An empirical analysis of the effectiveness of structure in classification problems

Michael Hatch
A09910883

Abstract

Structural learning is an extension of multiclass learning that allows classification algorithms to consider the context of a data point in classifying it, rather than relying solely on individual data points to make predictions. Structural learning allows for a computational modeling of important concepts in language, top-down visual processing, and other programmatic approximations of meaning in real-world contexts.

In this paper, we will compare a “standard” multiclass approach to structured learning for optical character recognition (OCR). We will train a Random Forest Classifier (RFC) with 10-fold cross-validation on a 5000-character dataset representing words (randomly subsampled from a 50,000-character set), and compared with the fixed-point structural learning algorithm run on the entire dataset.

Methods

We will be working with a dataset prepared by Stanford University[1] for OCR. The dataset includes approximately 50,000 letters, and approximately 130 features related to the position of the letter in the word, the next letter in the sequence, and the word it is a part of. In the interest of time, the RFC was trained with a 10% random subsampling of the dataset.

By way of comparison, the Fixed Point[2] algorithm’s demo was run on the full dataset, although Fixed Point trained on a sample of roughly the size of the full RFC subset and tested on the rest of the data. The demo was run largely as presented, however the window size was varied from 0-4, and cross-validation was reduced to single-fold due to time constraints.

Results

Results are provided in the table below. With comparable training sizes, and much less optimization and tuning, Fixed Point did much better than RFC on a much larger testing set. If accuracy is our only concern, then Fixed Point is clearly the superior algorithm.

Algorithm	Accuracy	Parameters
RFC	79.6%	Optimal # of Trees: 189;
Fixed Point	89.3%	Max Window Size: 4

Discussion

Although Fixed Point gives superior accuracy on a larger dataset with less tuning and comparable training to RFCs, it also took much longer to compute (less than an hour for RFC vs. over 3 hours for Fixed Point). However, due to the difference in the size of the training sets, it's difficult to say that Fixed Point is slower than RFC.

With time being an indeterminate factor, the conclusion of this study is that structure is likely an improvement in all cases over naive multiclass, and should be implemented when available.

References:

1: http://ai.stanford.edu/~btaskar/ocr/

2. http://ai.stanford.edu/~btaskar/ocr/

RFC Code:

https://github.com/TrackAddict/Cogs185/blob/master/Homework%202.ipynb