Improved one-class SVM classifier for sounds classification
Asma Rabaoui, Manuel Davy, Stéphane Rossignol, Zied Lachiri, Noureddine Ellouze

To cite this version:
IMPROVED ONE-CLASS SVM CLASSIFIER FOR SOUNDS CLASSIFICATION

A. Rabaoui$^1$ M. Davy$^2$ S. Rossignol$^2$ Z. Lachiri$^1$ N. Ellouze$^1$

$^1$Unité de recherche Signal, Image et Reconnaissance des formes, ENIT, BP 37, Campus Universitaire, 1002 le Belvédère, Tunis Tunisia

$^2$LAGIS, UMR CNRS 8146, and INRIA Sequel Team, BP 48, Cité Scientifique, 59651 Villeneuve d’Ascq Cedex, Lille France

Abstract

This paper proposes to apply optimized One-Class Support Vector Machines (1-SVMs) as a discriminative framework in order to address a specific audio classification problem. First, since SVM-based classifier with gaussian RBF kernel is sensitive to the kernel width, the width will be scaled in a distribution-dependent way permitting to avoid under-fitting and over-fitting problems. Moreover, an advanced dissimilarity measure will be introduced. We illustrate the performance of these methods on an audio database containing environmental sounds that may be of great importance for surveillance and security applications. The experiments conducted on a multi-class problem show that by choosing adequately the SVM parameters, we can efficiently address a sounds classification problem characterized by complex real-world datasets.

1 Introduction

Support Vector Machine (SVM) classifiers have been shown to provide better performance than more traditional techniques in many problems, thanks to their ability to generalize. In many cases, SVMs outperform most state-of-the-art classifiers. In recent years, many pattern recognition problems have been tackled using SVMs, ranging from computer vision to text classification. However, their application to computer auditory scene analysis (CASA) and, more specifically to sound recognition problems, has been very limited.

In this paper we focus on the specific task of impulsive sounds classification for a surveillance application using several one class SVMs (1-SVMs) [22]. Though they have been less studied than two-class SVMs (2-SVMs), 1-SVMs have proved extremely powerful in audio applications [22, 10, 9]. Moreover, we found that the 1-SVM methodology is more easy to implement and tune, while having a reasonable training cost. The application addressed here concerns sound classification. In real environment, there might be many sounds which do not belong to one of the pre-defined classes, thus it is necessary to define a rejection class, which may gather all sounds which do not belong to the training classes. An easy and elegant way to do so consists of estimating the regions of high probability of the known classes in the space of features, and considering the rest of the space as the rejection class. Training several 1-SVMs does this automatically.

As many others, the 1-SVM approach is dependent on a set of user-tuned parameters which have a regularization effect during training. Also, since these parameters are not learnt, they may be chosen inadequately and lead to poor performance on a given data set. Thus, a systematic methodology that allows for parameter optimization is quite useful. Our investigations will show that one effective method for improving 1-SVMs consists of well selecting the 1-SVM kernel function, and thus improving sound class prediction.

In the literature, several attempts have been made to improve class-prediction accuracy of SVMs [3, 19, 16, 29]. In [1, 2], a conformal transformation method was proposed to change the kernel function. Another method is proposed in [15] which is based on the kernel-alignment idea but with a simple transformation of the ideal target kernel, to adapt the kernel in the imbalanced training-data problem. Compared to [15], our method deals with the entire training dataset and not just the class-boundary data. Moreover, this solution modify the kernel function instead of the kernel matrix as in [15].

In this paper, we assume that the kernel function is a Radial Basis (RBF) function and we follow [8] to deal with the optimal $\sigma$ which is the tradeoff between the over-fitting loss in dense areas and the under-fitting loss in sparse areas, by scaling the kernel width in a distribution-dependent way. In addition, we will introduce an advanced decision function formulation allowing for a sophisticated dissimilarity measure.

The remainder of this paper is organized as follows. Section 2 gives an overview of the 1-SVM-based sound classi-
2 One-class SVM classifier

The 1-SVM approach\(^1\) [27] has been successfully applied to various learning problems [9, 20, 6, 28, 13]. It consists of learning the minimum volume contour that encloses most of the data in a dataset. Its original application is outlier detection, to detect data that differ from most of the data within a dataset.

