Gender Prediction from Handwriting

Final Project for ISE 5103

Student X

University of Oklahoma, Industrial and Systems Engineering, Norman
studentx@ou.edu

Executive Summary

Writer classification has been an interesting field of handwriting analysis due to its many useful applications. This project is based on a 2013 Kaggle data analysis competition to predict if a handwritten document has been produced by a male or female writer. The competition was based on correctly classifying writers based on gender when given four writer samples per writer in two different languages, Arabic and English. The raw photo images are pre-processed for the competition and the features were based on a geometric feature extraction method from previous contests (Hassaïne, Al-Maadeed, & Bouridane, 2012). The goal of this project is to understand the impact feature selection had on the overall accuracy of the analysis when comparing different models. For purposes of this project, the winner's feature selection method used Gradient Boost Decision Trees (GBDT), and was used as a benchmark to compare different combinations of features and classification models.

The data set used for this project was the training set provided for the competition and was composed of 282 writers with 4 samples per writer which summed up to 1,128 total observations. The data was scaled and centered with all near zero variance columns removed. The data was then split by writer with 80% of the writers in the training set and 20% of the writers in the test set. The data was split further by language under the hypothesis that language would have an impact on gender prediction. To reduce the 4,332 features, a wrapper random forest feature selection algorithm known as Boruta was used on the full, English, and Arabic training set to extract three different sets of features. These three different feature sets were used along with the winner's features to build four different classification models.

Four models were used to predict gender based on four different sets of features. These four models are: Logistic Regression (LogR), Decision Trees (DT), Gradient Boost Decision Trees (GBDT), and Random Forests (RF). The average predicted probabilities from the writer's four samples were then compared to the true results to calculate the accuracy of each feature-model combination. The results show that none of the features selected with the Boruta algorithm were more powerful than the winner's features. The highest accuracy achieved based on Arabic features applied on all training data with a GBDT model had a 71.43% accuracy which is less than the winner's features with the same GBDT model accuracy of 80.35%. This results shows the impact of feature selection on model accuracy as different features used on the same model had drastically different results. Another interesting results was that Arabic features had the most accurate results from the Boruta algorithm which could suggest a trace language accent present in English samples which allows Arabic features to accurately classify gender regardless of language.

Areas of future work would focus on different feature selection techniques such as Minimum Redundancy Maximum Relevance (mRMRe) or decision tree variable importance. An extremely interesting hypothesis as a result of this project would be the affect that extracting Arabic features from GBDT model, and then tested on all data would compare with the winner's results. This project successfully proved that feature selection has a high impact on model accuracy and was the key to success for the winner in the kaggle competition.

Problem Description

Background

Handwriting analysis is an old area of research of extracting information from handwritten text to draw conclusions about the author. For example, handwriting analysis can be used to predict personality traits of the author such as social skills or work habits (Prasad, Singh, & Sapre, 2010). Other applications include forensic analysis of handwriting to help identify traits of the authors used in other applications like criminal investigations. Several types of analyses can be associated with handwriting analysis depending on the information extracted, and can be separated by recognition, interpretation, and identification. Handwriting recognition focuses on transforming the graphical marks on the page into the symbolic representation of a particular language. Handwriting interpretation determines the meaning of the body of handwriting such as a handwritten address. Handwriting identification is the task of extracting key features of writing used to classify writers (Plamondon & Srihari, IEEE Transactions). This report will focus on the classification aspect of handwriting identification. With the advancement of technology, handwriting analysis has become automated to provide fast and accurate results that allows us to draw deeper insights into the information we can extract from handwriting.

