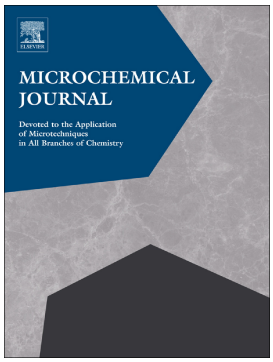
Accepted Manuscript



Validity of the best practice in splitting data for hold-out validation strategy as performed on the ink strokes in the context of forensic science

Loong Chuen Lee, Choong-Yeun Liong, Abdul Aziz Jemain

PII: S0026-265X(17)31300-0
DOI: doi:10.1016/j.microc.2018.02.009
Reference: MICROC 3048

To appear in: *Microchemical Journal*

Received date: 20 December 2017
Revised date: 8 February 2018
Accepted date: 8 February 2018

Please cite this article as: Loong Chuen Lee, Choong-Yeun Liong, Abdul Aziz Jemain , Validity of the best practice in splitting data for hold-out validation strategy as performed on the ink strokes in the context of forensic science. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Microc(2017), doi:10.1016/j.microc.2018.02.009

Validity of the best practice in Splitting Data for Hold-out Validation strategy as performed on the ink strokes in the context of forensic science

Loong Chuen Lee[1,2], Choong-Yeun Liong*[2], Abdul Aziz Jemain[2]

[1]*Forensic Science Program, FSK, Universiti Kebangsaan Malaysia, Jalan Raja Muda Abdul Aziz, 50300 Kuala Lumpur, Malaysia*
[2]*School of Mathematical Sciences, FST, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia*

Abstract

External testing (ET), known also as the hold-out validation, is currently considered to be one of the most reliable ways to estimate predictive ability of a statistical model. One safeguard to prevent impermissible peeking in ET is to ensure all replicates of a particular sample is only included in either the test or the training set. Assuming a sample X1 consists of two replicates (i.e. X1a and X1b). The model is claimed to enjoy impermissible peeking if the X1a and X1b are split into the training and the test sets, respectively. Eventually, the resulting prediction model is expected to predict the test sets easily and presents an over-optimistic model performance. In forensic document examinations, an individual pen (IP) can be used to produce multiple ink strokes. In real-world practice, pens are manufactured via bulk production such that one big tank of ink is used to produce a wealth of IPs. In other words, ink strokes produced by varying IPs but of the same pen model are indeed originated from one single source (i.e. the same tank of ink). Eventually, with respect to the aforementioned safeguard, how shall one treat the ink strokes? Are they replicates or independent samples? In this context, the aim of the work is to investigate the validity of the safeguard in splitting dataset for hold-out validation strategy (i.e. ET) in the domain of forensic pen ink analysis. An infrared (IR) spectra of blue gel pen inks was used to demonstrate the practical aspect. The IR spectral data were collected from 1361 ink strokes that originated from 273 IPs of 23 pen models and 10 pen brands. Iterative stratified random sampling was employed to prepare 1000 pairs of training and test sets that were split at ratio 7:3 using two different principles: (a) set IP - selection was conducted at IP level to ensure all the ink strokes originated from a particular IP must be included into either the training or the test sets only; and (b) set NIP - ink strokes of a particular IP were allowed to be spread between the training and the test sets. For each dataset, a series of 50 PLS-DA models were constructed by including the first 50 PLS components incrementally, which were then validated via auto-prediction and ET. Following that, the performances between IP and NIP model series were compared with respect to: (a) model accuracy; (b) model stability; and (c) model fitting. In conclusion, the NIP model series do not show any evidence of advantages from the impermissible peeking since both the NIP and IP model series exhibit quite similar performances in all the three model aspects.

Keywords: PLS-DA, replicates, data splitting, model validation, IR spectrum, forensic science

*Corresponding author

E-mail address: lc_lee@ukm.edu.my (LC Lee); lg@ukm.edu.my (C-Y Liong)*; azizj@ukm.edu.my (AA Jemain)