ICDAR2013 - Gender Prediction from Handwriting

Anil Thomas

1. Summary

This document describes the winning model in the ICDAR2013 Gender Prediction Contest organized by Kaggle. The provided data consisted of handwriting samples from 475 individuals. Two samples in Arabic and two samples in English were collected from each writer. The gender of 282 writers was identified in the training data. The objective of the contest was to predict the gender of the remaining 193 writers.

2. Features Selection / Extraction

The model was built using the features provided along with the images. Details of the feature extraction process are available in Hassaïne et al. (2012). The 7066 features given in the data set were ranked according to the relative influence determined by the Gradient Boosted Decision Trees (GBDT) algorithm. For every tree, the relative influence of a feature is computed as the empirical improvement gained by splitting on that feature. The values across all the generated trees are then averaged to determine the importance of a feature.

After ranking the features in descending order of importance, the first N features were selected for subsequent regressions. From cross validation, a value of 80 was found to be near optimal for N.

As an example of the association between handwriting features and gender, Figure 1 shows that the female writers tend to have lower values of the “directions_hist1a2a3a4a5a6a7a8a9a10_220.7.” feature and higher values of the "directions_hist1a2a3a4a5a6a7a8a9a10_220.187." feature on an average.

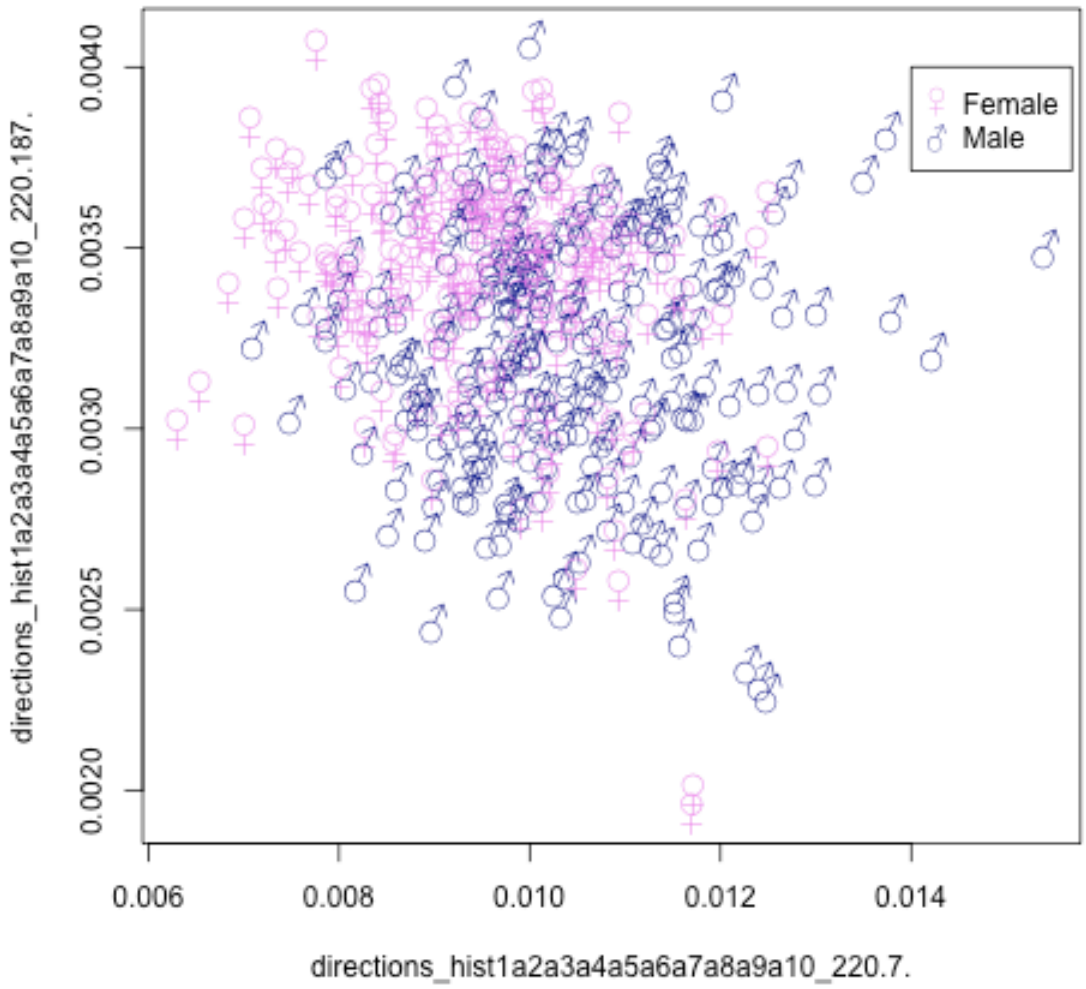

Figure 1. Influence of Gender on Two of the Features Selected as Examples.

3. Modeling Techniques and Training

The given training data had both Arabic and English samples. Before training the models, the data set was divided into two parts – one containing Arabic samples and the other one with English samples. Regression was performed separately on each subset. The rationale was that the same geometrical feature could have different implications depending on the script. However, some of the benefit of doing separate regressions is cancelled out due to the number of samples available for each regression getting effectively halved. This makes the regressions more susceptible to overfitting[1]. Cross validation showed a small improvement in accuracy from separating out the samples. The outcomes for the Arabic samples in the test data were predicted after training on the Arabic training subset and those for the English samples were predicted after

[1] In a non-contest setting, this is an issue that can be overcome by collecting handwriting samples from more subjects.