Kaggle Competition

This report is based around a 2013 kaggle competition to predict gender from handwriting (Kaggle, 2014). Given a data set of handwriting, the competition asks participants to classify writers by gender and provided predicted probabilities for each writer. The competitors were evaluated using a log loss metric which was very sensitive to overconfident predictions. The competition provides data from 475 writer samples in both English and Arabic. Since the competition only provided the true gender for 282 writers in their training set, the focus on this paper will only deal with that data set. In order to assist participants, digital image processing is not in the scope of the competition and the competition provides features extracted from the image data. These features are described in detail in a previous publication in 2012, and will be discussed further in this report (Hassaïne, Al-Maadeed, & Bouridane, 2012). One key challenge to this competition was feature selection as the features provided by the competition greatly outnumbered the number of observations and made this a “big data” problem.

Problem Definition

The goal of this report is to explore the impact feature selection has on prediction accuracy. Due to the large number of features relative to number of observations, feature selection is a very important step in this problem and is believed to have a high impact on classification accuracy. The goal of the project is to determine if different features impact accuracy when used to build controlled models. To give the project perspective, the winner of the kaggle competition's solution method was analyzed and used as a benchmark throughout the rest of the report to compare models and features. This project will provide key insight to the winner's success, and create a foundation for approaching similar data analysis problems.

Benchmark

Throughout the project the output of the model created is compared to the winner of the event. The winner of the event will be addressed just as winner in future reference. The winner was able to produce a model with the smallest log loss value amongst all the competitors. The winner reduced his data features to just 80 from 7,066 using a gradient boosted decision tree algorithm. With the help of this model, he ranked his features using their relative influence in descending order. The first 80 features were selected for final model building. There were 35 direction, 8 curvature and 37 chain code features selected by the winner. It was interesting to see that tortuosity was removed completely.

Data Exploration

The features applied to the data set in this competition is described in, "A set of geometrical features for writer identification" (Al-Máadeed, Ayouby, Hassaïne, & Al'Jaam, 2012). The data set was collected by Qatar University and is saved as Qatar University Writer Identification dataset (QUWI). It contains both Arabic and English handwriting. For the purpose of the contest hosted in Kaggle, only 475 of the 1,017 writers were taken. Each writer produced four sets of samples, two in English and two in Arabic. Each participant wrote similar text in one sample of English and one sample of Arabic. For the other two samples the participants were asked to write using their own imagination. The images were acquired using an EPSON GT-S80 scanner, with a 600 DPI resolution. Images were provided in JPG uncompressed format. The training set consisted of the first 282 writers. The total observations were 1,128 for the training set. Kaggle provided us with a dataset of features extracted from all the images. Those features were curvature, chain code, direction and tortuosity. These features are the probability distribution function of several values and added to a total of 7,066 values that had a value between 0 and 1. The target variable of the dataset was male which was set to 1, and female which was set to 0. The same text samples were coded as 1 and the other was coded as 0. There was a total of 5,020 chain code, 1,106 direction, 900 curvature, and 40 tortuosity variables.

Feature

It is essential to understand the dataset before we can create models for gender prediction. The data set values were the probability distribution function extracted from the handwriting images to characterize the sample individually. The images were first binarized using the Otsu thresholding algorithm and these categories were used to describe them:

- Direction. : Direction characterizes writers well. The Zhang skeleton of the binarized images were computed. This skeleton was segmented at junction pixels and the tangent direction of the middle axis of text was measured.
- Chain codes: They are generated by browsing the contour of the text and assigning a number to each pixel according to its location with respect to the previous pixel.
- Curvature: Is one of the most widely accepted features for characterization of forensic documents and measures the curvature of the text.
- Tortuosity: This is the feature which helps experts to distinguish between fast writers who produce smooth handwriting and slow writers who produce twisted handwriting.

Data Transformation

Exploring the data it was observed that there were no missing values but there were a lot of columns with zero variance. The function nearZeroVar() was used from the Package caret to delete columns that had an overall variance near zero and could be considered empty. As a result 2,734 columns were deleted. The column page_id was deemed as not a valuable field, hence was deleted. The language field had characters in it and they were changed to numeric, Arabic=1 and English=1. For proper modeling of the data it was necessary to scale and center the values. Scale() function was used for this purpose.

