, thereby, a natural technique needed for solving ill-defined (unbalanced) detection problems. Note, however, that standard two-class classifiers should be preferred if the non-target class is well represented.

Although such problems are often met in practice, representative standard datasets do not exist yet. They should be based on the raw data and various dissimilarity measures should be available. Our procedures here are not intended for general multi-class problems for which other, more suitable, techniques exist. Our methodology is applicable to difficult problems where the target examples are provided with or without additional outlier examples. For that reason, the effectiveness of the proposed procedures is illustrated with just a single, yet complex, application, i.e. the detection of diseased mucosa in oral cavity.

Two approaches are compared within this application. The first one focuses on combining dissimilarity representations into a single one, while the second approach considers a combiner operating on the outputs of the OCCs. This study extends results of our earlier work [9] devoted to usual classification tasks. Note, however, that OCCs do not directly estimate the posterior probabilities since they rely on information on a target class. OCCs output a sort of a signed distance to the boundary.

2 One-Class Classifiers for Dissimilarity Representations

Consider a representation set $R = \{p_1, p_2, \ldots, p_n\}$, which is a set of representative objects. $d(x, p_i)$ denotes a dissimilarity between the objects x and p_i, independently from their initial representations. In general, we do not require metric properties of d, since non-metric dissimilarities may arise when shapes or objects in images are compared; see e.g. [6]. The usefulness of d is judged by its construction and a fit to the problem; d should be relatively small for objects resembling each other in reality and large for objects that differ. Obviously, the non-negativity and reflexivity, i.e. $d(x, y) \geq 0$ and $d(x, x) = 0$ are taken as granted. Thereby, a dissimilarity representation (DR) of an object x is expressed as a vector $D(x, R) = [d(x, p_1),\ d(x, p_2), \ldots, d(x, p_n)]$. For a collection of training objects $T = \{t_1, t_2, \ldots, t_N\}$, it extends to a $N \times n$ dissimilarity matrix $D(T, R)$. In general, R might be a subset of T ($R \subseteq T$) or they might be distinct sets.

There are three principal learning approaches, referring to three interpretations of DRs, for which a particular methodology can be adapted. In the *pre-*

topological approach (I), the dissimilarity values are interpreted directly, hence they can be characterized in pretopological spaces [7,12], where the neighborhoods play a significant role. The *embedding* approach (II) builds on a spatial representation, i.e. an embedded pseudo-Euclidean configuration such that the dissimilarities are preserved [5,8]. In the *dissimilarity space* approach (III), one considers $D(x,R): \mathcal{X} \rightarrow \mathcal{R}^r$ as a data-depending mapping to the so-called dissimilarity space. In such a space, every dimension corresponds to a dissimilarity $D(\cdot,p_i)$ to a particular object $p_i \in R$. So, the dimensions convey a homogeneous type of information. The property that dissimilarities should be small for similar objects (belonging to the same class) and large for distinct objects, gives a possibility for a discrimination. Thereby, $D(\cdot,p_i)$ can be interpreted as an attribute.

Below, some exemplar one-class classifiers are described, which in practice rely on some proximity $f_{\text{prox}}(x,\omega_T)$ of an object x to the target class ω_T is computed. To decide whether an object belongs to the target class or not, a threshold γ on f_{prox} should be determined. A standard way is to supply a fraction r_{fn} of (training) target objects to be rejected by the OCC (a false negative ratio) [14,13]. This means that γ is set up such that $\int \mathcal{I}(f_{\text{prox}}(x,\omega_T) > \gamma)\, d\mu(x) = r_{\text{fn}}$, where $\mathcal{I}$ is the indicator function and μ is some measure. r_{fn} is a small value to prevent a high acceptance of outliers as targets. In other cases, γ can be determined as the $(1-r_{\text{thr}})$-percentile of the sorted sequence of the proximity outputs computed for the training (target) examples. r_{thr} is then a user-specified fraction. Unless stated otherwise, $R \subseteq T$ consists of the target examples only.

I. Neighborhood-Based OCC. The nearest-neighbor data description (NNDD) is realized by the classifier $\mathcal{C}_{\text{NNDD}}$ indirectly built in a pretopological space. The proximity function relies on the nearest neighbor dissimilarities. For n target training objects t_i, a vector of averaged nearest neighbor dissimilarities $d_{nn}(t_i,R) = \frac{1}{k}\sum_{j=1}^{k} d(t_i, p_{t_i}^j)$, where $p_{t_i}^j$ is the j-th nearest neighbor of t_i in R, is obtained. Then, a threshold γ is determined based on the $(1-r_{\text{thr}})$-th percentile of the sorted sequence of d_{nn}. The classifier becomes then:

$$\mathcal{C}_{\text{NNDD}}(D(x,R)) = \mathcal{I}(d_{nn}(x,R) \le \gamma) = \mathcal{I}(\frac{1}{k}\sum_{j=1}^{k} d(x,p_x^j)) \le \gamma), \quad p_x^j \in R. \qquad (1)$$

