Selection of Best Bases for Classification and Regression

Ronald R. Coifman[1] and Naoki Saito[1,2]

[1] Department of Mathematics, Yale University, New Haven, CT 06520

[2] Schlumberger-Doll Research, Old Quarry Rd., Ridgefield, CT 06877

Abstract — **We describe extensions to the "best-basis" method to select orthonormal bases suitable for signal classification (or regression) problems from a collection of orthonormal bases using the relative entropy (or regression errors). Once these bases are selected, the most significant coordinates are fed into a traditional classifier (or regression method) such as Linear Discriminant Analysis (LDA) or a Classification and Regression Tree (CART). The performance of these statistical methods is enhanced since the proposed methods reduce the dimensionality of the problems by using the basis functions which are well-localized in the time-frequency plane as feature extractors.**

I. Summary

The *best-basis* algorithm of Coifman and Wickerhauser [3] was developed mainly for signal compression. This method first expands a given signal into a *dictionary* of orthonormal bases, i.e., a redundant set of wavelet packet bases or local sine/cosine bases having a binary tree structure. The nodes of the tree represent subspaces with different time-frequency localization characteristics. Then a complete basis called a *best basis* which minimizes a certain information cost function (e.g., entropy) is searched in this binary tree using the divide-and-conquer algorithm. This cost function measures the flatness of the energy distribution of the signal so that minimizing this leads to an efficient representation (or coordinate system) for the signal. Because of this cost function, the best-basis algorithm is good for signal compression but is not necessarily good for classification or regression problems.

For classification, we need a measure to evaluate the discrimination power of the nodes (or subspaces) in the tree-structured bases. There are many choices for the discriminant measure $\mathcal{D}$ (see e.g., [1]). For simplicity, let us first consider the two-class case. Let $p = \{p_i\}_{i=1}^n$, $q = \{q_i\}_{i=1}^n$ 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 coordinate system). One natural choice for $\mathcal{D}$ is *relative entropy*: $D(p,q) \triangleq \sum_{i=1}^n p_i \log(p_i/q_i)$. If a symmetric quantity is preferred, one can use the *J-divergence* between p and q: $J(p,q) \triangleq D(p,q) + D(q,p)$. The measures D and J are both *additive*: for any j, $1 \le j \le n$, $\mathcal{D}(p,q) = \mathcal{D}(\{p_i\}_{i=1}^j, \{q_i\}_{i=1}^j) + \mathcal{D}(\{p_i\}_{i=j+1}^n, \{q_i\}_{i=j+1}^n)$. For measuring discrepancies among L distributions, one may take $\binom{L}{2}$ pairwise combinations of $\mathcal{D}$. The following algorithm selects an orthonormal basis (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: normalizing each signal by the total energy of all signals of that class, expanding that signal into the tree-structured subspaces, and 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 functions by their discrimination power and use* $k\,(\ll n)$ *most discriminant basis vectors for constructing classifiers.*

For regression problems, we use the same algorithm by modifying Step 2 and 3 above. In Step 2, we compute the prediction (or regression) error at each node instead of the time-frequency energy distributions. In Step 3, we prune the binary tree by comparing the prediction errors of each parent node and the union of its two children nodes: eliminate the children nodes if their prediction error is larger than their parent node. We call the basis so obtained a *local regression basis* (LRB). One disadvantage is that the prediction error is not an additive measure so that the algorithm is slower than the LDB algorithm.

We tested our method using the triangular waveform classification (three-class problem) described in [2]. We first generated 100 training signals and 1000 test signals for each class. Then, we supplied the raw signals to LDA and CART and obtained the misclassification rates 20.90%, 29.87%, respectively, using the test signals. Finally, we computed the LDB from the wavelet packet dictionary with the 6-tap coiflet filter, and supplied five most discriminant coordinates to LDA and CART. The misclassification rates become 15.90% and 21.37%. Note that the Bayes error of this example is about 14% [2]. The details as well as other examples and applications of LDB/LRB can be found in [4], [5], and [6].

References

[1] M. Basseville, *Distance measures for signal processing and pattern recognition*, Signal Processing **18** (1989), no. 4, 349–369.

[2] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Chapman & Hall, Inc., New York, 1993, previously published by Wadsworth & Brooks/Cole in 1984.

[3] R. R. Coifman and M. V. Wickerhauser, *Entropy-based algorithms for best basis selection*, IEEE Trans. Inform. Theory **38** (1992), no. 2, 713–719.

[4] R. R. Coifman and N. Saito, *Constructions of local orthonormal bases for classification and regression*, Comptes Rendus Acad. Sci. Paris, Série I **319** (1994), no. 2, 191–196.

[5] N. Saito, *Local Feature Extraction and Its Applications Using a Library of Bases*, Ph.D. thesis, Dept. of Mathematics, Yale University, New Haven, CT 06520 USA, 1994.

[6] N. Saito and R. R. Coifman, *Local discriminant bases*, Mathematical Imaging: Wavelet Applications in Signal and Image Processing (A. F. Laine and M. A. Unser, eds.), Jul. 1994, Proc. SPIE 2303.