training on the English subset. The probability outcomes were then averaged to arrive at the final predictions for each writer.

A publicly available implementation of GBDT, the gbm package in R was used for performing regressions. The hyper parameters were tuned with the aid of N-fold cross validation.

4. How To Generate the Solution

1. Make sure that R and the gbm package are installed on the target machine.
2. Copy the source files (genpred.r and features.csv) to the same directory containing the data set provided as part of the contest (train.csv, train_answers.csv and test.csv).
3. Launch R and run genpred.r.
4. After the program exits in a few minutes, a file called subm.csv containing the test predictions may be found in the same directory.

5. Additional Comments and Observations

Various error metrics obtained from cross validation are given in the table below. In order to discriminate between the two genders, a cutoff value of 0.5 was chosen for the probability outcomes. The last row shows the metrics after the probability values for Arabic and English were averaged.

	Precision %	Recall %	TrueNegRate%	Accuracy %	Log Loss
Arabic	65	85	67	74	0.7605
English	73	80	78	79	0.6874
Combined	76	95	78	85	0.3585

Table 1. Error Metrics on Cross Validation Data.

Note that the hyper parameters were tuned to minimize the logarithmic loss on the cross validation set. The error metrics on an unseen data set are likely to be not as good. The disproportionate amount of improvement in logarithmic loss obtained by averaging the results for both the languages may be attributed to idiosyncrasies of the metric. Confident false predictions are heavily penalized in this metric. Averaging of the probability outcomes from the two language subsets causes extreme outcomes to be pulled towards the mean, thereby dramatically improving the logarithmic loss.

An alternative approach to capitalize on the availability of samples in multiple languages from the same subjects would be as follows: Treat the same features obtained from multiple language samples as different. For example, instead of having a single *tortuosityDirectionHist10* feature, there could be a *tortuosityDirectionHist10**Arabic*** feature and a *tortuosityDirectionHist10**English*** feature. This approach does have the drawback of doubling the dimensionality of the feature space.

The winning model used the same feature set for both the languages. A possible improvement would be to select the features separately for each language.

6. References

[1] Hassaïne, A., Al-Maadeed, S. and Bouridane, A., *A Set of Geometrical Features for Writer Identification*. Neural Information Processing. Springer Berlin/Heidelberg, 2012.