II. Generalized Mean-Class OCC (GMDD). Assume a symmetric representation $D(R,R)$, where R consists of the targets only. Any such matrix D can be embedded in a pseudo-Euclidean space[1] given dissimilarities are preserved perfectly [5,8,7]. ($\mathcal{E}$ becomes Euclidean iff D is Euclidean.) If $D(T,R)$, $R \subset T$, is given, then $\mathcal{E}$ is determined by $D(R,R)$ and the remaining $T \backslash R$ objects are then projected to $\mathcal{E}$. In the embedded space $\mathcal{E}$, a simple OCC can be designed relying on the distance to the mean vector of the target class. This can be, however, carried out without performing the exact embedding. It can be proved

[1] A pseudo-Euclidean space $\mathcal{E} := \mathcal{R}^{(p,q)}$ is a non-degenerate indefinite inner product space such that the inner product $\langle\cdot,\cdot\rangle_{\mathcal{E}}$ is positive definite on $\mathcal{R}^p$ and negative definite on $\mathcal{R}^q$. $\langle \boldsymbol{x},\boldsymbol{y}\rangle_{\mathcal{E}} = \sum_{i=1}^{q} x_i y_i - \sum_{i=p+1}^{p+q} x_i y_i$ and $d_{\mathcal{E}}^2(\boldsymbol{x},\boldsymbol{y}) = ||\boldsymbol{x}-\boldsymbol{y}||_{\mathcal{E}}^2 = \langle \boldsymbol{x}-\boldsymbol{y}, \boldsymbol{x}-\boldsymbol{y}\rangle_{\mathcal{E}} = d_{\mathcal{R}^p}^2(\boldsymbol{x},\boldsymbol{y}) - d_{\mathcal{R}^q}^2(\boldsymbol{x},\boldsymbol{y})$. Since $\mathcal{E}$ is a linear space, many properties based on inner products can be appropriately extended from the Euclidean case.

that the proximity function $f_{\text{prox}}(x,\omega_T) = ||\boldsymbol{x}_{\mathcal{E}} - \overline{\boldsymbol{x}}_{\mathcal{E}}||^2_{\mathcal{E}}$ in $\mathcal{E}$ (where $\boldsymbol{x}_{\mathcal{E}}$ is the projection of $D(x,R)$ to $\mathcal{E}$) is equivalently computed by the use of square dissimilarities as $f_{\text{prox}}(x,\omega_T) = \frac{1}{n}\sum_{i=1}^n d^2(x,p_i) - \frac{1}{2n^2}\sum_{i=1}^n \sum_{j=1}^n d^2(p_i,p_j)$; see [8, 9, 7] for details. Then, a threshold γ is determined as the (1–r_{thr})-th percentile of the sorted sequence of $f_{\text{prox}}(t_i,\omega_T)$. The generalized mean-class data description (GMDD) becomes then:

$$\mathcal{C}_{\text{GMDD}}(D(x,R)) = \mathcal{I}(\frac{1}{n}\sum_{i=1}^n d^2(x,p_i) - \frac{1}{2n^2}\sum_{i=1}^n \sum_{j=1}^n d^2(p_i,p_j) \le \gamma). \qquad (2)$$

III. Linear Programming Dissimilarity-Data Description (LPDD). This OCC was proposed by us in [9]. It is designed as a hyperplane $H\colon \boldsymbol{w}^T D(x,R) = \rho$ in a dissimilarity space that bounds the target data from above (we assume that d is bounded) and which is attracted towards the origin. Non-negative dissimilarities impose both $\rho \ge 0$ and $w_i \ge 0$. This is achieved by minimizing $\rho/||\boldsymbol{w}||_1$, which is the max-norm distance of the hyperplane H to the origin in the dissimilarity space. Hence, H can be determined by minimizing $\rho - ||\boldsymbol{w}||_1$. Assuming that $||\boldsymbol{w}||_1 = 1$ (to avoid any arbitrary scaling of $\boldsymbol{w}$), H is found by the minimization of ρ only. A target class is then characterized by a linear proximity function on dissimilarities with the weights $\boldsymbol{w}$ and the threshold ρ. The LPDD is then defined as:

$$\mathcal{C}_{\text{LPDD}}(D(x,R)) = \mathcal{I}(\sum_{w_j \ne 0} w_j D(x,p_j) \le \rho), \qquad (3)$$

where w_j are found as the solution to a soft-margin linear programming formulation (the hard-margin case is then straightforward) with $\nu \in (0,1]$ being the upper bound on the target rejection fraction in training (here $\nu := r_{fn}$ is used) [9]:

$$\begin{aligned} &\min \rho + \tfrac{1}{\nu N}\textstyle\sum_{i=1}^N \xi_i \\ &\text{s.t. } \boldsymbol{w}^T D(p_i,R) \le \rho + \xi_i, \quad \textstyle\sum_j w_j = 1,\ w_j \ge 0,\ \rho \ge 0,\ \xi_i \ge 0, \quad i = 1,2,..,N. \end{aligned}$$