Training and Test

Since the test set did not have target values it was decided to split the given train set of 282 writers. Every writer wrote four samples so there was a chance that a powerful model will recognize the writer, not the gender. It was decided to keep all the samples of the writers together in either training or test set. Of the original train set, 80% was set as the new train set and 20% was set as the new test set. The train set had 904 observations and test set had 224 observations.

To check whether a more efficient model could be created, the train and test set were again split by language. The idea was to create separate models for each language and average the predictions in hope that more accurate models could be created. The train set for both Arabic and English had 452 observations and the test set for Arabic and English had 112 observations each.

Analysis Plan

Feature Extraction

The data set had too many variables for practical model building. To see if the features were relevant to classification and to avoid overfitting, it was essential to do feature extraction. We used an R package called Boruta for feature extraction.

Boruta is a wrapper random forest feature selection algorithm and is explained in a previous publication in 2010. The algorithm iteratively removes the features which are provided by a statistical test to be less relevant than random probability. Classification is done by voting of multiple unbiased weak classifiers. These trees are independently developed on different bagging samples of the training set. Z-score values for each attribute are noted. One of the fundamental components of Boruta is the use of shadow attributes. Shadow attributes are pseudo-features that are added to the information system, and produced by taking existing features from the original data-set and shuffling the values of those features between the data points. Then the Z-score for each shadow' attribute is calculated. Both the Z-scores are compared, if the shadow Z-score is less than the original Z-score then the feature is seen as unimportant. If the Z-score is greater than the original Z-score then the features is considered important (Kursa & Rudnicki, 2010).

The function used for the combined, English, and Arabic training data set is shown below.

```
>set.seed(1)
>Boruta.train<-Boruta(male~.,data=training, doTrace=2,ntree=500)
>features<-getSelectedAttributes(Boruta.train, withTentative = TRUE)
>features
```

When calling the function, Boruta essentially runs a random forest wrapper algorithm on the entire data and runs it until it finds relevance of all the attributes. This might take hours or even days depending on the complexity of the data set. The way we can control Boruta is by specifying the number of trees it should build. Once we specify the number of trees, Boruta stops functioning when it reaches that limit. The attributes that it is not able to classify as relevant is put into a tentative list and can be accessed for future reference.

Boruta was run on the full training set and 174 variables were recognized as relevant to make up the combined feature set. It took a total of 28 hours for Boruta to finish selection. When Boruta was executed for the Arabic training set it chose 82 variables, and for English it chose 90 variables. It took more than 10 hours for feature selection of both English and Arabic individually. It was noted that the features selected for English and Arabic were quite different from each other. An even more interesting find was that the selected features for the entire training set by Boruta only had 34 common variables with the winner's features. Of the 82 variables chosen for Arabic, only 9 were common with the winner's 80 features. English had 24 common variables with the winner features. English and Arabic features only had 2 common variables between them which suggests that dividing the training set by language has a high impact on the features selected.

Model Selection

To test each set of features, four models were chosen: logistic regression, random forest, decision trees, and gradient boost decision trees. Logistic regression (LogR) is a desirable choice due to model simplicity. If a robust logistic regression model could be created, each variable's relative importance can be easily deduced from the model. Random forest (RF) is a useful model when the relationship between the target variable and predictor variables is not linear. Additionally, logistic regression and random forests are submitted on the competition website as bench marks, so they were included in the model selection process to see how they performed relative to the winner. Through research, the winner of the competition used gradient boost decision trees. To test the performance of the winner's model, a simpler form of decision trees (DT) and gradient boost decision trees (GBDT) are included in the analysis. Each of these models have advantages and disadvantages, but are all well-known classification models. In addition, all of the models used in this analysis incorporated some form of re-sampling method such as K-fold cross validation or boosting methods of the training data when building the models. This was to insure the most accurate and reliable results when comparing models.

