Multivariate Calibration Domain Adaptation with Unlabeled Data

Robert Spiers, John H. Kalivas
Department of Chemistry
Idaho State University
921 S. 8th Ave., STOP 8023
Pocatello, ID 83209, USA
spierob2@isu.edu, kalijohn@isu.edu

Idaho State University

Abstract

Multivariate calibration is about modeling the relationship between a substance's chemical profile and its spectrum (here, near-infrared) in order to predict the concentration of new samples with known spectra. However, these new samples are often measured under different conditions than the *primary* conditions; different instruments, instrument drift, and temperature all affect the measurement conditions. Domain adaptation (DA) methods force the model to ignore these differences in order to generate an accurate model for the new domain (*secondary* conditions). There are two fundamental DA processes that individual methods can be classified under. One augments a few samples from the secondary domain with chemical reference values (labels) to the primary data and the other augments only secondary spectra (unlabeled data). In this work, we compare two existing labeled DA methods and two existing unlabeled DA methods to two novel labeled methods and a novel unlabeled approach. Since DA methods require selection of hyperparameters, a model selection framework based on model diversity and prediction similarity (MDPS) is applied to the DA methods. Regardless of the DA method, the MDPS process is shown to select models more accurate than the first quartile of all models generated by the DA process in three near-infrared datasets.

Objective

- Develop domain adaptation protocol for use with multivariate calibration data (near-IR spectroscopy, corresponding concentration profiles)
- Compare against traditional methods of domain adaptation
- Apply the novel model diversity and prediction similarity (MDPS) framework to select models from the domain adaptation methods

Domain Adaptation Methods

Labeled Secondary

- Mean Centering:
  - Local (LMC) and Global (GMC)

$$\begin{pmatrix}\mathbf{y}_P\\ \lambda\mathbf{y}_S\end{pmatrix}=\begin{pmatrix}\mathbf{X}_P\\ \lambda\mathbf{X}_S\end{pmatrix}\mathbf{b}$$

- Note: Local indicates that $\mathbf{X}_P$ and $\mathbf{X}_S$ are centered locally, rather than to the global combination

Unlabeled Secondary

Null Augmentation Regression (NAR)

$$\begin{pmatrix}\mathbf{y}_P\\ \mathbf{0}\end{pmatrix}=\begin{pmatrix}\mathbf{X}_P\\ \lambda\mathbf{R}\end{pmatrix}\mathbf{b}$$

- NAR-Centroid (NAR-C)

$$\mathbf{R}_C=\left(\boldsymbol{\mu}_P-\boldsymbol{\mu}_{SU}\right)^{\mathrm{T}}$$

- NAR-Covariance:
  - Local (LNAR-Cov) and Raw (RNAR-C)

$$\mathbf{R}_{Cov}=\frac{1}{m_P}\mathbf{X}_P^{\mathrm{T}}\mathbf{X}_P-\frac{1}{m_{SU}}\mathbf{X}_{SU}^{\mathrm{T}}\mathbf{X}_{SU}$$

Hybrid Labeled and Unlabeled Secondary

- NAR-Hybrid:
  - Local (LNAR-H) and Global (GNAR-H)

$$\begin{pmatrix}\mathbf{y}_P\\ 0\\ \tau\mathbf{y}_S\end{pmatrix}=\begin{pmatrix}\mathbf{X}_P\\ \lambda\mathbf{R}_C\\ \tau\mathbf{X}_S\end{pmatrix}\mathbf{b}$$

*All equations solved by Partial Least Squares (PLS)

Result Validation:

Accuracy of selected models is verified using a subset of secondary whose analyte values did not go into forming the model

$$RMSEV=\sqrt{\frac{\sum_{n=1}^{m}\left(\hat{y}_n-y_n\right)^2}{m}}$$

Real Life Applicability

- **Labeled secondary methods**
  - Effective when 5-10 samples are measured under the new (secondary) conditions, known analyte
  - Original (primary) and new (secondary) conditions can be quite different
- **Hybrid labeled/unlabeled secondary methods**
  - Useful with few (1-5) samples measured under new conditions
  - Primary and secondary conditions should be fairly similar
- **Unlabeled secondary methods**
  - Is performed when no samples are available in secondary conditions
  - Primary and secondary conditions must be quite similar

Model Selection by MDPS

General Theory

Get every possible combination of two models, want diverse combinations of models that retain similar predictions