As a result, sparse solutions are obtained, i.e. only some w_j are non-zero. Objects of R corresponding to such non-zero weights are called *support objects* (SO). The LPDD can be extended to handle example outliers as well. A label variable $y_i \in \{+1, –1\}$ is used to encode the targets (1) and outliers (–1). The formulation above remains the same, but the main constraint changes to $y_i\,(\boldsymbol{w}^T D(p_i,R)) \le y_i\rho + \xi_i$.

2.1 How Good Is an OCC?

To study the behavior of an OCC, the ROC curve [2, 14] is often used. It is a function of the true positive (target acceptance) versus the false positive (outlier acceptance) ratio. Example outliers are necessary for its evaluation. In principle, an OCC is trained with a fixed target rejection ratio r_{fn} for which the threshold is determined. This OCC is then optimized for one point on the ROC curve.

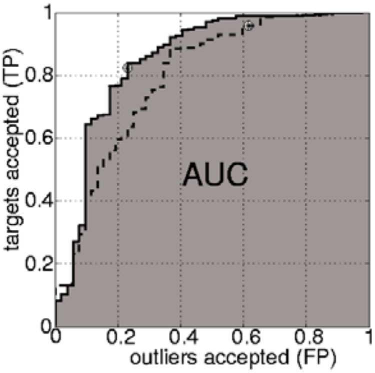

Fig. 1. A ROC curve for the LPDD.

To compare the performance of various classifiers, the AUC measure is used [1]. It computes the Area Under the Curve, which is the total OCC's performance integrated over all the thresholds. The larger AUC, the better the OCC; e.g. in Fig. 1, the solid curve (the LPDD trained using outliers) indicates a better performance than the dashed curve (the LPDD trained on the targets only). The stars indicate points for which the thresholds were optimized.

2.2 Combined Representations

Learning from distinct DRs can be realized by combining them into a new representation and then training a single OCC. As a result, a more powerful representation may be obtained, allowing for a better discrimination. Suppose that K representations $D^{(\tau)}(T,R)$, $\tau = 1, 2, \ldots, K$, all based on the same R, are given. Assume that the dissimilarity measures are similarly bounded (if not they can be scaled appropriately), since only then we can somehow relate their values to each other (otherwise we would need to compare not the direct values but the corresponding percentiles). The DRs can be combined, for instance, in the following ways:

D_{comb}	Expression
Avr	$D_{\mathrm{avr}}(t_i,p_j) \;= \frac{1}{K}\sum_{\tau=1}^{K} D^{(\tau)}(t_i,p_j)$
Prod	$D_{\mathrm{prod}}(t_i,p_j) = \sum_{\tau=1}^{K} \log\left(1 + D^{(\tau)}(t_i,p_j)\right)$
Min	$D_{\min}(t_i,p_j) \;= \min_{\tau}\{D^{(\tau)}(t_i,p_j)\}$
Max	$D_{\max}(t_i,p_j) \;= \max_{\tau}\{D^{(\tau)}(t_i,p_j)\}$

The DRs are combined into one representation by using a sort of fixed rules, usually applied when outputs of two-class classifiers are combined. Note that a DR can be interpreted as a collection of weak classifiers, where each of them is understood as a dissimilarity $D^{(\tau)}(\cdot,p_i)$ to a particular object p_i. In contrary to probabilities, a small dissimilarity value $D^{(\tau)}(t_j,p_i)$ is an evidence of a good 'performance', indicating here that the object t_j is similar to the target p_i. In general, different dissimilarity measures focus on different aspects of the data. Hence, each of them estimates a proximity of an object x to the target p_i as $D^{(\tau)}(x, p_i)$. So, D_{avr} yields an average proximity estimator. When, the dissimilarity measures are independent (e.g. one built on statistical and the other on structural object properties), the product combiner can be of interest. Logically, both D_{avr} and D_{prod} should integrate the strengths of various representations. Here, D_{prod} is expressed such that very small numbers are avoided (they could arise when multiplying close-to-zero dissimilarities). The min operator chooses the minimal dissimilarity value $D^{(\tau)}(x, p_i)$, $\tau = 1, \ldots, K$, hence the maximal evidence for an object x resembling the target t_i. The max operator works the other way around.

