2010 International Conference on Pattern Recognition

Hydroacoustic signal classification using kernel functions for variable feature sets

Matthias Tuma, Christian Igel*
Institut für Neuroinformatik
Ruhr-Universität Bochum, Germany
{ *Matthias.Tuma, Christian.Igel* }*@ini.rub.de*

Mark Prior†
CTBTO Preparatory Commission
Vienna International Centre, Austria
Mark.Prior@ctbto.org

***Abstract*—Large-scale geophysical monitoring systems raise the need for real-time feature extraction and signal classification. We study support vector machine (SVM) classification of hydroacoustic signals recorded by the Comprehensive Nuclear-Test-Ban Treaty's verification network. Due to constraints in the early signal processing most samples have incomplete feature sets with values missing not at random. We propose kernel functions explicitly incorporating Boolean representations of the missingness pattern through dedicated sub-kernels. For kernels with more than a few parameters, gradient-based model selection algorithms were employed. In the case of binary classification, an increase in classification accuracy as compared to baseline SVM and linear classifiers was observed. In the multi-class case we evaluated four different formulations of multi-class SVMs. Here, neither SVMs with standard nor with problem-specific kernels outperformed a baseline linear discriminant analysis.**

***Keywords*-support vector machine; missing data; underwater sound; CTBTO; treaty verification**

I. Introduction

The Preparatory Commission for the Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO) is currently installing a global verification system to monitor the earth for nuclear explosions [1], [2]. Across the oceans, hydroacoustic sensors are placed at depths of 1 km and less, where the *deep sound channel* acts as a waveguide for underwater acoustic signals. Observed signals stem from a wide variety of sources. For direct verification of the treaty, the binary classification task of identifying signals with explosive signature (underwater volcanoes, chemical or nuclear explosions) is paramount. Additionally distinguishing signals caused by earthquakes may benefit later processing stages, leading to a three-class problem of noise-like, earthquake-caused, and explosion-like underwater signals. Due to constraints in the early processing any two samples may have non-identical and even non-intersecting feature sets. We focus on soft margin support vector machine (SVM) classification, a state-of-the-art approach in pattern recognition. Support vector machines implement regularized risk minimization, are well-rooted in statistical learning theory, and allow for flexible re-representation of input data via kernel functions.

*This work was carried out within the CTBTO's International Scientific Studies 2009 project, having called for independent studies on the CTBTO verification regime [1]. M.T. gratefully acknowledges a scholarship from the German National Academic Foundation.

†The views expressed herein are those of the authors and do not necessarily reflect the views of the CTBTO Preparatory Commission.

II. Feature set characteristics

Raw sensor data is from the outset filtered into eight partially overlapping frequency bands between 1 and 100 Hz (1-2, 2-80, 3-6, 6-12, 8-16, 16-32, 32-64, and 64-100 Hz). In each band, detection and feature extraction are carried out independently. A detection in a band is made when the ratio of a short-term average to long-term average exceeds a station-specific threshold, with time windows of 10s and 150s, respectively. A signal is defined by one or more contemporaneous detections across frequency bands. For each signal a fixed set of 16 identically calculated features is extracted from every band with a detection and their union used for representation. This leads to a situation where $n \cdot 16$ real-valued features ($1 \leq n \leq 8$) are associated with a sample and thus any two signals' feature sets may differ and not even overlap. Features are listed in Table I and can be categorized into (i) time-related, (ii) energy-related, (iii) statistical moments, and (iv) cepstral. The latter are present in two variants, calculated from a low-pass filtered and a detrended spectrum. In total 778 labeled samples were available.

