Analyst

Supervised Discretization for Decluttering Classification Models

Journal:	*Analyst*
Manuscript ID	AN-ART-05-2023-000770.R2
Article Type:	Paper
Date Submitted by the Author:	15-Sep-2023

Complete List of Authors:	Jordan, James; USGS Celani, Caelin; University of Delaware, Department of Chemistry & Biochemistry Ketterrer, Michael; Northern Arizona University Lavine, Barry; Oklahoma State University, Department of Chemistry Booksh, Karl; University of Delaware, Department of Chemistry and Biochemistry

Supervised Discretization for Decluttering Classification Models

James A. Jordan[a], Caelin P. Celani[b], Michael Ketterer[c], Barry K. Lavine[d], and K. S. Booksh[b*].

a. United States Geological Survey, Reston, VA.
b. University of Delaware, Department of Chemistry and Biochemistry, Newark, DE
c. Northern Arizona University, Department of Chemistry and Biochemistry, Flagstaff, AZ
d. Oklahoma State University, Department of Chemistry, Stillwater, OK

Abstract

Presented here is the first demonstration of supervised discretization to 'declutter' multivariate classification data in chemical sensor applications. The performance of multivariate classification models is often limited by the non-informative chemical variance within each target class; decluttering methods seek to reduce within-class variance while retaining between-class variance. Supervised discretization is shown to declutter classes in a manner that is superior to the state-of-the-art External Parameter Orthogonalization (EPO) by constructing a more parsimonious model with fewer parameters to optimize and is, consequently, less susceptible to overfitting and information loss. The comparison of supervised discretization and EPO is performed on three classification applications: X-ray fluorescence spectra of pine ash where the pine was grown in three distinct soil types, laser induced breakdown spectroscopy of colored artisanal glasses, and laser induced breakdown spectroscopy of exotic hardwood species.

1. Introduction

Discretization is a set of techniques for conversion of continuous (or near continuous) variables into discrete variables while minimizing information loss.[1], [2] In machine learning, discretization offers several advantages. Methods such as Classification and Regression Trees (CART) and Naïve Bayes (NB) classifiers only work with discrete value data.[2, 3, 4] Discretization reduces the dimensionality of data and increases the speed of learning.[5] Additionally, decision trees constructed from discretized data tend to be more compact and accurate than rule structures developed from continuous data.[2, 5, 6]

Discretization methods can be characterized based on their algorithmic implementation; discretization methods are either supervised or unsupervised and are either univariate or multivariate. An unsupervised discretization method may set all observations within a fixed range to have a single nominal value whereas a supervised discretization would use class information to more optimally adjust the ranges prior to setting all observations within each range to a single nominal value. Univariate discretization methods operate on one variable at a time and do not consider information content from the other variables, whereas multivariate discretization methods consider relations among multiple variables during discretization. Additionally, parametric discretization methods require input from the user such as setting the number or frequency of bins, whereas nonparametric methods only use information from the data. A more complete taxonomy further contrasts discretization algorithms based on the relationships among the observed variables or between the variables and the classification model during the discretization process.[1]

The goal of a discretization algorithm is to optimize the number of discrete intervals, defined by their boundaries, across the range of each continuous variable. Many supervised discretization algorithms exist that take into account the relationship between class identity and the measurement variables, the so-called class-attribute interdependence. These algorithms optimize the discretization intervals based on information theory[5, 7, 8], statistics[9, 10], or empirical heuristics about the class attribute-interdependence.[11, 12] Class-attribute interdependence maximization (CAIM) is a supervised, univariate, nonparametric discretization algorithm.[13, 14] CAIM is heuristic-based and strives to simultaneously maximize the interdependency between the class labels and the continuous value attribute (variable) while minimizing the number of discrete intervals.

A powerful strategy that improves classification models is using multivariate filtering methods to identify and remove unwanted covariance structures that limit model performance. Popular strategies for multivariate filtering include various algorithms for Orthogonal Signal Correction (OSC),[15, 16, 17] Tikhonov regularization[18], and External Parameter Orthogonalization (EPO)[19, 20]. OSC seeks to successively identify and remove the largest direction of variance in the variable space that is orthogonal to the property of interest in the sample space. OSC was designed for calibration applications but can be appropriated for classification models. By contrast, EPO was designed for classification applications. EPO envisions the ideal class to be a collection of identical points in the variable space; any deviation from the class mean is viewed as 'clutter' to be removed. Consequently, EPO performs Principal Component Analyses (PCA) on the within class variance to determine which variance to remove from the variable space. For both EPO and OSC, the number of components removed is determined by the analyst. If all possible components worth of variance are removed to declutter the classes, EPO reduces to an Extended Mixture Model filter.[21] Tikhonov regularization augments the data collection with a small number of 'clutter' spectra that span the interferent space to create a model that is desensitized to those sources of variance.