2.3 Combined Classifiers

One usually combines classifiers based on their posterior probabilities. The outputs of the OCCs may be converted to estimates of probabilities [14] and standard fixed combiners, such as mean, product and majority voting, can be considered. Here, we also like to proceed with the exact OCCs outputs. For this reason, we focus the LPDDs. Each LPDD is determined by a hyperplane $H^{(\tau)}$ in the dissimilarity space $D^{(\tau)}(T, R)$. The distances to the hyperplane are realized by weighted linear combinations of the form $d_H^{(\tau)}(t_i) = \sum_{w_j^{(\tau)} \neq 0} w_j^{(\tau)} D^{(\tau)}(t_i, p_j) - \rho$. As a result, one may construct an $n \times K$ dissimilarity matrix $D_H = [d_H^{(1)}(T), \ldots, d_H^{(K)}(T)]$ expressing the non-normalized signed distances between the n training objects and K 'base' classifiers. Hence, again an OCC can be trained on D_H. This means that an OCC becomes a *trained combiner* now, re-trained by using the same training set (ideally, an additional validation set should be used). The LPDD can be used again, as well as some other feature-based OCCs. (Although the values of D_H become negative for the targets and positive for the outliers, they are bounded, so the LPDD can be constructed.) Additionally, two other standard data descriptions (OCCs) are used, where a proximity of an object to the target class relies on the k-mean information or density estimation by the Parzen kernels [13], respectively (the appropriate thresholds are set up as described in section 2).

3 Experiments and Results

The data consist of autofluorescence spectra acquired from healthy (target) and diseased (outlier) mucosa in the oral cavity [11, 16]. The measurements were taken at 11 different anatomical locations using six excitation wavelengths 365, 385, 405, 420, 435 and 450 nm. We will denote them by $v_1 - v_6$. After preprocessing [16], each spectrum consists of 199 bins. In total, 856 and 132 spectra representing healthy and diseased tissue, respectively, are given for each wavelength. The spectra are normalized to have a unit area; see also Fig. 2. Two cases are here investigated: combining various DRs for a fixed wavelength of 365 nm

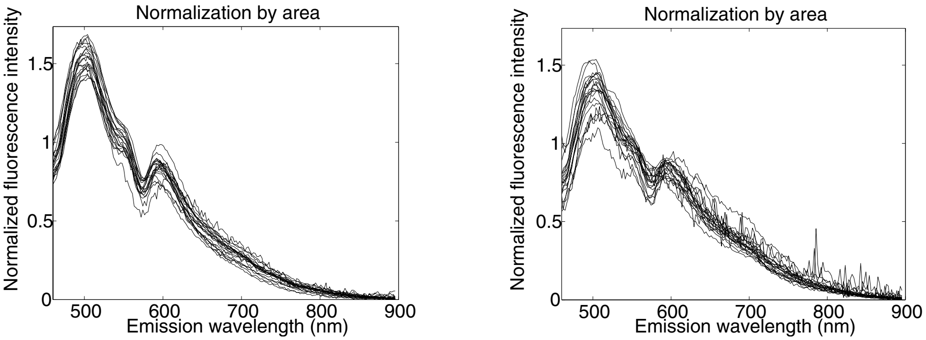

Fig. 2. Examples of normalized autofluoresence spectra for healthy (left) and diseased (right) patients for the excitation wavelength of 365 nm.

(experiment I) and combining representations derived for all the wavelengths (experiment II).

The objects are 30 times randomly split into the training set T and the test set S in the ratio of 60% : 40%, respectively. R, $R \subset T$, consists of the targets only, while T contains additional outliers. $|R| = 514$, $|T| = 594$ and $|S| = 394$ (337/57 healthy/diseased patients). If an OCC cannot use outlier information in the training stage, then it relies on $D^{(\tau)}(R,R)$ only. In the testing stage, $D^{(\tau)}(S,R)$ are used. Since we want to combine the representations directly, they should have a similar range. This is achieved by scaling all the initial $D^{(\tau)}$ by the maximal value of $D^{(\tau)}$ determined on the training data. So, further on, $D^{(\tau)}$ are assumed to be scaled appropriately. The LPDD is trained with $\nu = 0.05$ and the 3-NNDD and the GMDD use the threshold of 0.05. If the LPDD is trained using outlier information, it is denoted as $\mathcal{C}_{\rm LPDD}^{\rm out}$, otherwise, as $\mathcal{C}_{\rm LPDD}$. Trained combiners use the zero threshold. All the experiments are done using DD-Tools [13] and PRTools [3].