Model Diversity

Cosine of the angle between the i^{th} and j^{th} models

$$\cos\left(\theta\right)_{i,j}=\frac{\left(\hat{\mathbf{b}}_i\right)^{\mathrm{T}}\left(\hat{\mathbf{b}}_j\right)}{\left\|\hat{\mathbf{b}}_i\right\|\left\|\hat{\mathbf{b}}_j\right\|}$$

Prediction Similarity

Secondary Prediction Difference (SPD):

Analyte prediction differences of the i^{th} and j^{th} models relative to the m secondary spectra

$$SPD_{i,j}=\sum_{n=1}^{m}\left|\hat{y}_{n,i}-\hat{y}_{n,j}\right|$$

RMSEC$_P$:

Prediction error of the model as it relates to the primary set of samples

$$\overline{RMSEC}_{P(i,j)}=\frac{RMSEC_{P(i)}+RMSEC_{P(j)}}{2}$$

Range-Scaled Weighted Fusion (ω)

Weight regression vector 2-norm to characterize overfitting
Incorporation of RMSEC$_P$ to account for underfitting

$$C_{i,j}=\frac{SPD_{i,j}-\max\left(SPD\right)}{\max\left(SPD\right)-\min\left(SPD\right)}+\omega\left(\frac{\left\|\hat{\mathbf{b}}\right\|_{i,j}-\min\left(\left\|\hat{\mathbf{b}}\right\|\right)}{\max\left(\left\|\hat{\mathbf{b}}\right\|\right)-\min\left(\left\|\hat{\mathbf{b}}\right\|\right)}+\frac{\overline{RMSEC}_{P(i,j)}-\min\left(\overline{RMSEC}_P\right)}{\max\left(\overline{RMSEC}_P\right)-\min\left(\overline{RMSEC}_P\right)}\right)$$

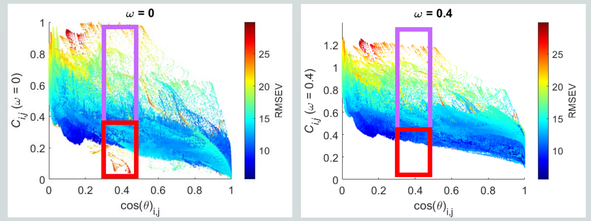


Figure 1. MDPS figures showing the organization of the prediction similarity against the model diversity for a combination of two models. Combinations are taken within the purple bucket and sorted to find the lowest 10% according to the red bucket

Data Description

Corn

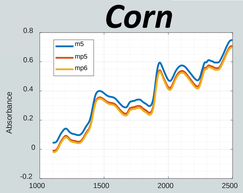


Figure 2. Corn: spectra of 80 corn samples measured on three instruments: m5, mp5, mp6. Analytes include moisture, oil, protein, and starch

Tablet

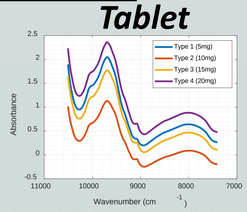


Figure 3. Tablet: spectra of 240 pharmaceutical tablets with analyte API. Samples grouped according to API. Primary is lab batch, secondary is full

Goat

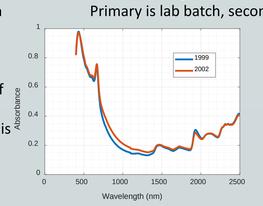


Figure 4. Goat: spectra of goat feces analyzed for juniper content. Primary is 1999, secondary is 2002