III. Dealing with incomplete data

In general, incomplete data sets can be grouped according to properties of the probability distribution assumed responsible for the pattern of missing features. If the missingness distribution is conditionally independent of the data or of the missing values given the observed ones, the data is coined *missing completely at random* (MCAR) or *missing at random* (MAR), respectively. CTBT data are neither MCAR or MAR but rather *missing not at random* (MNAR) as seen in section II. Common practice for samples with missing values is to either discard them or choose from a set of imputation methods, where missing values are filled in by some strategy. As less than 5% of CTBT samples have

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.253

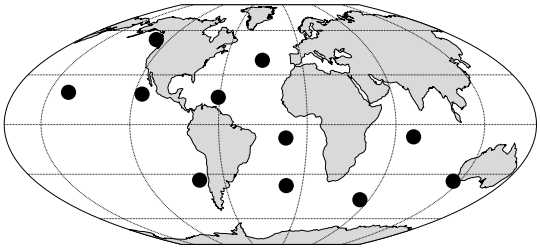

Figure 1. Locations of hydroacoustic stations defined by the treaty.

values for all features, the former is not an option. Simple imputation procedures can bias the estimates, and usually more so with "less random" missingness distributions. More elaborate techniques include Bayesian multiple imputation and maximum likelihood estimations, but these almost always rely on MAR data or only weak deviations thereof (see [3] for a review). Due to the scarcity of data, estimating joint densities in a probabilistic framework (e.g., [4]) does not seem feasible in the present case. In the context of SVM classification one recent approach [5] avoids imputation by altering the SVM margin interpretation to directly deal with incomplete samples. However, it is best suited for features *structurally absent* rather than MNAR data and leads to non-convex optimization problems. Dick et al. [6] optimize the assumed imputation distribution and SVM mutually by marginalizing kernels over the former. Yet, the approach operates on MAR data and cannot handle incomplete test sets. We here take an approach that maintains convexity and is applicable to test sets by directly passing representations of the missingness pattern to appropriate kernel functions which account for it through a specific structure.

IV. Kernel functions for variable feature sets

In kernel-based algorithms, inner products in a feature space are replaced by positive definite kernel functions. Among the most commonly used are radial basis function (RBF) kernels of the form $k(x,z) = e^{-\gamma\|x-z\|^2}$, with $x, z \in \mathbb{R}^n$ and the bandwidth γ as single free parameter. A second common class are polynomial kernels of the form $k(x,z) = (\langle x,z\rangle + c)^d$, with offset c and integer exponent d. In the following we take an approach in which kernel functions make use of both the information contained in the values of present features as well as in their missingness pattern. We introduce eight additional Boolean features b_i, $1 \leq i \leq 8$, indicating whether the features of band i are present or not. We write the joint feature space $\mathcal{F}$ of Boolean and real values as the Cartesian product $\mathcal{F}_b \times \mathcal{F}_r$ of the space of Boolean and real-valued features, respectively, and a sample of the extended feature set as $x = (x_b, x_r)$ with $x_b \in \mathcal{F}_b$ and $x_r \in \mathcal{F}_r$. It is evident that in order to exploit the information held by the Boolean values, they should be processed differently from the real-valued ones. Thus we employ kernel functions with a *bipartite structure*, where

Table I
Features extracted from each band with a detection.

Temporal	Energy	Statistical	Cepstral (2x)
Peak Time	Peak Level	Time Spread	Position
Mean Arrival	Total Energy	Skewness	Level
Total Time	Avg. Noise	Kurtosis	Variance
Crossing Rate			

one sub-kernel k_b operates on the Boolean features $x_b \in \mathcal{F}_b$ and another sub-kernel k_r on their real counterparts $x_r \in \mathcal{F}_r$. As we use standard SVMs that cannot directly deal with missing features, an imputation procedure for $\mathcal{F}_r$ still has to be chosen. We impute a single value of zero, because in the limit case of a zero-threshold detector, it constitutes the physically or statistically plausible continuation for some of the hydroacoustic features, including peak and total energy, skewness, and total time. A function f combining the sub-kernels k_b and k_r into one, $k : (\mathcal{F}_b \times \mathcal{F}_r) \times (\mathcal{F}_b \times \mathcal{F}_r) \to \mathbb{R}$, $k(x,z) = f(k_b(x_b,z_b), k_r(x_r,z_r))$, must be chosen such that the overall kernel k remains positive definite. Two intuitive possibilities, here using a polynomial and an RBF sub-kernel, are a direct product kernel

$$k(x,z) = (\langle x_b, z_b\rangle + c)^d \cdot e^{-\gamma\|x_r - z_r\|^2} \tag{1}$$