Five dissimilarity representations $D^{(1)} - D^{(5)}$ are considered for the normalized spectra in experiment I (wavelength 365 nm). The first three DRs are based on the l_1 (city block) distances computed between the smoothed spectra themselves ($D^{(1)}$) and their first and the second order Gaussian-smoothed ($\sigma = 3$ samples) derivatives ($D^{(2)}$ and $D^{(3)}$, respectively). The zero-crossings of the derivatives indicate the peaks and valleys of the spectra, so they are informative. The differences between the spectra focus on the overlap, the differences in first derivatives emphasize the locations of peaks and valleys, while the differences in second derivatives indicate the tempo of changes in spectra. $D^{(4)}$ is based on the spherical geodesic distance $d_{(4)}(\boldsymbol{x},\boldsymbol{y}) = r \arccos(\boldsymbol{x}^T\boldsymbol{y})/1^2$. $D^{(5)}$ is based on the Bhattacharyya distance, a divergence measure between two probability distributions. This measure is applicable, since the normalized spectra, say, s_i, can be considered as unidimensional histogram-like distributions. They are constant on disjoint intervals $I_1, \ldots, I_N$, such that $s_i(x) = \sum_{z=1}^{N} h_z^i \, \mathcal{I}(x \in I_z)$, where $h_z^i \geq 0$. The Bhattacharyya distance [4] is then: $d_{(5)}(s_i, s_j) = -\log\,(\sum_{z=1}^{N} (h_z^i h_z^j)^{1/2}\,|I_z|$.

Table 1. Experiment I: the AUC performances (in %), averaged over 30 runs, of OCCs built either on the combined DRs or fixed and trained combiners applied to the OCCs outputs. All DRs are considered for the excitation wavelength of 365 nm. SO denotes support objects. The standard deviations of the means are in parenthesis.

Single DRs: OCCs trained on $D^{(\tau)}$						
DR	$\mathcal{C}_{3-\text{NNDD}}$ (1)	$\mathcal{C}_{\text{GMDD}}$ (2)	$\mathcal{C}_{\text{LPDD}}$ (3)	#SO	$\mathcal{C}^{\text{out}}_{\text{LPDD}}$ (3)	#SO
$D^{(1)}$	80.9 (0.5)	77.0 (0.6)	72.3 (0.7)	2.5	79.6 (0.5)	5.5
$D^{(2)}$	86.0 (0.4)	78.4 (0.5)	72.0 (0.7)	2.8	83.1 (0.5)	5.8
$D^{(3)}$	86.7 (0.4)	78.1 (0.6)	78.1 (0.7)	2.9	84.2 (0.5)	5.3
$D^{(4)}$	81.8 (0.5)	76.6 (0.6)	68.0 (0.9)	2.9	80.2 (0.5)	6.1
$D^{(5)}$	85.5 (0.4)	77.3 (0.5)	75.1 (0.6)	2.1	80.1 (0.5)	2.5
Combined DRs: OCCs trained on D_{comb} ($D^{(1)}-D^{(5)}$)						
D_{comb}	$\mathcal{C}_{3-\text{NNDD}}$ (1)	$\mathcal{C}_{\text{GMDD}}$ (2)	$\mathcal{C}_{\text{LPDD}}$ (3)	#SO	$\mathcal{C}^{\text{out}}_{\text{LPDD}}$ (3)	#SO
Avr	95.5 (0.2)	94.6 (0.3)	93.0 (0.3)	4.1	93.4 (0.3)	5.1
Prod	95.7 (0.2)	94.9 (0.3)	93.6 (0.3)	4.6	93.6 (0.4)	7.6
Min	85.6 (0.4)	84.6 (0.4)	84.7 (0.5)	14.6	87.1 (0.9)	15.7
Max	93.5 (0.3)	90.6 (0.4)	84.7 (0.8)	7.1	89.0 (0.6)	10.5
Fixed combiners built on the OCCs outputs from $D^{(1)}-D^{(5)}$						
Combiner	$\mathcal{C}_{3-\text{NNDD}}$ (1)	$\mathcal{C}_{\text{GMDD}}$ (2)	$\mathcal{C}_{\text{LPDD}}$ (3)		$\mathcal{C}^{\text{out}}_{\text{LPDD}}$ (3)	
Mean	98.0 (0.2)	94.4 (0.4)	90.7 (0.6)	—	93.8 (0.3)	—
Prod	98.0 (0.1)	81.3 (0.6)	87.8 (0.5)	—	91.1 (0.3)	—
Voting	98.3 (0.1)	95.9 (0.2)	95.5 (0.2)	—	97.0 (0.2)	—
Trained combiners built on the LPDDs outputs from $D^{(1)}-D^{(5)}$						
Combiner	$\mathcal{C}_{3-\text{NNDD}}$ (1)	$\mathcal{C}_{\text{GMDD}}$ (2)	$\mathcal{C}_{\text{LPDD}}$ (3)	#SO	$\mathcal{C}^{\text{out}}_{\text{LPDD}}$ (3)	#SO
LPDD	—	—	90.1 (0.5)	4.9	95.8 (0.2)	5.0
5-means	—	—	88.0 (0.4)	—	91.1 (0.4)	—
Parzen	—	—	90.5 (0.4)	—	94.5 (0.3)	—