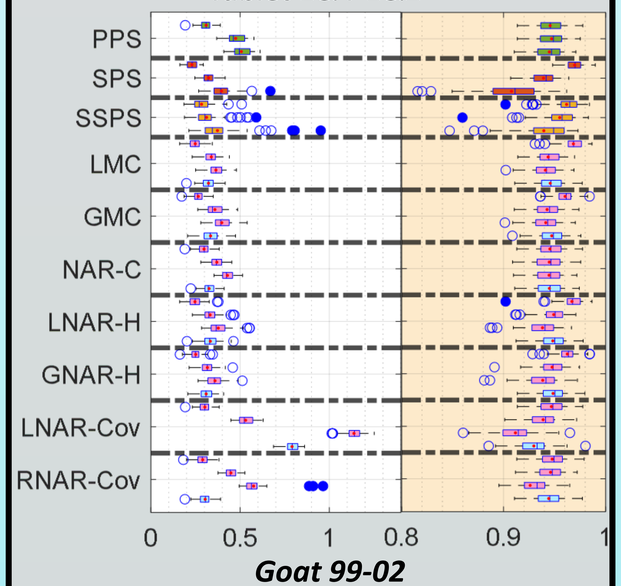
Division of samples for updating

	Primary	Secondary	Validation
Corn	40	5	20
Tablet	60	6	24
Goat	61	5	20

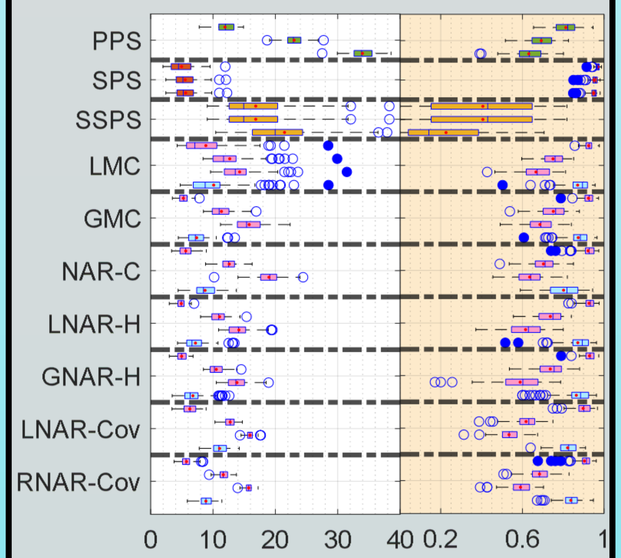
Table 1. Numeric division of samples into primary and secondary to preserve result consistency

Results

Tablet 1&4-1&2



Goat 99-02



Figures 5 and 6. Boxplots of (left) RMSEV and (right) R^2 of models generated by the domain adaption methods and selected by MDPS. PPS, SPS, and SSPS correspond to the baseline model generation methods, where we expect to perform better than PPS and SSPS, and no better than SPS. The first three boxes in every block correspond to the minimum, first quartile, and median of all models generated, respectively. Blue boxes correspond to models selected by MDPS.

Takeaway:

- Models selected by MDPS (blue) perform at or better than the first quartile of all generated models
- Model updating performs almost as good as the state-of-the-art incredibly expensive method of SPS
- Great predictions are achieved using unlabeled secondary data (NAR-C, NAR-Cov1, NAR-Cov2)

Results

Corn m5-mp5 Moisture

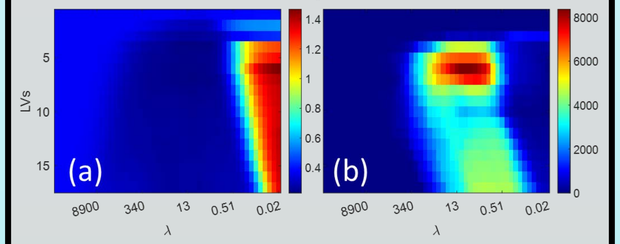


Figure 7. Heatmap of (a) RMSEV and (b) frequency of a given model being selected by MDPS for Corn m5-mp5 moisture, in the NAR-Cov1 updating situation.

Takeaway:

- Model selection by MDPS selects the most accurate models, as evidenced by the darkest blue (lowest RMSEV) models on the left side being most frequently selected on the right side

Conclusions

- Domain adaptation using very few or no reference values for the secondary domain achieves great accuracy
  - In datasets where primary and secondary are similar, the unlabeled secondary methods can outperform labeled
- Model selection using MDPS achieves performances at or better than the first quartile of all generated models in every domain adaptation situation
- When primary and secondary are sufficiently similar, complete model recalibration should never be necessary

Potential Applications

- Rapid analysis of tablet dosage even as the production method shifts slightly
- Real-time analysis of agricultural nutrient as the near-IR instrument degrades over time
- Standardization of handheld spectrometers (e.g. in a smartphone) even as instruments become damaged or lenses get smudged

Acknowledgements

This material is based upon work partially supported by the National Science Foundation under Grant Nos. CHE-1506417 (co-funded by CDS&E) and CHE-1904166 (co-funded by CDS&E and the Office of Investigative and Forensic Sciences in the National Institute of Justice).