Pattern Recognition 44 (2011) 2614–2629

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Wavelet kernel learning

F. Yger, A. Rakotomamonjy *

Université de Rouen, LITIS EA 4108, 76800 Saint-Etienne du Rouvray, France

ARTICLE INFO

Article history:
Received 18 August 2010
Received in revised form
3 February 2011
Accepted 5 March 2011
Available online 21 March 2011

Keywords:
Wavelet
Multiple kernel learning
SVM
Quadratic mirror filter

ABSTRACT

This paper addresses the problem of optimal feature extraction from a wavelet representation. Our work aims at building features by selecting wavelet coefficients resulting from signal or image decomposition on an adapted wavelet basis. For this purpose, we jointly learn in a kernelized large-margin context the wavelet shape as well as the appropriate scale and translation of the wavelets, hence the name "wavelet kernel learning". This problem is posed as a multiple kernel learning problem, where the number of kernels can be very large. For solving such a problem, we introduce a novel multiple kernel learning algorithm based on active constraints methods. We furthermore propose some variants of this algorithm that can produce approximate solutions more efficiently. Empirical analysis show that our active constraint MKL algorithm achieves state-of-the art efficiency. When used for wavelet kernel learning, our experimental results show that the approaches we propose are competitive with respect to the state-of-the-art on brain–computer interface and Brodatz texture datasets.

1. Introduction

In any pattern recognition problem, the choice of the features used for characterizing an object to be classified is of primary importance. Indeed, those features largely influence the performance of the pattern recognition system. Usually, features are extracted from original data and it can be a tricky task to craft them, so that they capture discriminative characteristics. Yet, this issue becomes even more complex when dealing with data corrupted by noise.

For instance, in brain–computer interface (BCI) problems or in other biomedical engineering classification problems like electrocardiogram (ECG) beat classification problems, several type of features have been proposed in the literature. In some cases, preprocessed time samples of the signal are directly used as features [28]. In other situations, classical signal transforms [8,15,42,13,16] such as wavelet transform or time-frequency transform are applied to the signal before extracting features from these novel representations. However, the choices of wavelet bases or time-frequency transforms used in those approaches are usually grounded on criterion adapted for signal representation or signal denoising and thus, they may not be optimal for classification.

In the same way, many works which dealt with texture classification used features extracted from wavelet decomposition [21,17]. In these two studies, the authors considered fixed wavelet bases such as *Coiflet* or *Daubechies* wavelets and have justified their choices based on the experimental results they achieved. However, there is no guarantee about the optimality of such wavelets, in the sense that some other wavelets with more appropriate waveform may lead to better classification performances. This difficulty of choosing a correct wavelet basis for texture classification was already noted by Busch et al. [7]. Indeed, their work clearly showed that basic wavelets such as Haar's wavelet may provide better features than complex ones. Moreover, they brought experimental evidences that combining simple bases may produce more efficient features. All these points emphasize the need for adapting the wavelet dictionary, or more generally the discriminant basis dictionary, to the classification problem at hand. This adaptation can be performed for instance by designing a pattern recognition system which jointly optimizes a dictionary and the classifier.

At the present time, this problem of learning discriminant dictionary adapted to a problem at hand has attracted few attentions. For instance, the trends followed by Huang and Aviyente [14] and Mairal et al. [23] are based on ideas from signal representation dictionary learning. Their approaches consist in selecting representative and discriminative features as atoms among an overcomplete codebook (which is not necessarily based on wavelet). In these works, the selection problem is cast as an optimization problem with respect to a criterion which takes into account signal representation error, discrimination power and sparsity.

Prior to these approaches, a stream of research [30,5] investigated the way of choosing a wavelet basis for classification among overcomplete wavelet packet decomposition [24]. This basis selection problem was grounded on several different criteria which only consider discrimination ability instead of representation one. In these works, however, the wavelet shape was kept fixed (classical wavelets were considered) and the best wavelet basis resulting

* Corresponding author.
E-mail address: alain.rakoto@insa-rouen.fr (A. Rakotomamonjy).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.03.006

from a wavelet packet decomposition was selected. Following these approaches of selecting optimal discriminant wavelet basis, some recent works considered discriminative criteria for generating discriminant wavelet waveforms so as to adapt the wavelet to the data to be classified. For instance, Strauss et al. [33,34] and Neumann et al. [25] tune their wavelet by maximizing the distance in the wavelet feature space of the means of the classes to be classified. Instead, Lucas et al. [22] consider the wavelet generating filter as a parameter of their kernel-based classifier (a support vector machine) and propose to select this parameter according to a cross-validation error criterion.

In this work, we address the problem of discriminant dictionary learning by following the road opened by Strauss et al. [33] and Lucas et al. [22]. Indeed, we consider the problem of wavelet adaptation for wavelet-based signal classification, but in addition, we propose to jointly

- learn the shape of the mother wavelet, since classical wavelet such as Haar, or Daubechies ones may not be optimal for a given discrimination problem,
- select the best wavelet coefficients that are useful for the discrimination problem. Indeed, we believe that among all the coefficients derived from a wavelet decomposition, most of them may be irrelevant,
- combine features obtained from different wavelet shapes and coefficient selections,
- and learn a large-margin classifier.