In experiment II, DRs are derived for all excitation wavelengths. The first three measures $D^{(1)}$ – $D^{(3)}$ are used. For each measure, six DRs are combined over the excitation wavelength $v_1 - v_6$ and, in the end, all 18 DRs are combined, as well.

Fixed combiners are also built on the outputs of single OCCs (the outputs need to be converted to posterior probabilities, e.g. as in [14]). Additionally, trained OCC combiners are constructed on the outputs of single LPDDs. The trained combiners are the LPDD and the k-means and Parzen data descriptions [13].

The following observations can be made from experiment I; see Table 1. Both an OCC trained on the combined representations and a trained or fixed combiner on the OCCs outputs improve the AUC performance of each single OCC trained on $D^{(\tau)}$. Concerning the combined representations, the element-wise average and product combiners perform better than the min and max operators. The 3-NNDD seems to give the best results; they are somewhat better than the ones obtained from the GMDD and and the LPDD trained on $D_{\text{comb}}(T, R)$. However, in the testing stage, both the 3-NNDD and the GMDD rely on computing dissimilarities to all 514 objects of R, while the LPDD is based on maximum 16 support objects (see $\#SO$ in Table 1; the SO are determined during training). Hence, if some outliers are available for training, the LPDD can be recommended from the efficiency point of view. The fixed and trained combiners on the OCCs outputs perform well. In fact, the best overall results are reached for the fixed majority voting combiner. However, combiners require more computations; first five OCCs are trained on each $D^{(\tau)}$ separately and then, the final combiner is applied. Yet, if the LPDD $\mathcal{C}^{\text{out}}_{\text{LPDD}}$ is used for training, then the testing stage is cheap: the dissimilarities to 27 objects have to be computed (sum of the support objects for single representations).

Due to lack of space, in Table 2 only some (the best) combining techniques are presented. Again, both an OCC trained on the combined representations (by the average and product) and a fixed or trained combiner on the OCCs outputs significantly improve the AUC performance (by more 10%) of each single OCC. By using all the six wavelengths and three dissimilarity measures (18 in total), all the combining procedures yield nearly perfect performances, i.e. mostly 99.5% or more. The trained combiners on the LPDDs outputs are somewhat worse (possibly due to overtraining) than the majority voting combiner, however, they are similar to the results of the mean combiner. Since the spectra derived from various wavelengths describe different information, an OCC built on their combined representation allows for reaching a somewhat better AUC performance than an OCC built on the DR combined for a single wavelength. From the computational point of view, either an LPDD trained on the combined DR or a fixed voting combiner on the LPDDs outputs should be preferred.

4 Conclusions

Here we study procedures of detecting one-class phenomena based on a set of training examples, performed in an unknown or ill-defined context of alternative phenomena. Since a proximity of an object to a class is essential for such a detection, dissimilarity representations (DRs) can be used as the ones which focus on object-to-target dissimilarities. The discriminative properties of various representations can be enhanced by a proper combining. Three different one-class classifiers (OCCs) are used: the NNDD (based on the nearest neighbor information), the GNMD (a generalized mean classifier in an underlying pseudo-Euclidean space) and the LPDD (a hyperplane in the corresponding dissimilarity space), which offers a sparse solution.

DRs directly encode evidences for objects which lie in close or far neighborhoods of the target objects. Hence, they can naturally be combined (after a proper scaling) into one representation, e.g. by an element-wise averaging. This is beneficial, since only one OCC can be trained, ultimately. From our study on the detection of diseased mucosa in oral cavity, it follows that DRs combined by average or product have a larger discriminative power than any single one. We also conclude that by combining information of DRs derived for spectra of different excitation wavelengths is somewhat more beneficial than by using only

Table 2. Experiment II: the AUC performances (in %), averaged over 30 runs. **Single DRs:** single OCCs built on DRs for six excitation wavelengths (only the worst and the best AUCs; $|\#SO| = 2-7$ for the LPDD). **Combined DRs:** OCCs built on the D_{comb} combined over six wavelengths and fixed $D^{(\tau)}$. **Fixed combiners:** fixed rules applied to the outputs of the trained OCCs and **trained combiners:** combiners trained on the outputs of the LPDDs, both combined over six wavelengths. 'ALL' refers to the results on all 6 × 3 (six wavelengths and three measures) DRs. SO denotes support objects.