More precisely, let \(X = \{x_1, \ldots, x_m\}\) a dataset in \(\mathbb{R}^d\). Here, each \(x_j\) is the full feature vector of a signal, i.e., each signal is represented by one vector \(x_j\) in \(\mathbb{R}^d\). The aim of 1-SVMs is to use the training data so as to learn a function \(f_X : \mathbb{R}^d \mapsto \mathbb{R}\) such that most of the data in \(X\) belong to the set \(\mathcal{R}_X = \{x \in \mathbb{R}^d| f_X(x) \geq 0\}\) while the volume of \(\mathcal{R}_X\) is minimal. This problem is termed minimum volume set (MVS) estimation, see [11], and we see that membership of \(x\) to \(\mathcal{R}_X\) indicates whether this datum is overall similar to \(X\), or not. Thus, by learning regions \(\mathcal{R}_X\), for each class of sound (\(i = 1, \ldots, N\)), we learn \(N\) membership functions \(f_{X_i}\). Given the \(f_{X_i}\)'s, the assignment of a datum \(x\) to a class is performed as detailed in Section 2.1.

1-SVMs solve MVS estimation in the following way. First, a so-called kernel function \(k(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}\) is selected, and it is assumed positive definite, see [27]. Here, we assume a Gaussian RBF kernel such that \(k(x, x') = \exp\left[-\|x - x'\|^2/2\sigma^2\right]\), where \(\|\cdot\|\) denotes the Euclidean norm in \(\mathbb{R}^d\). This kernel induces a so-called feature space denoted \(\mathcal{H}\) via the mapping \(\phi : \mathbb{R}^d \mapsto \mathcal{H}\) defined by \(\phi(x) \triangleq k(x, \cdot)\), where \(\mathcal{H}\) is shown to be reproducing kernel Hilbert space (RKHS) of functions, with dot product denoted \(\langle\cdot, \cdot\rangle_{\mathcal{H}}\). The reproducing kernel property implies that \(\langle \phi(x), \phi(x') \rangle_{\mathcal{H}} = k(x, x')\) which makes the evaluation of \(k(x, \cdot)\) a linear operation in \(\mathcal{H}\), whereas it is a nonlinear operation in \(\mathbb{R}^d\). In the case of the Gaussian RBF kernel, we see that \(\|\phi(x)\|^2_{\mathcal{H}} = \langle \phi(x), \phi(x) \rangle_{\mathcal{H}} = k(x, x) = 1\), thus all the mapped data are located on the hypersphere with radius one, centered onto the origin of \(\mathcal{H}\) denoted \(S(0, R=1)\), see Fig. 1. The 1-SVM approach proceeds in \(\mathcal{H}\) by determining the hyperplane \(W\) that separates most of the data from the hypersphere origin, while being as far as possible from it. Since in \(\mathcal{H}\), the image by \(\phi\) of \(R_X\) is included in the segment of hypersphere bounded by \(W\), this indeed implements MVS estimation [11]. In practice, let \(W = \{h(\cdot) \in \mathcal{H} | \langle h(\cdot), w(\cdot) \rangle_{\mathcal{H}} - \rho = 0\}\) , then its parameters \(w(\cdot)\) and \(\rho\) results from the optimization problem

\[
\min_{w, \xi, \rho} \left\{ \frac{1}{2} \|w(\cdot)\|^2_{\mathcal{H}} + \frac{1}{m} \sum_{j=1}^{m} \xi_j - \rho \right\}
\]

subject to (for \(i = 1, \ldots, m\))

\[
\langle w(\cdot), k(x_j, \cdot) \rangle_r \geq \rho - \xi_j, \quad \xi_j \geq 0
\]

where \(\nu\) tunes the fraction of data that are allowed to be on the wrong side of \(W\) (these are the outliers and they do not belong to \(R_X\)) and \(\xi_j\)’s are so-called slack variables. It can be shown [27] that a solution of (1)-(2) is such that

\[
w(\cdot) = \sum_{j=1}^{m} \alpha_j k(x_j, \cdot)
\]

where the \(\alpha_j\)’s verify the dual optimization problem

\[
\min_{\alpha} \frac{1}{2} \sum_{j, j'=1}^{m} \alpha_j \alpha_j' k(x_j, x_{j'})
\]

subject to \(0 \leq \alpha_j \leq \frac{1}{m}, \sum_{j} \alpha_j = 1\). Finally, the decision function is \(f_X(x) = \sum_{j=1}^{m} \alpha_j k(x_j, x) - \rho\) and \(\rho\) is computed by using that \(f_X(x)=0\) for those \(x_j\)’s which are located on the boundary, i.e., those that verify both \(\alpha_j \neq 0\) and \(\alpha_j \neq 1/m\). An important remark is that the solution is sparse, i.e., most of the \(\alpha_j\)’s are zero (they correspond to the \(x_j\)’s which are inside the region \(R_X\), and they verify \(f_X(x) > 0\)).