Each model was built using training data, and then used on test data to calculate predicted values in the form of probabilities. As discussed earlier in this report, each writer provides four samples of writing that are used to predict the writer's gender. Only one predicted value per writer is needed to assess model performance, so all four writer samples' predicted values are averaged together. The project goal is to asses feature selection impact on model accuracy, so four different sets of features are tested on every selected model as shown in the figure below.

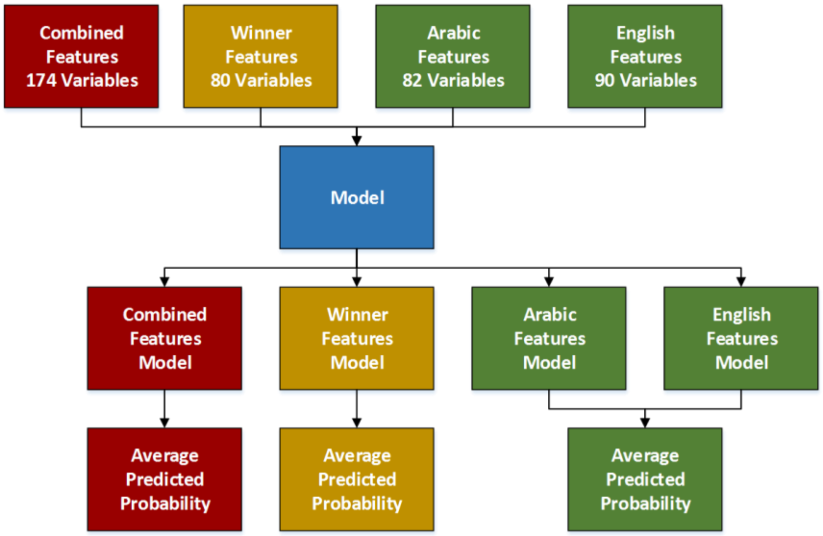
Combined Features 174 Variables
Winner Features 80 Variables
Arabic Features 82 Variables
English Features 90 Variables
Model
Combined Features Model
Winner Features Model
Arabic Features Model
English Features Model
Average Predicted Probability
Average Predicted Probability
Average Predicted Probability

Figure 1 Diagram representation of models created by feature set.

As you can see from the figure, the same process is repeated for all models. The combined features and winner features have the exact same process in that they use all training data to build the model then use the model to predict gender on the test data. Each writer in the test data has four sample predicted probabilities that are averaged together. However, English and Arabic features in green are treated differently because of the need to split the data by language. Each model built in this process will only be built with their respective language and features. This will result in only two predicted probabilities per writer per model. Once both models are built, the average of each model’s predicted probabilities are used to evaluate performance. This was all done under the assumption that language would have a high impact on feature selection and result in completely different models.

Validation Plan

Prediction accuracy is used to evaluate the performance of the combination of feature selections and models. Prediction accuracy can be calculated by the equation below,

$$Accuracy\ \% = \frac{True\ positive + True\ negative}{Total\ population}$$

True positives refer to the correctly predicted male values, and true negatives refer to the correctly predicted female populations. For example, the following R code creates a confusion table. The values from this confusion table can be used to calculate accuracy.

```
>tableconf <- confusionMatrix(predictFactor, test.indexAnswers$male, dnn =
      c("Predict",                                                "Answer"))
>tableconf

```

The confusion table is shown below and has the respective true positive and true negatives based on test data predictions. Interpreting the table, the model predicted 22 females and 34 males in the test data set. Using the true answers to calculate results, the matrix represents the overall accuracy of the model.

```
Confusion Matrix and Statistics

       Answer
Predict  0  1
      0 14  8
      1 10 24
```

Using the accuracy equation, the accuracy for this particular combination of features and model is 67.86%. This statistic is calculated for each model and feature combination and is used to provide a benchmark to evaluate how different feature selections across different models affect prediction accuracy.