Single DRs: OCCs trained on $D^{(\tau)}$ for different v_i								
	$D^{(1)}$		$D^{(2)}$		$D^{(3)}$		ALL	
$\mathcal{C}_{3-\text{NNDD}}$	80.9 - 84.8 (0.5)		82.8 - 87.0 (0.5)		83.5 - 88.8 (0.5)		80.9 - 88.8 (0.5)	
$\mathcal{C}_{\text{GMDD}}$	77.0 - 79.4 (0.7)		77.9 - 81.7 (0.6)		75.4 - 81.6 (0.6)		75.4 - 81.7 (0.7)	
$\mathcal{C}_{\text{LPDD}}$	62.8 - 72.4 (0.8)		65.5 - 72.8 (0.8)		70.7 - 77.5 (0.8)		62.8 - 77.5 (0.8)	
$\mathcal{C}^{\text{out}}_{\text{LPDD}}$	78.3 - 81.7 (0.9)		73.5 - 83.1 (0.7)		77.7 - 83.2 (0.6)		73.5 - 83.2 (0.6)	
Combined DRs: OCCs trained on D_{comb} combined over $v_1 - v_6$								
$\mathcal{C}_{3-\text{NNDD}}$, D_{comb}	$D^{(1)}$		$D^{(2)}$		$D^{(3)}$		ALL	
Avr	97.7 (0.2)	—	97.6 (0.2)	—	96.8 (0.1)	—	99.6 (0.0)	—
Prod	97.7 (0.2)	—	97.7 (0.2)	—	96.9 (0.1)	—	99.7 (0.0)	—
$\mathcal{C}_{\text{GMDD}}$, D_{comb}	$D^{(1)}$		$D^{(2)}$		$D^{(3)}$		ALL	
Avr	97.2 (0.2)	—	97.2 (0.2)	—	96.0 (0.1)	—	99.6 (0.0)	—
Prod	97.3 (0.2)	—	97.4 (0.2)	—	96.3 (0.1)	—	99.6 (0.0)	—
$\mathcal{C}_{\text{LPDD}}$, D_{comb}	$D^{(1)}$	#SO	$D^{(2)}$	#SO	$D^{(3)}$	#SO	ALL	#SO
Avr	96.6 (0.3)	5.2	97.1 (0.3)	4.2	95.6 (0.2)	3.6	99.5 (0.1)	4.3
Prod	96.9 (0.2)	5.7	97.2 (0.3)	4.0	95.8 (0.2)	3.7	99.6 (0.0)	4.9
$\mathcal{C}^{\text{out}}_{\text{LPDD}}$, D_{comb}	$D^{(1)}$	#SO	$D^{(2)}$	#SO	$D^{(3)}$	#SO	ALL	#SO
Avr	96.7 (0.1)	5.1	97.1 (0.1)	4.0	95.6 (0.1)	3.6	99.5 (0.0)	4.5
Prod	96.8 (0.1)	7.3	97.2 (0.2)	5.8	95.8 (0.1)	5.0	99.6 (0.1)	6.6
Fixed combiners applied to the OCCs outputs								
$\mathcal{C}_{3-\text{NNDD}}$ outputs	$D^{(1)}$		$D^{(2)}$		$D^{(3)}$		ALL	
Mean	97.8 (0.1)	—	98.0 (0.1)	—	98.2 (0.2)	—	99.6 (0.1)	—
Prod	98.6 (0.1)	—	98.5 (0.1)	—	98.6 (0.1)	—	99.6 (0.0)	—
Voting	97.6 (0.1)	—	98.7 (0.1)	—	98.6 (0.1)	—	99.8 (0.0)	—
$\mathcal{C}_{\text{GMDD}}$ outputs	$D^{(1)}$		$D^{(2)}$		$D^{(3)}$		ALL	
Mean	94.3 (0.4)	—	94.2 (0.3)	—	94.3 (0.3)	—	98.3 (0.2)	—
Prod	96.0 (0.2)	—	96.4 (0.1)	—	96.7 (0.1)	—	99.7 (0.0)	—
Voting	96.7 (0.2)	—	97.4 (0.1)	—	97.6 (0.1)	—	99.6 (0.1)	—
Fixed and trained combiners applied to the $\mathcal{C}_{\text{LPDD}}$ outputs								
Combiner	$D^{(1)}$	#SO	$D^{(2)}$	#SO	$D^{(3)}$	#SO	ALL	#SO
Mean	92.7 (0.4)	—	92.9 (0.4)	—	91.8 (0.3)	—	94.5 (0.2)	—
Prod	95.7 (0.9)	—	95.7 (1.0)	—	95.7 (0.5)	—	98.7 (0.6)	—
Voting	95.7 (0.4)	—	96.8 (0.2)	—	97.9 (0.1)	—	99.3 (0.1)	—
LPDD	89.3 (0.4)	5.9	91.5 (0.4)	5.9	94.6 (0.2)	5.9	96.6 (0.3)	13.2
Parzen	92.1 (0.3)	—	94.4 (0.3)	—	94.9 (0.3)	—	98.2 (0.1)	—
Fixed and trained combiners applied to the $\mathcal{C}^{\text{out}}_{\text{LPDD}}$ outputs								
Combiner	$D^{(1)}$	#SO	$D^{(2)}$	#SO	$D^{(3)}$	#SO	ALL	#SO
Mean	93.7 (0.4)	—	93.6 (0.5)	—	95.6 (0.4)	—	98.8 (0.3)	—
Prod	95.4 (0.8)	—	96.2 (0.9)	—	97.2 (0.5)	—	99.5 (0.6)	—
Voting	96.3 (0.4)	—	96.8 (0.2)	—	98.0 (0.1)	—	99.5 (0.1)	—
LPDD	95.7 (0.2)	6.0	96.5 (0.2)	6.0	95.8 (0.2)	6.0	99.1 (0.1)	16.3
Parzen	95.5 (0.2)	—	96.8 (0.2)	—	96.2 (0.2)	—	98.9 (0.1)	—