This work presents the first demonstration of supervised discretization to 'declutter' multivariate classification data in chemical sensor applications. Although supervised discretization has been successfully used to denoise data prior to analyses, the ability of discretization to mitigate the deleterious effects of uncontrolled chemical and instrumental effects (e.g., moisture, temperature, or sample matrix composition) has not been explored to date. Through three examples of classification by spectra collected on real-world samples with field portable instrumentation, supervised discretization 'declutters' classes are shown to be superior to the state-of-the-art EPO. Supervised discretization presents a more parsimonious model with fewer parameters to optimize and is, consequently, less susceptible to overfitting and information loss. The comparison of supervised discretization and EPO is performed of three classification applications: X-ray fluorescence spectra of pine ash where the pine was grown in three distinct soil types, laser induced breakdown spectroscopy of colored artisanal glasses, and laser induced breakdown spectroscopy of exotic hardwood species.

2. Class attribute interdependence maximization (CAIM)

Input data for the CAIM algorithm takes the form of a data matrix and the class vector. When discretization is initiated, the algorithm iteratively cycles through each variable, $\mathrm{F_i}$, of the data matrix, sorting it in descending order, then calculates the minimum value (d_0), the maximum value (d_n), and

midpoints (B) between each pair of successive observed values. This parsing provides the data needed to establish a single interval discretization scheme (D) for the variable that covers the full range of the data (i.e., $D_{initial}$, which is the mathematical set spanned by $\{[d_0,d_n]\}$). Successive iterations of D will parse the span of the variable into contiguous segments (e.g., D_3 refers to the set of intervals $\{[d_0, d_1],\ (d_1,\ d_2],\ (d_2,\ d_3]\}$). After these parameters are established, the continuous variable F_i is transformed into a quanta matrix (Table 1). Features of a quanta matrix include rows that correspond to the *S* unique classes in the input dataset (C), columns that correspond to the *n* intervals in the discretization scheme, the final row that is the sum of each column or the number of objects that belong to each discretized interval, the final column that is the sum of each row or the number of objects that belong to each class, and the bottom right corner of the quanta matrix that is the total number of objects or samples in the original dataset.

Table 1: The quanta matrix is the basis for visualizing multiple different supervised discretization schemes including CAIM.

Class	**Interval**					**Class Total**
	$[d_0, d_1]$	**…**	**$(d_{r-1}, d_r]$**	**…**	**$(d_{n-1}, d_n]$**	
C_1	q_{11}	…	q_{1r}		q_{1n}	M_{1+}
…	…	…	…	…	…	…
C_i	q_{i1}	…	q_{ir}		q_{in}	M_{i+}
…	…	…	…	…	…	…
C_s	q_{s1}	…	q_{sr}		q_{Sn}	M_{S+}
Interval Total	M_{+1}	…	M_{+r}	…	M_{+n}	M

A score is calculated from the quanta matrix to assess the value of the discretization. The quanta matrix can be used to determine many different scoring metrics: Shannon's entropy, Class-Attribute Information (INFO)[22], Class-Attribute Mutual Information[14], Class-Attribute Interdependence Redundancy (CAIR)[23, 24], and Class-Attribute Interdependence Uncertainty (CAIU)[14]. Many of these metrics are related; for example, CAIR is the ratio of Class-Attribute Mutual Information and Shannon's entropy whereas CAIU is the ratio of INFO and Shannon's entropy.

In this study, the CAIM heuristic[13, 14],

$$CAIM(C,D \mid F) = \frac{\sum_{r=1}^{n} \frac{(q_{ir})_{max}^{2}}{M_{+r}}}{n} \qquad \text{(Eq. 1)}$$

is used. In Eq. 1, the algorithm iterates through all *n* intervals in the quantum matrix where *r* = 1, 2, … *n*. The term $(q_{ir})_{max}^{2}$ is the square of the maximum q_{ir} value in the r^{th} interval. The term M_{+r} is the total number of samples that fall within the interval *r* for a particular variable. In this manner, CAIM seeks to make each interval as pure as possible with respect to the assigned classes. The squaring of $(q_{ir})_{max}$ serves to reward the algorithm for having fewer well-populated intervals over smaller sparsely

populated intervals. The CAIM score is initially determined for $D_{initial}$. For each successive iteration, provisional CAIM scores are determined from iteratively creating provisional boundaries between sequential continuous observations for a particular variable. The largest CAIM score from all the tested provisional boundaries is compared to the CAIM score from the previous iteration. If the new CAIM score is greater than the old CAIM score, the iterative process continues with D updated, unless the set maximum number of boundaries has been reached. Generally, the maximum number of inner boundaries is set to be the number of classes. If the new CAIM score is not greater than the old CAIM score, the algorithm terminates for this variable and the previous D is retained. This process is then repeated for each variable of the original dataset until all variables are discretized with their respective discretization schemes.

