B197 Machine-Learning (ML) Approaches for Source Attribution of Forensic-Relevant Materials

Josh Dettman, PhD, Massachusetts Institute of Technology Lincoln Laboratory, Lexington, MA 02420; Amanda M. Casale, Massachusetts Institute of Technology Lincoln Laboratory, Lexington, MA 02421*

Learning Overview: After attending this presentation, attendees will better understand how Artificial Intelligence (AI) ML algorithms are enabling new applications requiring the classification of large data sets, including extracting conclusions on the source of materials using chemical and physical signature data.

Impact on the Forensic Science Community: Applications of AI and ML techniques are growing rapidly, including in forensic science. This presentation will impact the forensic science community by increasing competence in ML by introducing potential ML approaches and disseminating methods for material source attribution to the operational community and other researchers.

Source attribution of materials involved in a crime using chemical and physical signatures can be an important technique for generating investigative leads on the origination location (source) of the material and potential association with a suspect. Proof-of-concept studies of this type have been performed for a variety of materials of forensic and commercial importance. Chemometric or, more generally, ML techniques can be used to extract conclusions from the relatively high-dimensional data collected and estimate the probability that a sample originated from a specific source. This presentation will describe the development of a composite ML sourcing algorithm utilizing multiple tiers of supervised ML classification methods, data conditioning, bootstrap supplementation of reference data, and probability-weighted results fusion to estimate the probability of source membership of fertilizer samples based on chemical and physical signature data.

Reference samples of fertilizer were obtained from known commercial and industrial sources. For most sources, only single samples were available (single-lot sources); however, nine to ten samples were collected over a period of five months from two sources (multi-lot sources) to estimate signature variance over time. After evaluation of many potential signature types, the final set of signatures collected from the reference samples was comprised of reflectance spectra (color), particle morphology (size/shape), and trace element composition. This data was subjected to limited conditioning (e.g., conversion from intensity to element concentration by a calibration curve, calculation of color, and shape summary variables). Additional bootstrapped reference data for algorithm training was performed based on the signature means (by source) and largest experimentally measured variance for each signature (by source and lot). Increasing the signature variance above what was empirically measured incorporates an estimate of source variance (not otherwise possible with the largely single-lot source sample set) and provides a more conservative and, notionally, realistic estimate of operational classification performance.

The relative probability that the unknown sample arose from one of the reference sources (or from a previously measured unknown source) is determined using a composite ML algorithm consisting of three parts: (1) elimination of the least likely sources using a *k*-Nearest Neighbors (*k*-NN) classifier for each of the three signature types; (2) a Multinomial logistic Regression (MNR) classifier to aid the examiner in assessing if a reference sample matching the unknown sample is present in the reference database or not; and (3) three random forest classifiers (one for each signature type) for final estimates of the relative match probability for the unknown sample to each of the remaining sources from the reference database. The source probabilities from the random forest classifier for each signature type are weighted by *z*-score across all of the source probabilities, then averaged and normalized to give "fusion scores" that are relative weighted probabilities. The individual signature and fused relative probabilities, as well as the raw signature data, are presented to the examiner in a user-focused web interface for further analysis and a source assignment decision.

To estimate the algorithm performance conservatively and as realistically as possible given inherent sample collection limitations, cross-validation testing was done by holding out signature data from entire lots (rather than a random subset of single samples) during training and using the entire held-out lots for subsequent testing. For the samples from 23 fertilizer sources that were tested, the correct source is chosen in cross-validation as the most likely source 87% of the time and is an average of 2.3 times as likely as the most probable incorrect source. These relatively accurate and confident sourcing results, even under conservative performance estimate conditions, indicate the promise of the signatures and composite ML sourcing algorithm for determining the source of unknown fertilizer samples.

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited. This material is based upon work supported under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force.

Machine Learning, Artificial Intelligence, Attribution