and a weighted direct sum kernel

$$k(x,z) = (\langle x_b, z_b\rangle + c)^d + we^{-\gamma\|x_r - z_r\|^2} \tag{2}$$

with weighting factor $w > 0$. Their structure stresses similarity of features across all bands on the one hand and inherent differences between Boolean and real-valued features on the other. Given that underwater signal propagation is highly frequency-dependent, one might demand that kernel parameters be allowed to incorporate information from different bands differently. From such a perspective we can view $\mathcal{F}$ as $\prod_{i=1}^{8}(\mathcal{F}_{b,i} \times \mathcal{F}_{r,i})$ and in analogy to eq. (1) introduce *band-wise bipartite kernels*,

$$k_i(x_i, z_i) = (\langle x_{b,i}, z_{b,i}\rangle + c)^d \cdot e^{-\gamma_i\|x_{r,i} - z_{r,i}\|^2} \quad , \tag{3}$$

where we use identical c and d, but different γ_i across all bands. We combine these eight sub-kernels through a convolution kernel [7] of integer degree M, $1 \leq M \leq 8$,

$$k_M(x,z) = \sum_{1 \leq j_1 < \cdots < j_M \leq 8} \prod_{m=1}^{M} k_{j_m}(x_{j_m}, z_{j_m}) \ . \tag{4}$$

Equation (4) adds up all M-th order monomials, multiples excluded, of sub-kernels k_i and hence constitutes a direct sum kernel for $M = 1$ and direct product kernel for $M = 8$. Since each $x_{b,i}$ in eq. (3) is a single Boolean, the corresponding scalar product can either yield 1 if both samples feature detections in band i or 0 if only one or none of them do. In the special case of $c = 0$, k_i will equal the respective value of $k_{r,i}$ if both samples were detected in

frequency band i or zero otherwise, and eq. (4) will yield all M-th order monomials of those $k_{r,i}$ for which both x and z have detections in all corresponding bands. By summing over M we derive a *cumulative convolution kernel* of integer degree N, $1 \leq N \leq 8$,

$$k_N(x,z) = \sum_{M=1}^{N} k_M(x,z) \ , \tag{5}$$

which includes contributions from all monomials with degree not higher than N.

V. Experiments and results

We conducted experiments with two different foci. On the one hand, we considered finding a suitable and well-performing binary classifier for the main task of identifying signals with explosive signature. With eqs. (1, 2, 4, 5) we have proposed four candidate classes of SVM kernels for this task. On the other hand, we explored four different SVM multi-class formulations for the extended task of discriminating noise-like, earthquake-caused, and explosive-like signals. All results in this section are averages over test errors obtained from five different splits into 80% training and 20% test data. In addition to this "outer" 5-fold cross-validation procedure, for each fold the model selection algorithms used another 5-fold "inner" cross-validation (splitting the respective training set) for determining SVM hyperparameters as described below.

A. Support vector machines

On the binary task of identifying explosive-like signals we used L_1-norm soft margin SVMs. In addition we considered four of the many qualitatively different ways to extend binary SVMs for multi-class tasks. We employed (i) the often-used one-versus-all approach (OVA); (ii) the theoretically more stringent approach independently proposed by Weston and Watkins, and Vapnik (WW, [8]); (iii) the variant by Crammer and Singer (CS, [9]) which relaxes some constraints of WW with the goal of increased learning speed; and (iv) multi-class classification with maximum margin regression at one-class cost as proposed by Szedmak, Shawe-Taylor, and Parado-Hernandez (MMR, [10]). The MMR algorithm considers less complex hypothesis classes without bias parameter and only learns ℓ parameters instead of $\ell \cdot q$ as OVA, WW, and CS, where ℓ and q denote the number of training examples and the number of classes, respectively.