For this purpose, we cast this problem as a multiple kernel learning, where each kernel is related to some wavelet coefficients resulting from a parametrized wavelet decomposition. To this end, we first show how to build kernels from a wavelet decomposition. Then, we describe how the problem of selecting optimal wavelet shape and coefficients can be related to a multiple wavelet kernel learning problem. As a side contribution, we propose an active constraint multiple kernel learning (MKL) algorithm grounded on the KKT conditions of the primal MKL problem that is proved to achieve state-of-the-art in term of computational efficiency compared to recent MKL algorithms [35,9]. We also discuss some variants of our MKL algorithm which are more efficient when the number of kernels become very large or infinite at the expense of providing an approximate solution of the learning problem. We want to emphasize that our approach differs from those of Lucas et al. [11], Strauss et al. [33] and Neumann et al. [25] as we essentially learn a combination wavelet coefficients obtained from different optimal wavelet waveforms while the mentioned works consider a single adapted wavelet. Furthermore, as detailed in the sequel, our approach is able to deal with wavelets built from longer quadrature mirror filters (which have better smoothness properties).

The paper is organized as follows. Section 2 reviews some backgrounds on quadrature mirror filters and parametrized wavelets and shows how kernels can be built from wavelet decomposition. Section 3 details the novel active kernel MKL algorithm that we use for jointly learning wavelet kernel combinations and the classifier. Experimental analysis on toy dataset, BCI problem and texture classification are given in Section 4. Section 5 concludes the paper and provides some final discussions and perspectives on this work. For a sake of reproducibility, the code used for this work is available at : http://asi.insa-rouen.fr/enseignants/∼ arakotom/code/wkl.html.

2. Wavelet kernels

In this section, we briefly review wavelets, quadrature mirror filter banks and wavelet decomposition. We also present a general way to extract features and kernels from such a decomposition.

2.1. Parametrized wavelet decomposition

Depending on the used wavelet basis, a signal or image wavelet decomposition will have different property (e.g. different sparsity pattern). Hence, as discussed in Mallat's [24] book for signal representation and denoising, the choice of the mother wavelet shape has a strong impact on wavelet-based features for discrimination. In this section, we explain why parametrized quadrature mirror filter banks are the adaptive tool we seek for generating mother wavelet waveforms and wavelet-based features.

Fast wavelet transform (FWT) algorithm computes a discrete wavelet transform (DWT) of a given signal using a quadrature mirror filter bank (QMF). A QMF consists of a couple of high-pass and low-pass filters h and g and is related to a single mother wavelet [24]. Hence, there is a sort of mapping between waveforms and QM filters. As we restrict here to orthonormal wavelet basis, such a basis can be fully described by the filter h of the QMF. Formulas linking filters h and g, mother wavelet ϕ and the related scaling function ψ are omitted and can be found in [24].

Using analytic formula of QM filters (of a given length) is a simple way for parametrizing QMF and thus wavelet waveform. Those formulas being specific to the filter length, they do not provide general framework for QMF generation. For instance, the following equations enable us to parametrize QMF of length $L=4$ [22]:

$$\begin{aligned} i=0,3: \quad & h[i]=\frac{1-\cos(\theta)+(-1)^i\sin(\theta)}{2\sqrt{2}} \\ i=1,2: \quad & h[i]=\frac{1+\cos(\theta)+(-1)^{i-1}\sin(\theta)}{2\sqrt{2}} \end{aligned} \tag{1}$$

where $\theta \in [0,2\pi[$ is a given angle. Formulas for other lengths of QMF can also be derived but a recurrence hardly appears for different lengths.

Among several possible parametrizations, we choose the angular parametrization of QMFs proposed by Sherlock and Monro [31]. In their paper, the authors have shown that any compactly supported orthonormal wavelet can be generated using a proper set of angles $\theta_i \in [0,2\pi[$. They also demonstrated that a $2M$ filter coefficients $\{h_i\}$ can be expressed in terms of M angular filters and proposed a recursive algorithm to compute the QMF. This algorithm is briefly exposed in the Appendix. Furthermore, they proved that in order for the QM filter to generate an orthonormal wavelet basis, the constraint $\sum_i \theta_i = \pi/4$ has to be satisfied, which reduces the choice to $M-1$ free parameters. Fig. 1 shows examples of wavelets generated by the Sherlock–Monro algorithm for $L=4$ (which gives $M=1$ for orthogonal wavelets). The two upper wavelets were generated with angular parameters $\pi/2$ and $\pi/3$ and the lower ones were selected during our experiments on a toy signal classification problem. We can note that the learned wavelet waveforms are very different to the classical ones.

The advantage of Sherlock and Monro's algorithm is that it only needs recursive sums of sine and cosine for generating a QMF, regardless of the filter length. Hence, it provides an elegant and general way for parametrizing QMF. In the sequel, in order to be independent of the QMF length, we used this algorithm as QMF parametrization and for generating wavelet waveforms.

2.2. From wavelet to features and kernels

Since we want to integrate the process of extracting wavelet features into the classifier learning process, our work can be interpreted as a method for selecting the best mother wavelet and the best elements of the resulting wavelet dictionary for a classification task. As described in the sequel, we address this problem by considering a multiple kernel learning approach. Hence, we introduce kernels derived from wavelet decomposition.