This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LSP.2018.2890765, IEEE Signal Processing Letters

Supervised Sparse Coding with Decision Forest

Yan Huang, Yuhui Quan* and Tao Liu

***Abstract*—By jointly conducting sparse coding and classifier training, supervised sparse coding has shown its effectiveness in a variety of recognition tasks. However, the existing supervised sparse coding methods often consider linear classification, which limits their discrimination in handling highly-nonlinear data. In this paper, we propose a new supervised sparse coding model by incorporating decision tree classifiers. Since decision trees can well deal with the non-linear properties of data, the introduction of decision trees to sparse coding can noticeably improve the discrimination of coding. Meanwhile, sparse coding is able to produce sparse de-correlated features that decision tree is favor of. For further improvement, we close the loop of sparse coding and decision tree learning with an ensemble framework, which alternatively learns a dictionary for sparse coding and a decision tree for classification. The resulting series of decision trees as well as series of dictionaries are used to construct a decision forest for classification. The proposed method was applied to face recognition and scene classification, and the experimental results have demonstrated its power in comparison with recent supervised sparse coding methods.**

***Index Terms*—sparse coding, dictionary learning, decision tree, decision forest, classification.**

I. Introduction

In recent years, sparse coding has become one of the most popular technologies for data analysis. By finding the most succinct yet effective representation of data in the coding space, sparse coding is capable of discovering and capturing the intrinsic subspaces of data. With such a capability, sparse coding has been widely used in pattern recognition, such as feature selection [1], subspace clustering [2], image analysis [3] and classification [4–7].

Given a set of input signals, the aim of the sparse coding is to represent each signal by the linear combination of a few atoms in a fixed or learned dictionary. The coefficients of the linear combination are expected to be sparse, *i.e.*, as few atoms as possible are selected to represent the signal, and thus the coefficients are also referred to as sparse codes. Generally, the sparse coding problem, *e.g.* the very popular method K-SVD [8], can be formulated as:

$$\min_{\boldsymbol{D},\boldsymbol{C}} \|\boldsymbol{Y}-\boldsymbol{D}\boldsymbol{C}\|_F^2, \quad \text{subject to} \quad \forall i, \|\boldsymbol{c}_i\|_0 \leq T, \qquad (1)$$

where $\boldsymbol{Y} \in R^{Q\times P}$ is a data matrix consisting of P signals, $\boldsymbol{C} = [\boldsymbol{c}_1,\ldots,\boldsymbol{c}_P] \in R^{M\times P}$ is the coding matrix which collects the corresponding sparse codes of $\boldsymbol{Y}$ as columns, and $\boldsymbol{D} \in R^{Q\times M}$ is the dictionary to be learned for maximizing the efficiency of sparse coding. The ℓ_0 pseudo norm $\|\cdot\|_0$ serves as a sparsity measure by counting the number of nonzero entries, and the threshold T controls the sparsity degree as well as the dimensionality of the coding results.

The above sparse coding model only minimizes the data reconstruction error of coding, which may make the results short of discriminability in classification tasks. To enhance the discrimination of the features from sparse coding, the supervised sparse coding approaches, *e.g.* [9–11], joint the sparse coding process and the classification process by enforcing some misclassification penalties on sparse codes when learning dictionaries. In the existing approaches, the linear classification is often used, which limits the power of these approaches to handle the data with highly-nonlinear properties, *e.g.* the linear classifier used in [9, 11] is not suitable for classifying nonlinearly-separable data, and the Fisher discriminant used in [12] is not suitable for characterizing mixtures of Guassians.

To further improve the performance of supervised sparse coding in classification, in this paper we propose a new supervised sparse coding method by incorporating decision tree classifiers in an ensemble framework, which alternatively conducts the sparse coding with dictionary learning and the decision tree construction. The basic idea is that the sparse codes under the learned dictionary are used as attributes for training a decision tree in the current stage, and the important attributes are identified during the construction of the decision tree for learning a better dictionary in the next stage of sparse coding. Repeating such a process results in multiple learned dictionaries as well as decision trees, which are then used to construct a sparse coding based decision forest via voting. The proposed method is applied to face recognition and scene classification, and the experimental results have shown the effectiveness of the proposed method.

The benefits of incorporating sparse coding and decision tree are two-way. On the one hand, by generating a tree model with an iterative attribute splitting process, the decision tree can well handle the nonlinear properties of data in classification. Moreover, the decision tree enjoys both the discriminability and the robustness to deal with irrelevant features [13]. Thus, by utilizing the feedback from the decision tree, the effectiveness of the sparse coding process can be significantly improved. On the other hand, the construction of a decision tree is actually to divide the feature space with disjoint hyper-rectangles. As a result, the lower correlation of the dimensions of input signals have in the feature space, the shallower and the more efficient a decision tree becomes. Fortunately, the sparse coding with dictionary learning can effectively de-correlate the input data in the coding space, as the data are aligned to their subspaces in sparse representations with the learned dictionary atoms as the coordinate axes. In other words, the dimensions of the features from sparse

All the authors are from School of Computer Science and Engineering, South China University of Technology, Guangzhou, 510006, China. Yuhui Quan would like to thank the support by National Natural Science Foundation of China (Grant No. 61602184, 61872151), Natural Science Foundation of Guangdong Province (Grant No. 2017A030313376), Science and Technology Program of Guangzhou (Grant No. 201707010147), and Fundamental Research Funds for the Central Universities (x2js-D2181690). Asterisk denotes the corresponding author.