SAITO, N.; COIFMAN, R. R.

On Local Feature Extraction for Signal Classification

This paper reviews the local discriminant basis *(LDB) method for signal classification problems and demonstrates its capability using a synthetic example. The LDB method rapidly selects an orthonormal basis suitable for signal classification problem from a large collection of orthonormal bases. The goodness of each basis in this collection is measured by the "difference" (e.g., relative entropy) of time-frequency energy distributions among signal classes. Once the LDB – which maximizes this measure – is selected, a small number of most significant coordinates are fed into a traditional classifier such as linear discriminant analysis (LDA) or classification tree (CT). The performance of these statistical methods is enhanced since the method reduces the dimensionality of the problems without losing important information for classification. Moreover, since the basis functions well-localized in the time-frequency plane are used as feature extractors, interpretation of the classification results becomes easier and more intuitive than using the conventional methods directly on the original coordinate system.*

1. Introduction

Important features for signal classification problems are often local both in time and frequency. Examples of these features are edges, spikes, or transients. Therefore, it is worthwhile to pursue a method to extract local features which are useful for classification tasks.

Let $\mathcal{L} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \mathcal{Y}$ be a "training" or "learning" dataset with N pairs of signals $\boldsymbol{x}_i$ and the corresponding outputs (class names) y_i. The spaces of input signals and output responses are denoted by $\mathcal{X}$ and $\mathcal{Y}$, respectively. In our context, $\mathcal{X} \subset \mathbb{R}^n$ where n is a number of time samples in each signal, and $\mathcal{Y} = \{1, \ldots, L\}$, a set of class names. The fundamental goal of any signal classification is to extract a map $d : \mathcal{X} \to \mathcal{Y}$ between these two spaces from the given training samples so that this map will be applicable to a newly acquired dataset, called "test" dataset. The high dimensionality of the input signal space prevents the direct application of any traditional classifier (e.g., LDA, CT, artificial neural network, etc.) from being successful. These classifiers are completely overwhelmed by the volume of the data. Moreover, the original signals contain both useful and useless information for classification; the useful part is often obscured by the useless part. Therefore, an important task for classification is to extract a small number (compared to the orignal signal dimension n) of useful features from the given training signals; once this is done, the final classification procedure can be greatly improved in its accuracy and efficiency. In this paper, we only consider the map of the following form:

$$d = g \circ \Theta^{(k)} \circ \Psi, \tag{1}$$

where $\circ$ shows the composite operation. The first component, $\Psi : \mathcal{X} \to \mathcal{X}$, is an orthogonal matrix selected from a *dictionary of orthonormal bases* which will be described in the next section. Each of the basis in the dictionary is essentially a rotation of the original coordinate system. In Section 3, we explain a rapid method of selecting a good basis for classification from such dictionary. The second component, $\Theta^{(k)} : \mathcal{X} \to \mathcal{F} \subset \mathbb{R}^k$, does the reduction of dimensionality: it selects most important $k(\ll n)$ features from n features. The space $\mathcal{F}$ is called a feature space. Finally, $g : \mathcal{F} \to \mathcal{Y}$ can be any conventional classifier. The feature extractor is defined as a map $f = \Theta^{(k)} \circ \Psi : \mathcal{X} \to \mathcal{F}$, which is a main topic in this paper.

For the details of the method as well as other applications and examples, we refer the reader to [6], [8], [9].

2. A Dictionary of Orthonormal Bases

We use wavelet packet and local trigonometric bases [11] as our coordinate systems since these bases can capture the signal information localized both in time and frequency and can be computed in an efficient manner (e.g., $O(n \log n)$ for a wavelet packet basis and $O(n[\log n]^2)$ for a local trigonometric basis). Both wavelet packet and local trigonometric transforms naturally generate the binary-tree-structured subspaces of the input signal space [11]. Each node of the tree represents a subspace spanned by the specific set of orthonormal basis functions (or the so-called *atoms*) with a specific time-frequency localization characteristic. This redundant set of atoms, arranged in

the binary-tree structure, is called a *dictionary of orthonormal bases* [7]. If the depth of the tree is J, it contains more than 2^{2^J} different bases [12].