The CAIM heuristic (Equation 1) normalizes the maximum value in a given interval (column) relative to the total instances that occur in that interval, then sums this metric across all intervals of a given quanta matrix before normalizing by the number of intervals in the given discretization scheme. In this way, the CAIM heuristic prioritizes information that distinguishes classes across intervals, or in other words, the CAIM heuristic is increased when one interval, or a subset of the intervals, exclusively describe information from a single class. Furthermore, the CAIM heuristic then prioritizes discretization schemes with fewer intervals by penalizing the CAIM heuristic through normalization of the class attribute interdependency metric (numerator) by the total number of intervals in a given scheme, that is, as the number of intervals increases, the CAIM heuristic decreases.

CAIM discretization, as opposed to other supervised discretization strategies, was chosen because CAIM tended to provide better performance metrics (e.g., accuracy with minimum number of rules) than other methods for parsing individual variables based on provided classes. Of note, the intended use of discretization here is conceptually different than previously published applications where CAIM was compared to other methods. First, here discretization is used as a preprocessing step for a partial least squares – discriminant analyses (PLS-DA) model, not as the final step in analyses. Second, discretization in this application is generally limited to two-way classifications in service to a PLS-DA decision tree or binary classification models. In traditional discretization literature, test applications would have three or more target classes. Consequently, the open question of which discretization criteria performs best with PLS-DA, support vector machine – discriminant analyses (SVM-DA), or other algorithms is not addressed in this study.

3. Experimental

3.1. Data analyses

All data were imported and analyzed in the Matlab (Matlab, Natick, Massachussetts, USA) operating environment. The CAIM discretization algorithm was written by Booksh at the University of Delaware based on the papers by Kurgan and Cios.[13, 14] Classification by PLS-DA and SVM-DA were performed in the PLS Toolbox (Eigenvector Inc., Chelan, Washington, USA). Additionally, the automatic asymmetric least-squares (Whittaker) filter[25], Savitzky-Golay[26], and external parameter orthogonalization (EPO)[19,20] methods were all used within the PLS toolbox.

3.2. X-Ray Fluorescence (XRF) spectra of pine ash

Live pine needles were collected from *Pinus ponderosa* in public open spaces across a narrow geographic region of the Colorado Plateau near Flagstaff, Arizona, USA. Trees were identified as growing in soil atop one of three well-defined parent rock archetypes: recent basalt/andesite volcanic rock, Kaibab Limestone, and Moenkopi Formation. The needles were cut into ~1 cm lengths and dry ashed over a hydrogen flame at 600 °C.

The pine needle ash was consequently pressed into a pellet for better handling while collecting XRF spectra. A base layer of 2.6 grams of confectioners' sugar was added to the cavity of a 13-mm stainless steel die press assembly on top of the polished face of a pressing disc and compacted. Approximately 0.5 g of ash was placed on top of the binder followed by the second pressing disc. The die assembly was placed into the VivTEK® (COL-INT TECH, USA) 12-ton, 2-pole hydraulic press and subjected to 10 tons of force for 30 sec. Additionally, sucrose blanks were also prepared to account for any potential background caused by the inclusion of trace contamination. However, analysis of the blanks indicated that sucrose binding does not impart any additive noise to the background and thus it was deemed unnecessary to continue accumulating data to provide for the subtraction of the backing matrix. XRF measurements were performed using an Olympus X-ray fluorescence analyzer Vanta C series running in the three beam GeoChem mode (50 kV). Acquisition method timings were adjusted to perform measurements using each of the 10-kV, 40-kV, and 50-kV beams for 60 sec each and processing by fundamental parameters. The instrument's calibration was routinely checked using manufacturer supplied calibration materials to ensure that the instrument's calibration was within the stated values included with the NIST certificate of calibration. Example raw spectra are presented in **Supplemental Information SI1**.

Data were pretreated by application of a Savitzky-Golay algorithm (7 point smooth, quadratic, first derivative) to minimize the effect of the XRF baseline on the subsequent analyses. A rough variable selection was performed by visually selecting ranges of energies spanning each of the 17 XRF peaks from the ensemble spectra; only the regions with an XRF signal were used. This reduced the length of each XRF spectrum from 2049 unique channels to 441 channels. Each reduced spectrum was normalized to unit area to account for variability in ash loading and sample placement across all collected spectra.

