Quantitative Support for Forensic Document Examination in an Open Set using RandomForests

Madeline Q. Johnson, Danica Ommen, Alicia Carriquiry

Funding Acknowledgement: This work was funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College and University of Nebraska, Lincoln.

Learning Overview: The goal of this poster is to illustrate how the statistical method of random forests can be used to quantify whether a pair of handwritten documents are from the same or different source based on clustered features in handwritten documents.

Impact on the Forensic Science Community: This presentation will impact the forensic science community by contributing to the work of current forensic statistics relating to comparing documents in an open set.

Abstract

Handwriting analysis is conducted through the expertise of Forensic Document Examiners (FDEs) by visually comparing writing samples. Through their training and years of experience, FDEs are able to recognize critical characteristics of writing to evaluate the evidence of writership. In recent years, there have been incentives to further investigate how to quantify the similarity between two written documents to support the conclusions drawn by FDEs.

One way to extract information from these documents is to extract various features within handwritten samples. Using an automatic algorithm with the `handwriter` package in R, a sample can split into "glyphs", which are small units of writing [2]. These glyphs are sorted into 40 exemplar groups or "clusters". The clusters have similar structures found in documents throughout a database with many writers. Previous work related to the number of glyphs per cluster focused on quantifying the probability a questioned document was written by one of the writers in a closed set. In these cases, all of the potential sources of the handwriting are assumed to be known [1]. This project aims to use simulated data to study how classification tools can be used to assess the within-writer versus between-writer hypotheses in an open set of documents.

Specifically, a statistical model can be used to study the proportion of these glyphs categorized within each of the 40 clusters for each document. Then, given two questioned handwritten documents, it is possible to quantify how similar the proportions across clusters are using a distance measure, such as the difference in proportions for each cluster. Since writers over time and across documents have similar writing patterns, it is expected that the proportion of glyphs classified to these clusters is comparable when written by the same person. Conversely, the proportion of glyphs by cluster will be less similar when the documents do not share the same source. A random forest is a statistical classification algorithm made up of many decision trees that is able to use these measures to classify pairs of documents as coming from the same source or different sources when trained on previous data where the true identification was known. Results of using random forest algorithms have shown clear discernment between simulated data distance measurements from the same writer and different writers.

Findings from this statistical research provide insight on another way to quantify the similarity between two questioned documents when all possible sources are unknown.

References

1. Crawford, Amy. "Bayesian hierarchical modeling for the forensic evaluation of handwritten documents." PhD diss., Iowa State University, 2020. https://lib.dr.iastate.edu/etd/18078
2. "Handwriter Package." handwriter, 2020. https://csafe-isu.github.io/handwriter/index.html.

Keywords: Statistics, Random Forest, Document Analysis