J17 The Application of t-Stochastic Node Embedding and Random Forest Statistical Methods to Classify Raman Spectra of Inkjet Printer Inks for Purposes of Identification and Production of Investigative Leads

Patrick Buzzini, PhD, Sam Houston State University, Chemistry and Forensic Science Bldg, 1003 Bowers Boulevard, Box 2525, Huntsville, TX 77341; James M. Curran, PhD, University of Auckland, Dept of Statistics, Private Bag 92019, Auckland 1142, NEW ZEALAND; and Carrie Polston, BA, Sam Houston State University, Chemistry and Forensic Science Bldg, 1003 Bowers Boulevard, Box 2525, Huntsville, TX 77341*

After attending this presentation, attendees will gain knowledge regarding the application of a relatively novel statistical approach to Raman patterns gathered *in situ* from micrometric colored spots of inkjet-printed documents from different sources.

This presentation will impact the forensic science community, with an emphasis toward questioned document examiners, by evaluating the utilization of a statistical approach to extract information from Raman patterns for investigative purposes. Although the main goal of this research is the improvement of the current examination of inkjet-printed counterfeit banknotes, the proposed approach is definitely relevant to areas of forensic science where spectroscopic data are used.

The printing process by means of inkjet technology involves the production of a constellation of ink spots of micrometric size. Raman spectroscopy proved to be a suitable method for rapidly obtaining a chemical signature *in situ* from the three main colored components (cyan, magenta, and yellow) of inkjet printer inks. The present step of this study seeks to evaluate whether the statistical methods of t-Stochastic Node Embedding (t-SNE) and random forest are suitable for classification to produce investigative leads in cases in which a suspected printer needs to be developed based on its detected Raman profiles.

One hundred fifty Raman spectra were captured from the cyan, magenta, and yellow spots of ten inkjet printer ink samples provided by the Treasury Obligations Section of the United States Secret Service, using a Near-Infrared (NIR) laser wavelength at 785nm. Given that inkjet ink samples generate Raman spectra that often can be differentiated on the basis of the presence of minor peaks only, a sensible classifier is then required for conducting spectral comparisons.

t-SNE is a dimension reduction technique that is primarily used for visualization of high-dimensional data. This technique seeks to preserve low-dimensional (potentially non-linear) groupings that may exist in high-dimensional data. Linear methods such as Principal Component Analysis (PCA) are unable to preserve these non-linear relationships. The method proceeds by converting Euclidean distances to conditional probabilities that represent similarities. These probabilities are computed in the original dimension and in the proposed low-dimensional representation. The optimal lower dimensional representation is the one that minimizes a well-known information theoretic criterion known as the Kullback-Leibler divergence. Van der Maaten and Hinton proposed a variation to t-SNE in which the Gaussian distribution used to compute the probabilities with a Student t-distribution is replaced and the joint probability distribution of pairs of points (rather than the conditional probabilities of one point given another) is used.[1] Both of these proposals lead to improved performance. The method of random forest is an ensemble approach to classification in which a group of low-dimension classifiers can be combined to provide a strong classifier (or learner). A random forest takes random subsets of the available variables to produce a tree, then aggregates (usually by averaging) the classifications over the trees to produce a classification.

The t-SNE visualization and PCA biplots of the Raman data reveal that in general, measurements from the same source (ink cartridge) cluster together. These methods also reveal observations that are behaving poorly. It is interesting to contrast the plots for all dye colors using t-SNE and PCA methods. As one might expect, the PCA plots emphasize the gross differences between the dye colors, whereas the t-SNE plots emphasize the clusters of samples within colors. Inspection of PCA biplots (using pairs of the first three principal components) reveal strong local structure aligned with the source of the measurement; that is, measurements from the same source appear to be close together in Euclidean space. This is an indication that classification techniques should be able to make a reasonable classification; in nearly each case, there are five measurements on each of ten sources. A 60:40 split for training and testing of a random forest model was used. The overall classification rate varied between 75% and 95%, providing at least initial evidence that this technique is a promising classifier for the samples of this study.

Reference(s):

1. Van der Maaten L.J.P, Hinton G.E. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research*. 2008; 9: 2579-2605.

Inkjet, Raman, Statistics

*Presenting Author