One of the key questions is, then, how to pick a basis which performs "best" for one's task from a large number of bases in such a dictionary. In order to compare the performance of each basis, we need a measure of efficiency/usefulness of a basis for that particular task. If one's task were to compress a given signal, an information cost such as entropy [4] may be appropriate since entropy measures the number of significant coordinates in a coordinate system (in fact, the flatness of the distribution of the signal's energy among the coordinates). Classification problems, however, are quite different from the compression problems. Important coordinates for compression which try to capture the main features of signals may be completely irrelevant for the classification problems where we need coordinates to see the "differences" among classes.

3. The Local Discriminant Basis

For classification, then, we need a measure of discrimination power of the nodes in the tree-structured bases. There are many choices for such discriminant measure $\mathcal{D}$ (see e.g., [1]) One natural choice is the so-called *relative entropy* (also known as *cross entropy* or *Kullback-Leibler distance*). For simplicity, let us first consider the two-class case. Let $\boldsymbol{p} = \{p_i\}_{i=1}^n$, $\boldsymbol{q} = \{q_i\}_{i=1}^n$ be two nonnegative sequences with $\sum p_i = \sum q_i = 1$ (which can be viewed as normalized energy distributions of signals belonging to class 1 and class 2 respectively in a particular coordinate system). Then, the relative entropy is defined as:

$$I(\boldsymbol{p}, \boldsymbol{q}) \stackrel{\Delta}{=} \sum_{i=1}^{n} p_i \log(p_i/q_i),$$

with the convention, $\log 0 = -\infty$, $\log(x/0) = +\infty$ for $x > 0$, $0 \cdot (\pm\infty) = 0$. It is clear that $I(\boldsymbol{p}, \boldsymbol{q}) \geq 0$ and equality holds iff $\boldsymbol{p} \equiv \boldsymbol{q}$. This quantity is not a metric since it is not symmetric and does not satisfy the triangle inequality. But it measures the discrepancy or deviation of $\boldsymbol{p}$ from $\boldsymbol{q}$. Note that if $q_i = 1/n$ for all i (i.e., q_is are distributed uniformly), then it reduces to $I(\boldsymbol{p}, \boldsymbol{q}) = \sum p_i \log p_i$, i.e., the negative of the entropy of the sequence $\boldsymbol{p}$ itself. If a symmetric quantity is preferred, one can use the *J-divergence* between $\boldsymbol{p}$ and $\boldsymbol{q}$: $J(\boldsymbol{p}, \boldsymbol{q}) \stackrel{\Delta}{=} I(\boldsymbol{p}, \boldsymbol{q}) + I(\boldsymbol{q}, \boldsymbol{p})$. Another important property of these measures I and J is *additivity*: for any j, $1 \leq j \leq n$,

$$\mathcal{D}(\boldsymbol{p}, \boldsymbol{q}) = \mathcal{D}(\{p_i\}_{i=1}^{j}, \{q_i\}_{i=1}^{j}) + \mathcal{D}(\{p_i\}_{i=j+1}^{n}, \{q_i\}_{i=j+1}^{n}). \tag{2}$$

For measuring discrepancies among L distributions, the simplest way is to take $\binom{L}{2}$ pairwise combinations of $\mathcal{D}$.

The following algorithm selects an orthonormal basis (Ψ matrix in (1)) from the dictionary which maximizes the discriminant measure on the time-frequency energy distributions of classes. We call this a *local discriminant basis* (LDB).

Algorithm 1. *Given L classes of training signals,*
Step 0: *Choose a dictionary of orthonormal bases (i.e., specify QMFs for a wavelet packet dictionary or decide to use either the local cosine dictionary or the local sine dictionary).*
Step 1: *Construct a time-frequency energy map for each class by: 1) normalizing each signal by the total energy of all signals of that class, 2) expanding that signal into the tree-structured subspaces, and 3) accumulating the signal energy in each coordinate.*
Step 2: *At each node, compute the discriminant measure $\mathcal{D}$ among L time-frequency energy maps.*
Step 3: *Prune the binary tree: eliminate children nodes if the sum of their discriminant measures is smaller than or equal to the discriminant measure of their parent node.*
Step 4: *Order the basis vectors (atoms) by their discrimination power and use $k (\leq n)$ most discriminant coordinates for constructing classifiers.*

The pruning process in Step 3 is fast, i.e., $O(n)$ if the measure $\mathcal{D}$ is additive as (2). After this step, we have a complete orthonormal basis LDB. We have now the following proposition:

Proposition 1. *The basis obtained by Step 3 of Algorithm maximizes the additive discriminant measure $\mathcal{D}$ on the time-frequency energy distributions among all the bases in the dictionary obtainable by the divide-and-conquer algorithm.*

See [6,Chapter 4], [8] for the proof.