All replicate spectra from approximately $1/4^{th}$ of the samples from each class were randomly selected and removed to form a validation set. The 129 spectra in the training set were composed of triplicate XRF spectra from ashes of 8 trees grown in soil derived from the Kaibab Limestone (hereafter Kaibab samples), 16 trees grown in soil derived from the Moenkopi Formation (hereafter Moenkopi samples), and 19 trees grown in basalt/andesite soil. The 39 validation set spectra were collected in triplicate from ashes of 2 trees grown in Kaibab soil, 5 trees grown in Moenkopi soils, and 6 trees grown in basalt/andesite soil.

3.3. Laser Induced Breakdown Spectroscopy (LIBS) Spectra of colored glasses

Soft glass (coefficient of expansion 104) samples were purchased from Devardi Glass (Sheridan, Oregon, USA) of the type appropriate for lampworking projects. The glasses were a "set of mixed reds" of various hues, both opaque and transparent, each approximately 25 cm long and 6 – 10 mm in diameter. Inspection of the 21 rods received and comparison to the Devardi catalog color chart indicates that the set contains 2 duplicate- and 1 triplicate-colored rods. Although other rods might be duplicate colors, they were not readily identified by visual inspection. Prior to analyses, the rods were designated 'A'

through 'U.' It was determined by visual inspection and knowledge of sample provenance prior to analyses that 'I'/'G' and 'K'/'S' were duplicate pairs and 'E'/'J'/'R' was a triplicate pair. Consequently, the set of 21 rods spans as many as 17 unique colors.

LIBS spectra were collected with a SciAps Z300 hand-held LIBS analyzer. Samples were aligned by manually holding each glass rod in the v-shaped alignment groove on the Z300 faceplate. Twelve spectra were collected from random locations on each glass rod. Three spectra of each rod were collected in one sitting. Nine spectra on rods 'A' through 'R' were subsequently collected at a later date. The remaining 9 spectra of rods 'S' through 'U' were collected in a single sitting at a different date. Consequently, each set of 12 spectra spans at least 2 collection periods. In total, 252 spectra were obtained for this data set. Example raw spectra are presented in **Supplemental Information SI2**.

The LIBS baseline contribution was minimized by a Savitzky-Golay algorithm (7-point window, quadratic fit, and first derivative). The derivatized signal at every wavelength was transformed by applying the square root of its absolute value; this normalized the error distribution across peaks of vastly different scale. Each spectrum was then normalized to unit area to account for efficiencies in placing the sample on the LIBS analyzer. Based on the mean spectra of the entire data collection, a threshold value of 0.004 units was determined to separate 'baseline' from 'LIBS' channels. In this manner the number of channels employed in each spectrum was reduced from 23,431 to 8,169.

3.4. LIBS spectra of Dalbergia

Collection and preprocessing of the *Dalbergia* spectra have been previously discussed.[27] LIBS spectra from 90 *Dalbergia* samples were collected and provided by the U.S. Forest Service. Samples consisted of seven classes of *Dalbergia* hardwoods and two classes of non-*Dalbergia* hardwoods. For each of the nine classes, one LIBS spectrum from approximately 10 distinct exemplars were recorded with a SciAps Z-200C LIBS analyzer.

The spectral baseline was removed by an asymmetric least squares (Whitaker) filter ($\lambda = 100$; $P = 0.001$) followed by a first derivative Savitzky-Golay (2nd order, 15 points) smoothing to help remove any residual baseline and better eliminate high-frequency noise. Each variable was normalized by taking the square root of the absolute signal intensity following baseline removal. Variables were down selected from 17,431 to 489 by removing all variables with an average intensity less than 0.5 units. All remaining variables were then autoscaled ($\mu = 0$, $\sigma = 1$) prior to analyses. The 90 spectra were organized into three different training and validation set combinations of 72 training samples and 18 validation samples by bootstrapped Latin partitions.[28] Each combination consisted of two samples from each class in the validation set.

4. Results and Discussion

4.1. Classification of pine ash

The challenge for determining original soil type from tree ash lies in the large variability of the XRF signal within each soil class. Comparing the mean ash spectrum from the 30 'Kaibab', 63 'Moenkopi Formation', and 75 'basalt/andesite' samples indicates a unique XRF signature from each soil of origin

(**Figure 1a**). Overlaying the 95% confidence interval for a sample from each class as determined by the standard deviation of all spectra in a class demonstrates how the natural spread of the data within a class overlaps the mean spectra of other classes (**Figure 1b-d**). Concurrently, calculation of variance between the means of the three classes, the mean variance within each of the three classes, and the mean variance of the triplicate spectra from each ash pellet shows that the largest source of variance is attributed to the natural spread of the spectra within a class (**Table 2, column 2**).

