Feature Learning for Datasets with Domain Variability

Michael Wilson

November 2022

Abstract

We present a general machine learning framework for simultaneous alignment, feature learning, and classification. We present results on real and simulated data.

1   Introduction

Feature learning after alignment gives simple, interpretable models with high test set accuracy for uni-modal shape distributions, but doesn't work well for multi-modal shape distributions. Shape distances are useful for separating multi-modal shape distributions into uni-modal clusters, but don't provide as good classification accuracy, and the features aren't as interpretable.

By using shape distances to separate data into uni-modal clusters, and then applying alignment and feature learning to each cluster individually, we are able to get simple, interpretable models with good classification accuracy on data drawn from multi-modal shape distributions.

2   Uni-modal Shape Distributions

2.1   Berkeley Growth Study

For our first example, we consider classifying the Berkeley Growth Study dataset into male and female. This is not intended to be the optimal machine learning pipeline for this data set; on the contrary, it is intended to be a simple illustration of a much more general pipeline that can be applied to any data set with domain variability. With that in mind, our pipeline for this data set will be as follows;

1) Domain Adaption: Multiple Alignment

2) Feature Extraction: Select extrema of mean function as features

3) Feature Selection: Forward selection, based on 1-cut Decision Tree test set accuracy

4) Classification: SVM with Gaussian Kernel for Classification
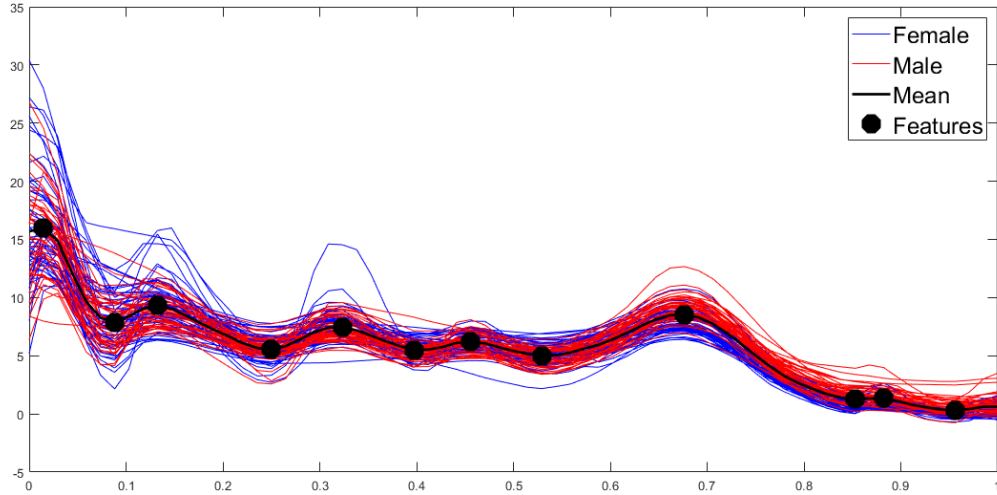
Aligned Functions

Figure 1: This is the caption

Feature:	1	2	3	4	5	6	7	8	9	10	11	12
1-cut Tree LOO Accuracy:	0.52	0.58	0.58	0.52	0.48	0.55	0.57	0.58	0.78	0.70	0.62	0.72

Table 1: Even though our feature extraction picks out 12 features, only features 9, 10, and 12 achieve higher than 0.7 percent test set accuracy. These scores are used to order our features for forward selection. Note that 57% of the labels are female, so 57% accuracy is no better than chance.

Our first step is to perform multiple alignment on our functions, and calculate a (centered) shape mean for the data. Given our shape mean, there are many ways to perform feature extraction. For simplicity, we will just treat the values of our aligned functions at time points corresponding to the local extrema of our shape mean as our features. The results of these steps are presented in Figure 1.

After alignment, our feature extraction method finds 12 features, denoted with black dots in Figure 1. Now we wish to determine which of our features are informative for classification. There are many ways to measure the informativeness of a feature, with classification accuracy of a 1-cut decision tree being among the simplest. The results of these models are presented in Table 1. Thus, our next step is to use forward selection (forward w.r.t. Leave-One-Out classification accuracy of 1-cut decision trees) to select a subset of maximally informative features, defined in terms of 10-fold cross validation performance.

Figure 2 shows the results of running a Gaussian Kernel SVM using only features 9 and 12, the two most common features selected in our validation models (in fact, selected by our learner in every validation model). Our learner achieves just over 86% Leave-One-Out test set accuracy. Furthermore, the features are very interpretable; feature 9 might be called an 'early puberty high', while feature 12 might be called a 'late puberty low'. Males tend to have higher growth rates at both of these time points, and a classifier built on
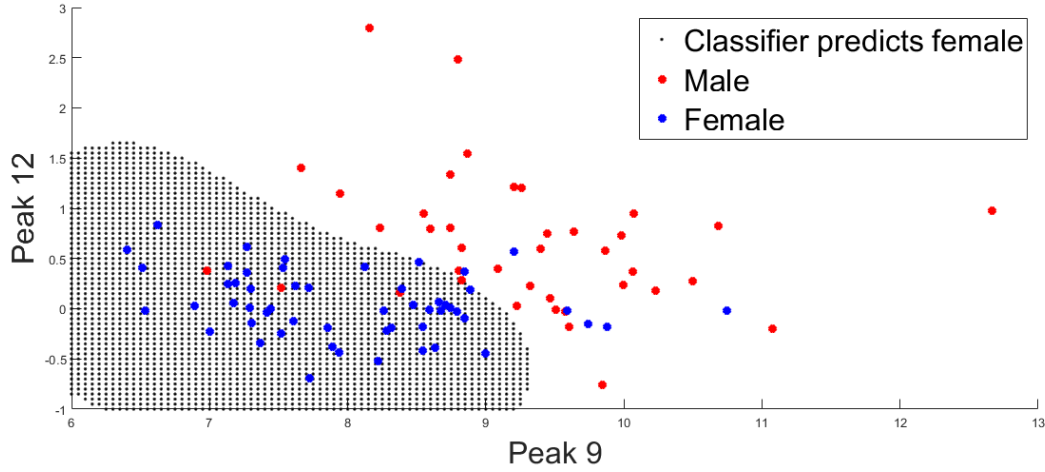
Figure 2: This is the caption

these two features gets high classification accuracy.

A pipeline using the same model selection criterion, using elastic pairwise distances as features yields a Leave-One-Out classification accuracy of 74.19%. While this classifier still does better than chance, it has significantly lower classification accuracy, and the features aren't as interpretable. We can show which growth curves are used in the classifier, but are left to speculate what features of a particular growth curve are informative. Note, these results are for a linear SVM; a Gaussian Kernel SVM on our shape distances performs much worse.

There is much more to say about this analysis, but our purpose here is simply to illustrate why this particular pipeline is useful for datasets which contain information about class in both shape and non-shape features.

3 Multi-modal Shape Distributions

3.1 Berkeley Growth Study

3.2 DTMRI Data

The basic pipeline is as follows;

1) Domain Adaption: Pairwise Functional Optimal Transport

2) Feature Extraction: k-means

3) Feature Selection: Select optimal k based on 1-cut Decision Tree test set accuracy

4) Classification: Decision Tree, $maxNumCuts = 4$