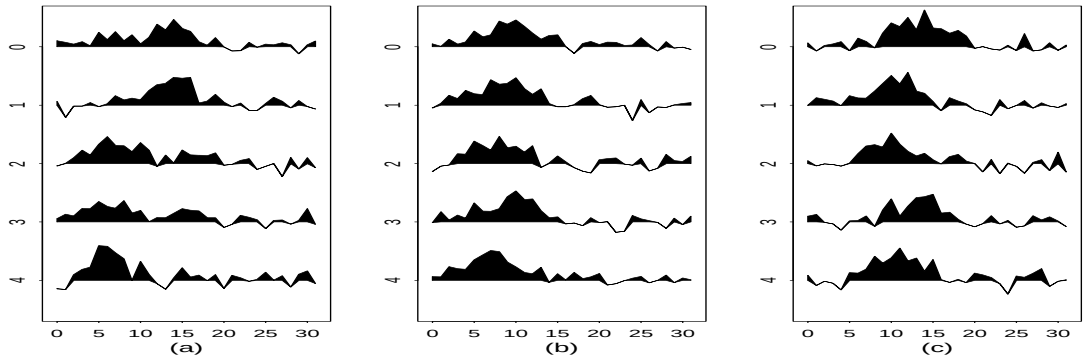

Figure 1: Five sample waveforms from (a) Class 1, (b) Class 2, and (c) Class3.

Step 4 is called "feature selection" or "selection-of-variables" (the operator $\Theta^{(k)}$ in (1)) and reduces the dimensionality of the problem from n to k (hopefully $k \ll n$). Thus many interesting statistical techniques (which are often computationally too expensive for n dimensional problems) may become feasible. How to select the truly-best k is a tough interesting question as described in the literature such as [10, Section 3.6], [4]. Some information-theoretic criteria, such as the minimum description length (MDL) criterion [5], [7], may be a good candidate for obtaining the optimal k. Here, however, we use one of the simplest approaches: first, we order the atoms in the LDB in terms of their individual discrimination power measured by computing $\mathcal{D}$ in a coordinatewise fashion; then pick the top k most individually-discriminant atoms. See [6, Chapter 4], [8], [9] for the other choices of ordering.

4. An Example Problem

We applied the LDB method to the triangular waveform classification problem often referred to as "waveform" described in [2]. The dimensionality of the signal was extended from 21 in [2] to 32 for the dyadic dimensionality requirement of the bases under consideration. We generated 100 training signals and 1000 test signals for each class by the following formula:

$$
\begin{aligned}
x^{(1)}(i) &= uh_1(i) + (1-u)h_2(i) + \epsilon(i) \quad \text{for Class 1,} \\
x^{(2)}(i) &= uh_1(i) + (1-u)h_3(i) + \epsilon(i) \quad \text{for Class 2,} \\
x^{(3)}(i) &= uh_2(i) + (1-u)h_3(i) + \epsilon(i) \quad \text{for Class 3,}
\end{aligned}
$$

where $i = 1, \ldots, 32$, $h_1(i) = \max(6 - |i-7|, 0)$, $h_2(i) = h_1(i-8)$, $h_3(i) = h_1(i-4)$, u is a uniform random variable on the interval $(0, 1)$, and $\epsilon(i)$ are the standard normal variates. Figure 1 shows five sample waveforms from each class. We first constructed an LDA-based classifier and Classification Trees (CTs) using the training signals represented in the original coordinate (i.e., standard Euclidean) system, and computed error rates of the training dataset. Then we fed the test signals into these classifiers and computed the error rates. Next we computed the LDB (with the 6-tap coiflet filter [11]) using the training signals. Then we selected five individually-most-discriminant basis vectors, and used these coordinates to construct an LDA-based classifier and CTs. Finally the test signals were projected onto the subspace spanned by these selected LDB vectors and then fed into these classifiers. In Figure 2, we compare a few most discriminant vectors from LDA and LDB. Only top two vectors were useful in LDA in this case. The top five LDB vectors look similar to the functions h_j or their derivatives whereas it is difficult to interpret the LDA vectors. We repeated the whole process 10 times by generating the training and test datasets. The averaged misclassification rates are summarized in Table 1. The best result so far was obtained by applying LDA to the top 5 LDB coordinates. We note that according to Breiman et al. [2] the Bayes error of this example is about 14 %.