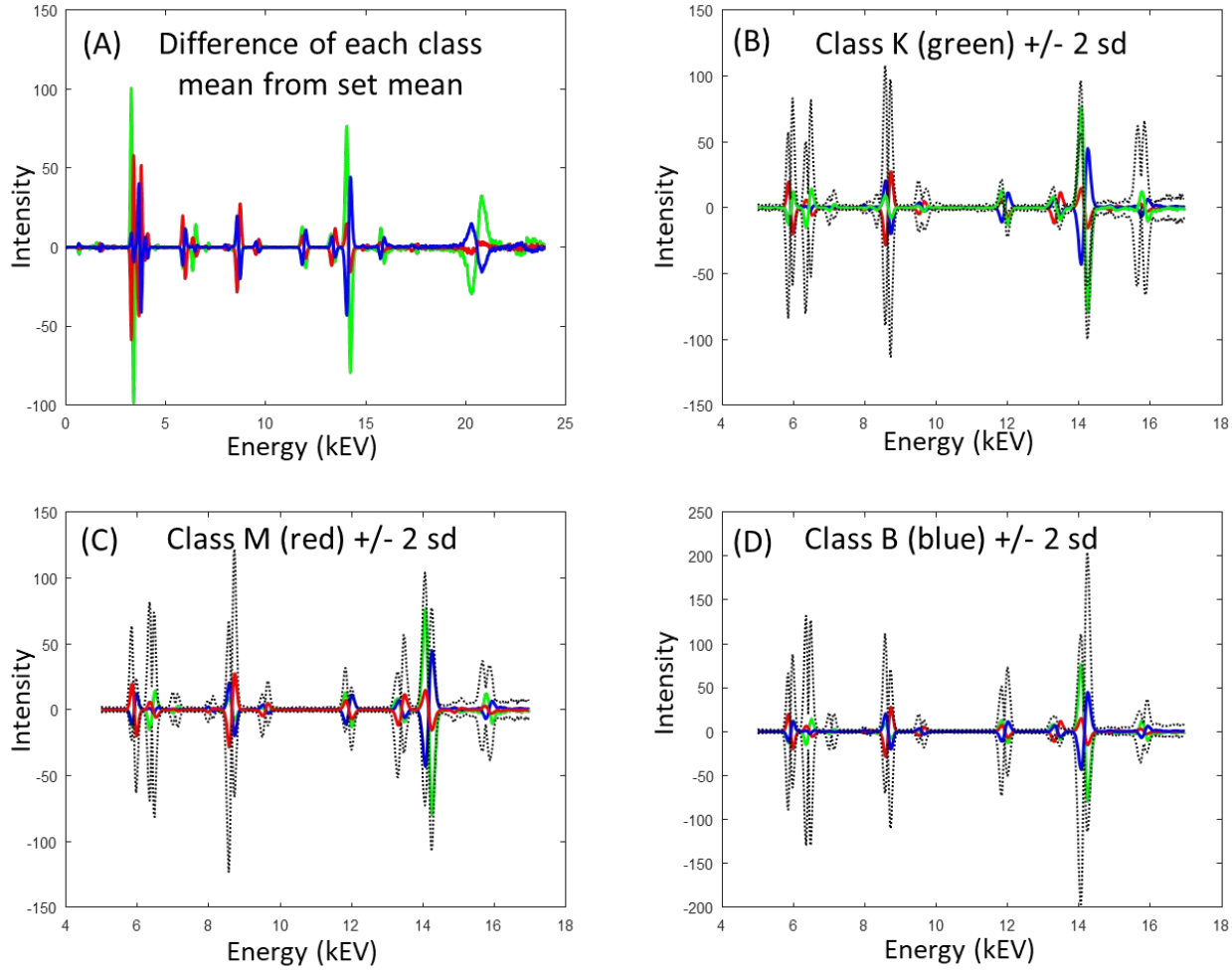

Figure 1. Comparison of the average mean-centered XRF spectra for ash from pine trees grown in soil derived from Kaibab Limestone (green), Moenkopi Formation (red), and basalt/andesite (blue) shows that the spectral profiles for these three classes are highly overlapped, yet each class has a distinct spectral signature (A). Including the +/- 2 standard deviation (sd) error bars at each kEV demonstrated that the within class variability is greater than the between class variability; in each case the confidence limits extend beyond the average spectrum for the other classes (B – D).

Table 2: Effect of External Parameter Orthogonalization (EPO) decluttering on distribution of variance across the pine ash data set and performance of Principal Component Analyses (PCA) and K-Nearest Neighbors (K-NN) modeling.