Results

The results of the analysis plan discuss above are shown in Table 1. Since accuracy is the evaluation metric for this project, the accuracy from the selected feature and model combination are shown. Each column represents the respective features used in each of the models. The column with combined uses the combined features from Boruta algorithm and applied those features to all of the training data. The Averaged column is the average of the English and Arabic predicted probabilities. As can be seen from the table, the winner's selected features result in the highest accuracy of 80.35% using GBDT. In fact, the winner's features generally had the highest accuracy regardless of the model selected with the exception of random forests which resulted in a tie. This supports the claim that feature selection had the highest impact on model accuracy. The best model produced with the Boruta algorithm selected features was created using the averaged model of English and Arabic features on random forests with an accuracy of 69.64%. When comparing the English and Arabic language models, Arabic overwhelming performed better than the English features at predicting gender. However, when the results of the English and Arabic features were averaged together, the results predicted probabilities were stronger for the Random Forests model than when they were alone. It appeared the Arabic features used on Arabic training data yielded the highest accuracy, which suggests they were the strongest features captured.

Table 1 Accuracy results from analysis

Features	Winner	Combined	Averaged	English	Arabic
Logistic Regress	76.78%	67.85%	62.50%	62.50%	67.85%
Pruned Decision Trees	69.64%	64.28%	62.50%	51.78%	64.28%
Gradient Boost DT	80.35%	66.07%	67.85%	60.71%	69.64%
Random Forest	69.64%	64.28%	69.64%	62.50%	67.85%

As a matter of interest, the Arabic features were used on all training set just like the combined features and the winner's features with the results shown below in Table 2. Interestingly enough, using these features resulted in the highest overall accuracy of 71.43%, but not the best when compared to the winner's 80.35%. Arabic features only had 9 common variables with the winner's feature set and definitely suggests that multiple features could be considered when predicting accuracy. There could be several reasons for this that could be explored outside of the constraints of the analysis and about the writers themselves. For example, if all of the writers are more familiar with writing in Arabic, maybe it is easier to detect gender features in the Arabic language because the writers are more comfortable writing in Arabic. Another possible reasoning behind the universality of the Arabic feature identification could be tied to which language the writers learned first. If all of the writers learned how to write in Arabic first, maybe the English letters had traces of Arabic features that the writers subconsciously displayed that analyzing English features alone could not detect. This subconscious use of Arabic features could be similar to an accent, so to speak. These results are interesting and open up opportunities for future research.

Table 2 Accuracy results from Arabic features used on all training data.

Features	Arabic - All Train
Logistic Regress	64.28%
Pruned Decision Trees	62.50%
Gradient Boost DT	71.43%
Random Forest	69.34%

Of all the features that were selected by Boruta, the majority of the features selected were directional. Tortuosity was ignored for all selections. Looking at the variable importance plot for random forest and gradient boost it was noted that curvature was the most relevant attribute for the models.

Conclusion

Issues

There were several issues encounter throughout this project that had to be overcome. Fortunately, the data itself did not need much scrubbing, but did need to be scaled. Due to some of the columns being non-numeric, some of the functions used to scale the data failed when they encountered non-numeric data. Failure to recognize that the data was not scaled after calling the function compromised the first analysis results and was a time-consuming error. In addition, the feature selection method for Boruta did not reveal better features than the GBDT algorithm and was computationally demanding with tests taking days to run. When performing this analysis, it became very useful to create functions to perform the analysis tasks and analyze the data. Otherwise, the R code would be lengthy, confusing, and very difficult to de-bug. Finally, even though every effort was made to set the random.seed option to ensure consistent results, there was some inconsistency between analysis runs. This variability of accuracy results did not change the overall conclusions, but did raise questions about the strength of the model and code that must be addressed. Recent results have been consistent across runs after a refresh of the R Environment, but have not yet been validated.