5. Conclusion

We have reviewed the LDB method which constructs adaptive local orthonormal bases for classification problems. The basis functions generated by this algorithm can capture relevant local features (in both time and frequency) in data. The LDB provides us with better insight and understanding of relationships between the essential features of the input signals and the corresponding outputs (class names), and permit us to build rudimentary data-driven models. Therefore, they can enhance both traditional and modern statistical methods.

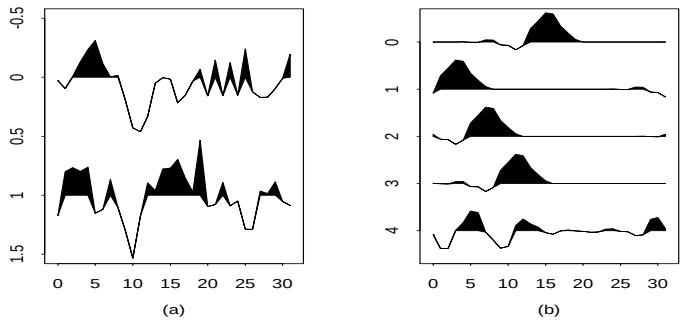

Figure 2: Plots from the analysis of the example "waveform": (a) Top 2 LDA vectors. (b) Top 5 LDB vectors.

Method	Error rate (%)	
	Training	Test
LDA on STD	13.33	20.90
CT on STD	6.33	29.87
LDA on LDB5	14.33	**15.90**
CT on LDB5	7.00	21.37
CT on LDB	7.33	23.60

Table 1: Misclassification rates of the example "waveform". In Method column, STD, LDB5, and LDB represent the standard coordinates, the top 5 LDB coordinates, and all the LDB coordinates, respectively. We do not show the error rates of LDA on all the LDB coordinates since this is the same as the ones of LDA on STD theoretically. The smallest error on the test dataset is shown in bold font.

Acknowledgements

The authors would like to thank Prof. Andrew Laine at University of Florida for providing us the digitized photographs of the Brodatz textures.

6. References

1 Basseville, M.: Distance measures for signal processing and pattern recognition; Signal Processing, **18** (1989), 349–369.

2 Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees; Chapman and Hall, Inc., New York 1993.

3 Coifman, R.R., Wickerhauser, M.V.: Entropy-based algorithms for best basis selection; IEEE Trans. Inform. Theory, **38** (1992), 713–719.

4 Cover, T.M.: The best two independent measurements are not the two best; IEEE Trans. Syst. Man Cybern., **SMC-4** (1974), 116–117.

5 Rissanen, J.: Stochastic Complexity in Statistical Inquiry; World Scientific, Singapore 1989.

6 Saito, N.: Local Feature Extraction and Its Applications Using a Library of Bases; PhD thesis, Dept. of Mathematics, Yale University, New Haven, CT 06520 USA, Dec. 1994.

7 Saito, N.: Simultaneous noise suppression and signal compression using a library of orthonormal bases and the minimum description length criterion; in Wavelets in Geophysics (Foufoula-Georgiou, E., Kumar, P., eds.), chapter XI, pages 299–324. Academic Press, San Diego, CA 1994.

8 Saito, N., Coifman, R.R.: Local discriminant bases and their applications; J. Mathematical Imaging and Vision, **5** (1995), invited paper, to appear.

9 Saito, N., Coifman, R.R.: Local feature extraction for classification and regression using a library of bases; in preparation.

10 Weiss, S.M., Kulikowski, C.A.: Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems; Morgan Kaufmann Publishers, Inc., San Francisco, CA 1991.

11 Wickerhauser, M.V.: Adapted Wavelet Analysis from Theory to Software; A K Peters, Ltd., Wellesley, MA 1994.

Addresses: Dr. Naoki Saito, Schlumberger-Doll Research, Old Quarry Road, Ridgefield, CT 06877, USA, `saito@ridgefield.sdr.slb.com`.

Prof. Dr. Ronald R. Coifman, Department of Mathematics, Yale University, New Haven, CT 06520, USA, `coifman@math.yale.edu`.