As plotted in Fig. 1, the MVS in \(\mathcal{H}\) may also be estimated by finding the minimum volume hypersphere that encloses most of the data (Support Vector Data Description (SVDD) [28]), but this approach is equivalent to the hyperplane one in the case of a RBF kernel. In order to adjust the kernel for optimal results, the parameter \(\sigma\) can be tuned to control the amount of smoothing, i.e. large values of \(\sigma\) lead to flat decision boundaries, as shown in Fig. 2. Also, \(\nu\) is an upper bound on the fraction of outliers in the dataset [27]. Fig. 2 displays an example of a 2-dimensional data to which a 1-SVM using an RBF kernel was applied for different values of \(\nu\) and \(\sigma\).

2.1 A dissimilarity measure

The 1-SVM described above can be used to learn the MVS of a dataset of feature vectors which relate to sounds. This is not yet a multiclass classifier, and we now describe how to build such a classifier, by adapting the results of [9, 12]. Assume that \(N\) 1-SVMs have been learnt from the datasets \(\{X_1, \ldots, X_N\}\), and consider one of them, with associated set of coefficients denoted \((\alpha_j)_{j=1, \ldots, m, \rho}\). In

\(^1\)Various terms have been used in the literature to refer to one-class learning approaches. The term single-class classification originates from Moya [21], but also outlier detection [25], novelty detection [10, 4] or concept learning [14] are used.

\(^2\)In the sequel, we will use \(H \equiv \mathbb{R}_+\) to denote the set of positive real numbers.
A Single-class SVM using an RBF kernel is applied to 2-dimensional data; the parameter $\nu$ characterizes the fractions of SVs (Support Vectors) and outliers (points which are on the wrong side of the hyperplane). The parameter $\sigma$ characterizes the kernel width. For these plots, the parameters are (a) $\nu = 0.1$, $\sigma = 1$ (b) $\nu = 0.3$, $\sigma = 1$ (c) $\nu = 0.3$, $\sigma = 4$.

In the feature space $\mathcal{H}$, the training data are mapped on a hypersphere $S(o,R=1)$. The 1-SVM algorithm defines a hyperplane with equation $W = \{ h \in \mathcal{H} \text{ s.t. } \langle w, h \rangle_{\mathcal{H}} - \rho = 0 \}$, orthogonal to $w$. Black dots represent the set of mapped data, that is, $k(x_j, \cdot)$, $i = 1, \ldots, m$. For RBF kernels, which depend only on $x - x'$, $k(x, x')$ is constant, and the mapped data points thus lie on a hypersphere. In this case, finding the smallest sphere enclosing the data is equivalent to maximizing the margin of separation from the origin.

To determine whether a new datum $x$ is similar to the set $X$, we define a dissimilarity measure, denoted $d(X, x)$ as follows

$$d(X, x) = -\log \left[ \sum_{j=1}^{m} \alpha_j k(x, x_j) \right] + \log[\rho]$$

in which $\rho$ is seen as a scaling parameter which balances the $\alpha_j$'s. Thanks to this normalization, the comparison of such dissimilarity measures $d(X_i, x)$ and $d(X_i', x)$ is possible. Indeed,

$$d(X, x) = -\log \left[ \frac{\langle w(\cdot), k(x, \cdot) \rangle_{\mathcal{H}}}{\rho} \right]$$

because $\| k(x, \cdot) \|_{\mathcal{H}} = 1$, where $w(\cdot) \angle k(x, \cdot)$ denotes the angle between $w(\cdot)$ and $k(x, \cdot)$. By doing elementary geometry in $\mathcal{H}$, we can show that $\frac{\rho}{\| w(\cdot) \|_{\mathcal{H}}} = \cos(\hat{\theta})$, see Fig.1. This yields the following interpretation of $d(X, x)$

$$d(X, x) = -\log \left[ \frac{\cos(\hat{\theta})}{\cos(\hat{\theta})} \right]$$

which shows that the normalization is sound, and makes $d(X, x)$ a valid tool to examine the membership of $x$ to a given class represented by a training set $X$.