Treatment	Un-decluttered	EPO (1 Factor)	EPO (6 Factors)	EPO (Full Rank)
Variance of class means	2.48×10^{-4} (32.7%)	2.11×10^{-4} (48.8%)	9.66×10^{-6} (35.0%)	1.33×10^{-6} (58.4%)
Mean variance within a class	5.08×10^{-4} (67.3%)	2.21×10^{-4} (51.1%)	1.78×10^{-5} (64.5%)	4.33×10^{-7} (19.0%)
Mean variance of replicates	1.02×10^{-6} (0.013%)	6.56×10^{-7} (0.15%)	1.37×10^{-7} (0.50%)	5.15×10^{-7} (22.6%)
PCA	# PC: 8 Cum Var: 95.2%	# PC: 7 Cum Var: 90.9%	# PC: 3 Cum Var: 44.4%	# PC: 3 Cum Var: 15.0%
K-NN(1) Misclassified (129:39 split)	0 Cal; 11 Pred	0 Cal; 13 Pred	13 Cal; 10 Pred	62 Cal; 12 Pred

EPO was conceived to reduce the 'clutter' within each class, shifting the distribution of variance from 'within classes' to 'between classes' and hence improving both the precision and accuracy of classification models. For the XRF ash data, EPO generally accomplishes the desired effect of minimizing the 'within classes' variance relative to the 'between classes' variance. Increasing the number of factors in EPO pretreatment increases the percent of variance 'between classes' from 32.7% with no EPO, to 48.8% with a 1-factor EPO decluttering, to 58.4% with a full EPO decluttering (**Table 2, row 1**). However, the decrease to 35.0% when using a 6-factor EPO treatment presages a limitation of EPO-based decluttering. EPO removes all variance that is colinear with the data clouds of each class. The removal of variance is apparent in the decrease of variance in all three categories (**Table 2, rows 1 – 3**). However, a successful EPO application assumes that the sub-space of 'between class' variance is largely orthogonal to the sub-space of the removed 'within classes' variance. When this assumption fails to hold, a significant portion of the discrimination 'between class' variance is removed and the proportion of 'between class' variance may decrease. Note in this example, after a large portion of the 'between class' variance is removed, the principal component (PC) space of the data decreases from 7 or 8 PCs to only 3 PCs (**Table 2, row 4**) and K-Nearest Neighbors (k-NN) can no longer reliably classify samples in the training set. The number of training set misclassifications increases from 0, to 13 with a 6-factor EPO, to 62 with a full-factor EPO decluttering (**Table 2, row 5**).

Using PLS-DA for classification of pine ash by original soil type highlights the potential benefit of decluttering with EPO and the concern for overfitting when decluttering with EPO. In this application, PLS-DA models for three one class versus all other classes were observed to perform better than a single flat PLS2-DA classifier with either no decluttering or with EPO decluttering. The number of factors in each model was based on the cross-validated class error for the training set by removing 10% of the spectra at each split. Comparing the validation set predictions across multiple levels of decluttering (**Table 3**) shows that either no decluttering by EPO or decluttering by a 1-factor EPO model yields the best results. No decluttering was needed to correctly classify all 6 Kaibab samples, each with greater than 90% probability. However, when a 1-factor EPO treatment was applied, 3 Moenkopi samples were incorrectly identified as Kaibab with 80% to 90% probability. Further increasing the number of EPO factors results in 3 Kaibab samples not identified as being classified in the Kaibab set. No decluttering

and 1-factor EPO models perform comparably for the Moenkopi samples; the models only slightly differ in their probabilities of classification. Similarly, the un-decluttered, 1-factor EPO and 6-factor EPO all correctly classify the basalt/andesite samples with EPO yielding classification results at higher probabilities.

Table 3: PLS-DA Classification of XRF ash data with no decluttering (Base), External Parameter Orthogonalization (EPO) decluttering, and CAIM discretization decluttering for 6 validation samples derived from the Kaibab Limestone, 15 validation samples derived from the Moenkopi Formation, and 18 basalt/andesite validation samples. Every model was constructed as a target class versus all other classes as a single group except for the final column which is a flat CAIM discretization decluttered classifier.