B. Model selection

The kernel function classes of eqs. (1, 2, 4, 5) are parametrized by increasing numbers of hyperparameters, namely offset and exponent of the polynomial sub-kernel as well as the RBF bandwidth for eq. (1); plus a weighting term for eq. (2); or plus seven bandwidth parameters and a degree for eqs. (4, 5). Of these, the exponent d and the degrees M, N must be constrained to integer values which we accounted for by using fixed values within individual runs. When used in an SVM, classifier performance crucially depends on the choice of these kernel parameters as well as on the regularization parameter C of the SVM itself. For the kernels given by eqs. (1, 2) we conducted standard grid search on the five-fold cross-validation error on the training set (CV-5) for parameter selection. For the kernels given by eqs. (4, 5) which have too many parameters for grid search to be applicable, we used two different gradient-based methods both implemented in the *Shark Machine Learning Library* [11]. For binary SVMs we employed a recently proposed and well-performing maximum likelihood based approach [12] mutually optimizing C and the kernel parameters. As this method has not yet been transferred to the multi-class setting we in that case chose kernel parameters by maximizing the *kernel-target alignment* (KTA, [13], [14]) and selected C through subsequent grid search on the CV-5.

The maximum likelihood approach to model selection has recently been proposed in [12]. It relies on an established approximation of the class conditional probabilities first suggested by Platt [15]. At the core of his approximation is a cross-validation based logistic regression estimate of the class conditionals (basically this is done by squashing the SVM decision function with a parametrized sigmoid). Given this estimate, a log-likelihood function is formulated that is differentiable and used as objective function for SVM model selection. This approach corresponds to a Bayesian interpretation of model parameters.

The KTA on the other hand is a classifier-independent approach to choosing kernel parameters. It is below outlined for binary data but is easily extended to the multi-class case. On ℓ consistent training samples we can measure the similarity of two kernel functions k_1 and k_2 by the normalized inner product of their Gram matrices K_1 and K_2 as $S(k_1,k_2) := \frac{\langle K_1,K_2\rangle}{\sqrt{\langle K_1,K_1\rangle\langle K_2,K_2\rangle}}$, where $\langle A,B\rangle := \sum_{n,m=1}^{\ell} A_{nm}B_{nm}$ for $A,B \in \mathbb{R}^{\ell\times\ell}$. The matrix Y, with $[Y]_{ij} = y_i y_j$ the product of the labels of training patterns i and j, can be viewed as Gram matrix of a kernel perfectly fitting the given data. This leads to the definition of the *kernel-target alignment* $\hat{A}(k) := \frac{\langle K,Y\rangle}{\ell\sqrt{\langle K,K\rangle}}$, which is differentiable w.r.t. kernel parameters, but independent of the classifier. After gradient ascent on the KTA we use one-dimensional grid search on the CV-5 to determine the SVM regularization parameter C.

C. Results

For the binary classification task, Table II lists test classification errors averaged over five trials, where LDA denotes linear discriminant analysis; *svm-i* an SVM with RBF kernel; *svm-s* an SVM with direct sum kernel (eq. 2); *svm-p* an SVM with direct product kernel (eq. 1); and *svm-c* an SVM with convolution kernel of degree one (eq. 4, 5). The LDA and the *svm-i* operated on the real-valued, zero-imputed data set and the last three SVM classifiers on the Boolean-extended data

Table II
AVERAGE CLASSIFICATION TEST ERRORS FOR THE BINARY CASE.

	Classifier				
	LDA	svm-i	svm-s	svm-p	svm-c
Error [%]	5.2	4.9	4.9	4.8	4.3

set. All SVMs performed better than the linear approach, and the two bipartite kernels from eqs. (1, 2) were on par with the baseline RBF kernel. Additionally passing the Boolean features to the latter did not change its performance. For a degree of one, the convolution-based kernels from eqs. (4, 5) are mathematically identical and performed best among all approaches employed. With increasing degrees, error rates tended to increase for both. At the highest possible degree of eight, they with 5.9% were higher than that of LDA.

In the multi-class task we compared LDA and four different multi-class SVM formulations (see section V-A). Each SVM was tested with the RBF kernel on the real-valued, zero-imputed data set as well as with the 1-degree convolution kernel on the extended data set. As shown in Table III, the LDA performed best among all the classifiers. This is due to overfitting of the multi-class SVMs as indicated by comparatively strong performance on the training data. The too few samples per class are not sufficient to identify the parameters of the flexible multi-class SVMs, especially in combination with parameter-rich kernels.