2.2 Multiple sound classes 1-SVM-based classification Algorithm

The sound classification algorithm comprises three main steps. Step one is that of training data preparation, and it includes the selection of a set of features (MFCCs in experiments) which are computed for all the training data on sliding windows. Thus, each sound is converted into a time series of such features, which are then averaged over the signal. The resulting feature vectors are centered and normalized to unit variances. The value of $\nu$ is selected in the reduced interval $[0.05, 0.8]$ in order to avoid edge effects for small or large values of $\nu$.

We emphasize the use of feature is confusing and that the features vectors are different from data in feature space, the former being finite dimensional vectors and the latter being functions. However, the word being used in both audio signal processing and VMs, we stick to these two uses.
3. Scaling the kernel width

We assume here that the Gaussian RBF kernel is used. This provides a powerful learner, however it is sensitive to the choice of the scale (or width) parameter \( \sigma \). This parameter controls the trade-off between faithfulness to the training data and smoothness of the decision surface in the space of the data. When \( \sigma \) is too small the SVM overfits the training data. If \( \sigma \) is too large then the decision boundary is not able to model the required decision boundary. In most cases, when using a Gaussian kernel the scale parameter \( \sigma \) must be specified in advance.

In a given classification task, picking the best kernel parameters is actually a non trivial model selection problem. It may be solved by either an exhaustive search over all parameter values or an optimization procedure that explores only a finite subset of the possible values. Generally, the parameter values or an optimization procedure that explores a set \( \sigma \) as half the average euclidean distance between any two points \( x_{i,j} \) and \( x_{i,j'} \), see [26]), and select \( \nu \in [0.05, 0.8] \).

Step 2: Training step

- For \( i = 1, \ldots, N \), solve the 1-SVM problem for the set \( X_i \), resulting in a set of coefficients \((\alpha_{i,j}, \rho_j)\), \( j = 1, \ldots, m_i \).

Step 3: Testing step

- For each sound \( s \) to be classified into one of the \( N \) classes, do:
  - compute its feature vector, denoted \( x \).
  - for \( i = 1, \ldots, N \), compute \( d(X_i, x) \) by using Eq. (5).
  - assign the sound \( s \) to the class \( \hat{i} \) such that \( \hat{i} = \arg\min_{i=1,\ldots,N} d(X_i, x) \)

Algorithm 1: Sound classification algorithm

**Step 1:** Data preparation

- Form the training sets \( X_i = \{x_{i,1}, \ldots, x_{i,m_i}\}, i = 1, \ldots, N \) by computing the feature vectors for all the training sounds selected.
- Set the parameter \( \sigma \) of the Gaussian RBF kernel to some pre-determined value (e.g., set \( \sigma \) as half the average euclidean distance between any two points \( x_{i,j} \) and \( x_{i,j'} \), see [26]), and select \( \nu \in [0.05, 0.8] \).

**Step 2:** Training step

- For \( i = 1, \ldots, N \), solve the 1-SVM problem for the set \( X_i \), resulting in a set of coefficients \((\alpha_{i,j}, \rho_j)\), \( j = 1, \ldots, m_i \).

**Step 3:** Testing step

- For each sound \( s \) to be classified into one of the \( N \) classes, do:
  - compute its feature vector, denoted \( x \).
  - for \( i = 1, \ldots, N \), compute \( d(X_i, x) \) by using Eq. (5).
  - assign the sound \( s \) to the class \( \hat{i} \) such that \( \hat{i} = \arg\min_{i=1,\ldots,N} d(X_i, x) \)

Algorithm 2: Scaling Kernel width

For each training set \( X_i, i = 1, \ldots, N \), do:

- For each training pattern \( x_{i,j} \in X_i, j = 1, \ldots, m_i \), the following steps are needed:
  - Find its \( k \)-NNs in \( X_i \), and compute \( \beta(x_{i,j}) \), according to Eq. (8).
  - Compute the mean of \( \beta(x_{i,j}) \) as
    \[ \bar{\beta} = (1/m_i) \sum_j \beta(x_{i,j}) \]  
  - Compute \( \tilde{\lambda}(x_{i,j}) \) by
    \[ \tilde{\lambda}(x_{i,j}) = 1 + \delta(\beta(x_{i,j}) - \bar{\beta}) \]  

where \( \delta \in [0, 1] \) is a factor for weighting intensity.
• For each testing pattern \( z_i \in \mathcal{X}' \), \( i = 1, \ldots, m' \)
  - Find its \( k \)-NNs in \( \mathcal{X} \), and compute \( \beta(z_i) \), according to Eq. (8)
  - Compute \( \tilde{\beta} \) using Eq. (10).
  - Compute \( \lambda(z_i) \), using Eq. (11).