Sample	Model	Confidence	Base	EPO(1)	EPO(6)	EPO (All)	CAIM (1 group versus 2 groups)	CAIM (3 Groups)
'Kaibab'	K vs. (M-K & B/A)	Correct	6	6	3	2	6	6
		Incorrect	0	0	3	4	0	0
		False Positive	0	3	3	0	0	1
		Not Classified	0	0	0	0	0	0
'Moen-Kopi'	M-K vs. (K & B/A)	Correct	12	12	9	7	12	9
		Incorrect	3	3	6	8	3	6
		False Positive	6	6	6	3	3	3
		Not Classified	0	0	0	0	0	0
'Basalt / Andesite'	B/A vs. (K & M-K)	Correct	18	18	18	16	18	18
		Incorrect	0	0	0	2	0	0
		False Positive	0	0	0	1	0	0
		Not Classified	0	0	0	5	0	0
'Kaibab' and 'Moen-Kopi'	Hierarchical. Remove B/A Classify K vs. M	Correct	15	15	16	13	19	
		Incorrect	6	6	5	2	2	
		False Positive	0	0	0	0	0	
		Not Classified	0	0	0	6	0	

Principal Component Analyses (PCA) plots illuminate how the application of EPO for decluttering can succeed or fail, dependent on the degree of decluttering. From the perspective of fit to the training set, increasing the degree of EPO decluttering condenses samples from each of the classes nearer to the class mean (**Figure 2 At, Bt, and Ct**). However, EPO can eliminate systematic variance that is useful for differentiating between classes, leaving only spurious correlations to define the classes. This process can be seen in the class locations of the validation sets (**Figure 2 Av, Bv, and Cv**) relative to the classes in the training sets. A 1-factor EPO sharpens the classes relative to the not decluttered data while maintaining colocalization of the validation set. However, the full-rank EPO largely leaves random correlations to define classes; consequently, the validation set exhibits greater spread across the PC space. As expected, the 6-factor EPO and full-rank EPO models perform much worse than the models with weaker decluttering.