one fixed wavelength, yet different dissimilarity measures. In the former case, all the OCCs on the combined representations performed about the same, while in the latter case, the LPDD trained on the targets seemed to be worse. The fixed OCC combiners have also been applied to the outputs of single OCCs. The overall best results are reached for the majority voting rule. The trained OCC combiners, applied to the outputs of single LPDDs, performed well, yet worse than the voting rule. Concerning the computational issues, either the LPDD on the combined representations should be used or the majority voting combiner applied to the LPDDs outputs.

Further studies on new problems need to be conducted in the future.

Acknowledgments

This work is supported by the Dutch Organization for Scientific Research (NWO) and the Dutch Organization for Applied Research (STW), which supplied the data [16].

References

1. A.P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997.
2. P.S. Bradley, O.L. Mangasarian, and W.N. Street. Feature selection via mathematical programming. *INFORMS Journal on Computing*, 10:209–217, 1998.
3. R.P.W. Duin, P. Juszczak, de D. Ridder, P. Paclík, E. Pękalska, and D. Tax. PR-Tools, a Matlab toolbox for pattern recognition, 2004.
4. F. Esposito, D. Malerba, V. Tamma, H.H. Bock, and F.A. Lisi. *Analysis of Symbolic Data*, chapter Similarity and Dissimilarity. Springer-Verlag, 2000.
5. L. Goldfarb. A new approach to pattern recognition. In *Progress in Pattern Recognition*, volume 2, pages 241–402. Elsevier Science Publishers B.V., 1985.
6. D.W. Jacobs, D. Weinshall, and Y. Gdalyahu. Classification with non-metric distances: Image retrieval and class representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(6):583–600, 2000.
7. E. Pękalska. *working title: Dissimilarity-based pattern recognition.* PhD thesis, Delft University of Technology, The Netherlands, expected in 2004.
8. E. Pękalska, P. Paclík, and R.P.W. Duin. A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, 2(2):175–211, 2001.
9. E. Pękalska, D.M.J. Tax, and R.P.W. Duin. One-class LP classifier for dissimilarity representations. In *NIPS*, pages 761–768. MIT Press, Cambridge, MA, 2003.
10. B. Schölkopf, J.C. Platt, A.J. Smola, and R.C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001.
11. M. Skurichina and R.P.W. Duin. Combining different normalizations in lesion diagnostics. In *Supplementary Proc. ICANN/ICONIP*, pages 227–230, Turkey, 2003.
12. B.M.R. Stadler, P.F. Stadler, G.P. Wagner, and W. Fontana. The topology of the possible: Formal spaces underlying patterns of evolutionary change. *Journal of Theoretical Biology*, 213(2):241–274, 2001.

13. D.M.J. Tax. DD-Tools, a Matlab toolbox for data description, outlier and novelty detection, 2003.
14. D.M.J. Tax and R.P.W. Duin. Combining one-class classifiers. In *Multiple Classifier Systems, LNCS*, volume 2096, pages 299–308. Springer Verlag, 2001.
15. D.M.J. Tax and R.P.W. Duin. Support vector data description. *Machine Learning*, 54(1):45–56, 2002.
16. D.C.G. de Veld, M. Skurichina, M.J.H. Witjes, and et.al. Autofluorescence characteristics of healthy oral mucosa at different anatomical sites. *Lasers in Surgery and Medicine*, submitted, 2003.