For each pair \((x, y)\), the modified Gaussian RBF kernel is formulated according to Eq. (9).

An important remark is that the dissimilarity in Eq. (7) is not properly defined anymore. However, if we understand the role \( \lambda(x, y) \) as locally stretching the hypersphere surface in \( \mathcal{H} \), the relative positions of the data in \( \mathcal{H} \) is not deeply modified, and the interpretation of \( d(\mathcal{X}, x) \) remains valid.

4. Experimental evaluation

For experimental evaluation, considering several sound libraries was necessary for building a representative, large, and sufficiently diversified database. The major part of the sound samples used in the recognition experiments was taken from different sound libraries [18, 24]. Some particular classes of sounds have been built or completed with sounds recorded by us. The selected sounds are impulsive and they are typical of surveillance applications. More precisely, the database consists of above 1,000 sounds belonging to 9 classes (Human screams, Gunshots, Glass breaks, Explosions, Door slams, Dog barks, Phone rings, etc) as described in the following Table 1.

Table 1: Classes of sounds and number of samples in the database used for performance evaluation.

<table>
<thead>
<tr>
<th>Classes</th>
<th>Train</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human screams</td>
<td>48</td>
<td>25</td>
<td>73</td>
</tr>
<tr>
<td>Gunshots</td>
<td>150</td>
<td>75</td>
<td>225</td>
</tr>
<tr>
<td>Glass breaks</td>
<td>58</td>
<td>30</td>
<td>88</td>
</tr>
<tr>
<td>Explosions</td>
<td>41</td>
<td>21</td>
<td>62</td>
</tr>
<tr>
<td>Door slams</td>
<td>209</td>
<td>105</td>
<td>314</td>
</tr>
<tr>
<td>Dog barks</td>
<td>36</td>
<td>19</td>
<td>55</td>
</tr>
<tr>
<td>Phone rings</td>
<td>34</td>
<td>17</td>
<td>51</td>
</tr>
<tr>
<td>Children voices</td>
<td>58</td>
<td>29</td>
<td>87</td>
</tr>
<tr>
<td>Machines</td>
<td>40</td>
<td>20</td>
<td>60</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>674</td>
<td>341</td>
<td>1015</td>
</tr>
</tbody>
</table>

All signals in the database have a 16 bits resolution and are sampled at 44100 Hz. A feature vector (of dimension 12) issued from Mel Frequency Cepstral Coefficients (MFCCs) is computed for each entire sample file. The analysis window length for all features was 25 ms and the used windowing function was Hamming. The overlap between successive frames was 50% of the frame length.

The objective of this experimental study is to evaluate the quality of the approach described above. The sound recognition ability of our method is evaluated by supervised classification using a 1-SVM-based classifier as described in Algorithm 1 in which the \( v \) value is fixed to 0.2. Results for Algorithm 2 are displayed in Table 2. The weighting intensity factor \( \delta \) is set to 0.5 in all experiments.

The results listed in Table 2 show that the performance with scaled width kernel is better than that with fixed width kernel and thus confirms that the pattern spatial distribution affects the sound recognition quality.

5. Conclusions

In this paper, we propose a system motivated by a practical surveillance application that uses a discriminative method based on one-class SVMs, together with a sophisticated dissimilarity measure. In order to enhance the discrimination ability of the proposed system, the kernel parameters are also studied. The under-fitting and over-fitting phenomenon is tackled by taking into account the patterns’ spatial distribution.

References

Table 2: Recognition accuracy using fixed and scaled-width kernel applied to 1-SVMs based classifier.

<table>
<thead>
<tr>
<th>σ</th>
<th>acc. with fixed width</th>
<th>acc. with scaled-width</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>92.33</td>
<td>91.56</td>
</tr>
<tr>
<td>10</td>
<td>92.80</td>
<td>93.66</td>
</tr>
<tr>
<td>8</td>
<td>92.80</td>
<td>93.73</td>
</tr>
<tr>
<td>6</td>
<td>91.64</td>
<td>92.99</td>
</tr>
<tr>
<td>4</td>
<td>91.83</td>
<td>91.89</td>
</tr>
<tr>
<td>2</td>
<td>90.79</td>
<td>91.22</td>
</tr>
<tr>
<td>1</td>
<td>88.23</td>
<td>88.23</td>
</tr>
<tr>
<td>0.5</td>
<td>87.66</td>
<td>87.66</td>
</tr>
<tr>
<td>0.1</td>
<td>88.90</td>
<td>89.33</td>
</tr>
</tbody>
</table>