VI. DISCUSSION AND CONCLUSIONS

Our experiments showed that support vector machines (SVMs) are well suited to automatically identify hydroacoustic signals with explosive signature as recorded by the Comprehensive Nuclear-Test-Ban Treaty's verification network. We proposed two classes of problem-specific kernel functions: first, kernels with a bipartite structure taking into account an individual sample's missingness pattern; and second, variants of convolution kernels, which combine several bipartite sub-kernels operating on features from a single frequency band. The latter improved classification accuracy compared to the baseline radial basis function (RBF) kernel, while the former performed on par. In the multi-class task, we examined the RBF kernel and the best-performing convolution kernel from the binary task in combination with four different SVM multi-class extensions. Here, none of the above outperformed a standard linear discriminant analysis. In future work, the approach taken in [12] could be extended to multi-class SVMs and be used in place of the kernel-target alignment. Also, variants of the kernel functions proposed here might be combined with the direct approach taken in [5], circumventing the auxiliary imputation step.

Table III
ERROR RATES FOR MULTI-CLASS CASE.

		SVM multi-class formulation			
Classifier	LDA	OVA	WW	CS	MMR
LDA	15.6	–	–	–	–
svm-i	–	16.4	16.4	17.0	17.4
svm-c	–	17.0	16.8	16.8	18.2

REFERENCES

[1] A. Thunborg, Ed., *Science for security. Verifying the Comprehensive Nuclear-Test-Ban Treaty*. Preparatory Commission for the CTBTO, Vienna, Austria, 2009.

[2] D. Hafemeister, "Progress in CTBT monitoring since its 1999 senate defeat," *Science and Global Security*, vol. 15, pp. 151–183, 2007.

[3] J. Schafer and J. Graham, "Missing data: our view of the state of the art," *Psychol Methods*, vol. 7, pp. 144–177, 2002.

[4] Z. Ghahramani and M. I. Jordan, "Supervised learning from incomplete data via an EM approach," in *NIPS 6*. Morgan Kaufmann, 1994, pp. 120–127.

[5] G. Chechik, G. Heitz, G. Elidan, P. Abbeel, and D. Koller, "Max-margin classification of data with absent features," *JMLR*, vol. 9, pp. 1–21, 2008.

[6] U. Dick, P. Haider, and T. Scheffer, "Learning from incomplete data with infinite imputations," in *Proc. of the 25th ICML*, New York, NY, USA, 2008, pp. 232–239.

[7] D. Haussler, "Convolution kernels on discrete structures," UCS-CRL-99-10, University of California at Santa Cruz., Tech. Rep., 1999.

[8] J. Weston and C. Watkins, "Support vector machines for multi-class pattern recognition," in *Proc. of the 7th ESANN*, 1999, pp. 219–224.

[9] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *JMLR*, vol. 2, pp. 265–292, 2002.

[10] S. Szedmak, J. Shawe-Taylor, and E. Parado-Hernandez, "Learning via linear operators: Maximum margin regression," PASCAL, Southampton, UK, Tech. Rep., 2006.

[11] C. Igel, T. Glasmachers, and V. Heidrich-Meisner, "Shark," *JMLR*, vol. 9, pp. 993–996, 2008.

[12] T. Glasmachers and C. Igel, "Maximum Likelihood Model Selection for 1-Norm Soft Margin SVMs with Multiple Parameters," *PAMI, doi: 10.1109/TPAMI.2010.95*, 2010.

[13] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola, "On kernel-target alignment," in *NIPS 14*. MIT Press, 2002.

[14] C. Igel, T. Glasmachers, B. Mersch, N. Pfeifer, and P. Meinicke, "Gradient-Based Optimization of Kernel-Target Alignment for Sequence Kernels Applied to Bacterial Gene Start Detection," *IEEE/ACM Trans Comput Biol Bioinf*, vol. 4, pp. 216–226, 2007.

[15] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. MIT Press, 1999, pp. 61–74.