Future Work

Results proved that good feature selection is key to building a good model. The models that were created did not perform well when compared to the winner, but they did have a reasonable accuracy when compared to random chance. When making the comparison between models, the only difference was the features used, so all features were tested on the same models. Different feature selection techniques like Minimum Redundancy Maximum Relevance (mRMRe) and decision trees variable importance should be explored. The winner decided on his features using Gradient Boost Decision tree.

It will be interesting to see how accuracy changes when new models are built. Neural Network and Support vector Machines traditionally give robust and accurate models, but were not in the scope of this project. These are two great models to build future work with. Another opportunity for future work would be different evaluation metrics. The only evaluation metric used here was accuracy due to simplicity of results. More evaluation metrics should be tried out because accuracy is not the best criteria to base your judgment on when considering model performance overall. D-statistic, lift and gain charts could also be included in the work.

All the issues that were not addressed in the current project should be fixed. Even though accurate ranking was produced every time the code was run we did encounter unexplained variability with the accuracy rate results. It was also seen that the features selected for Arabic were quite significant and produced more accurate results than English features. It will be interesting to study more datasets were writers were asked to make samples in different languages in comparison with their native languages. Based on the results of this analysis, another interesting hypothesis would be to try to detect native language of the writer based on different language samples, almost like trying to detect a writer’s accent.

Conclusion

The aim of the project was to understand the effect of feature selection and its impact on model accuracy. Three different feature subsets were created based on the results from the Boruta algorithm which represented combined, English, and Arabic observations. Overall, four different models were built and tested with each feature set which included the winner’s features created from a GBDT analysis. It was observed that a GBDT model using Arabic features applied to all training data resulted in the highest accuracy of 71.43%. These results suggest that Arabic features might be present in the English language due to a writing accent from the writers. This must be validated with future work focused on a more varied data set with different native languages. Overall, the results of the project support the hypothesis that feature selection has a high impact on classification accuracy regardless of model chosen.

References

Al-Máadeed, S., Ayouby, W., Hassaïne, A., & Al'Jaam, J. M. (2012). QUWI: An Arabic and English Handwriting Dataset for Offline Writer Identification. *ICFHR*, 746-751.

Bandi, K., & Srihari, S. N. (2005). Writer demographic classification using bagging and boosting. *Proc. International Graphonmics Society Conference (IGS)*, 133-137.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 1157-1182.

Hassaïne, A., Al Maadeed, S., Aljaam, J., & Jaoua, A. (2013). ICDAR 2013 Competition on Gender Prediction from Handwriting. *Document Analysis and Recognition (ICDAR)*, 1417-1421.

Hassaïne, A., Al-Maadeed, S., & Bouridane, A. (2012). A set of geometrical features for writer identification. *Neural Information Processing*, 584-591.

Kaggle. (2014, 12 9). *ICDAR2013 - Gender Prediction from Handwriting*. Retrieved from Kaggle: https://www.kaggle.com/c/icdar2013-gender-prediction-from-handwriting

Kuhn, M., & Kjell, J. (2013). *Applied Predictive Modeling.* New York: Springer.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta package. *Journal of Statistical Software*.

Liwicki, M., Schlapbach, A., Loretan, P., & Bunke, H. (2007). Automatic detection of gender and handedness from on-line handwriting. *Proc. 13th Conf. of the Graphonomics Society*, (pp. 179-183).

Liwicki, M., Schlapback, A., & Bunke, H. (2011). Automatic gender detection using on-line and off-line information. *Pattern Analysis and Applications*, 87-92.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*.

Plamondon, R., & Srihari, S. (IEEE Transactions). Online and off-line handwriting recognition: a comprehensive survey. *Pattern Analysis and Machine Intelligence*, 63-84.

Prasad, S., Singh, V., & Sapre, A. (2010). Handwriting Analysis based on Segmentation Method for Prediction of Human Personality using Support Vector Machine. *International Journal of Computer Application* , 25-29.

Zygmunt, Z. (2013, 01 14). *Feature selection in practice*. Retrieved from FastML: http://fastml.com/feature-selection-in-practice/