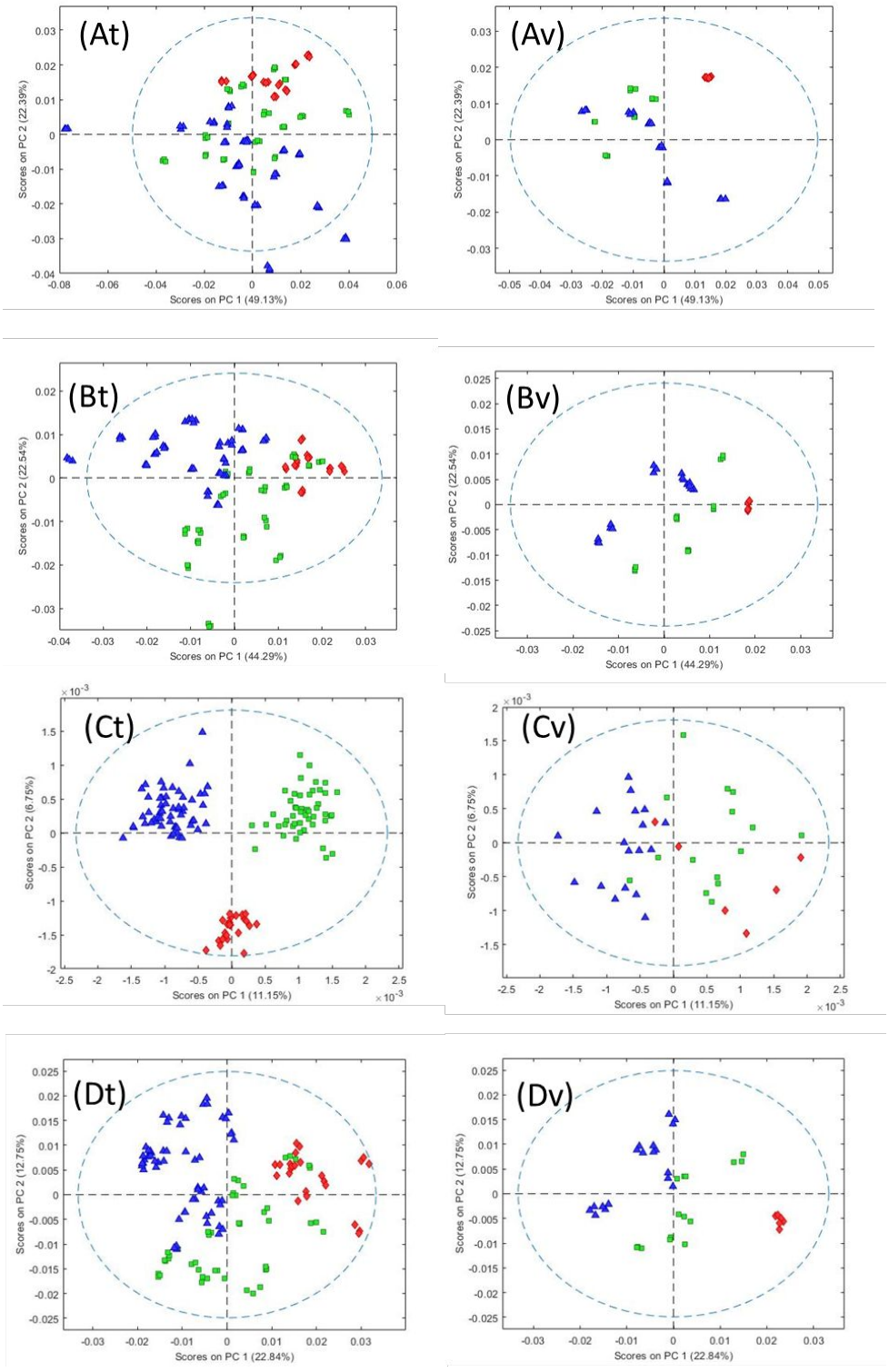

Figure 2. PCA scores plots of four XRF ash training sets (_t) and validation sets projected into the training set space (_v) for the untreated data (A_), data following 1-factor EPO (B_), data following full-factor EPO (C_), and data following CAIM discretization (D_). Samples were from three classes: soil derived

from Kaibab Limestone (red diamonds), soil derived from Moenkopi Formation (green squares), and basalt/andesite soil (blue triangles).

CAIM discretization of the pine ash XRF data results in better model performance than EPO decluttering (**Table 3, column 8 v. Table 3, columns 4-7**). By design, the CAIM algorithm strives to discretize each variable to be aligned with the known training set classification. Although CAIM was conceived to help normalize the variance within each class, the method has the added benefit of simultaneously reducing clutter within each of the classes. Three modeling strategies were investigated with CAIM discretization. It is observed that CAIM works best when discretizing for a binary classifier; in this case, the model seeks to distinguish one target class from all other classes combined as a group. As set of binary classifiers, CAIM outperforms both undecluttered PLS-DA and EPO(1) filtered PLS-DA, returning half the number of false positives and more samples classified with greater than 90% probability (**Table 3, column 8 versus Table 3, columns 4-5**).

CAIM filtering also performs better than no decluttering and EPO(1) filtering when the data are modeled as a hierarchical tree of binary classifiers (**Table 3, last row**). All methods can distinguish basalt/andesite from the other soils with 100% selectivity and sensitivity. However, when distinguishing between Kaibab and Moenkopi, CAIM had only 2 misclassifications out of 24 samples while the other two methods each had 6 misclassifications.

CAIM discretization avoids the issue of losing systematic variance needed for classification, unlike EPO decluttering (**Figure 2 Dt and Dv**). With CAIM discretization the distributions of the classes, in both the training and validation sets, are sharper and better resolved than no decluttering or EPO filtered classes. Comparing the EPO results (**Figure 2C**) to CAIM results (**Figure 2D**), it is clear that samples in the validation set lie outside of the boundaries of the training set in the PC space for EPO, but not for CAIM. This is especially evident in the training and validation set locations for the Kaibab Limestone (red) samples and basalt/andesite (blue) samples. The disparity in training vs validation set locations in the PC space is evidence of overfitting during EPO decluttering.

One caveat to the use of CAIM is that CAIM often performs worse when more than two functional classes are in the model. Applying CAIM here to resolve all three classes in a single flat model resulted in twice the number of misclassified samples than with a set of binary classifiers (**Table 3, column 9**). With binary classifier models, a unique set of discretization intervals is found to maximize Equation 1 for each model. With three or more classes, CAIM is less likely to derive a discretization scheme that is optimal for every class.

4.2. Classification of colored glasses

Preliminary analysis of the variance sources within this data set indicates a high probability for successful classification. Treating each rod as a unique class prior to elimination of uninformative variables, the average variance within the 12 replicates from each glass rod is only 22.1% of the variance among the observed class means. Eliminating the 15,262 uninformative variables eliminates only 0.0012 units of variance observed both from within and between classes. This reduces the mean interclass variance to 20.7% of the total variance with 79.3% of the total variance being between the classes (**Table 4, column 2**). Use of Principal Components Analyses (PCA) and Hierarchical Cluster Analyses

(HCA) presents a visual snapshot of the observed class overlap for the 21 glass rods (**Figure 3a and 3b**). Despite of this overlap, a 2 PC model describes ~80% of the total variance and a K-Nearest Neighbors with K = 1 (K-NN-1) model (removing 1/4th of each class to form a test set) accurately classifies all but 1 sample in each of the training and test sets (**Table 4, rows 3 and 4**).

Table 4: Effect of External Parameter Orthogonalization (EPO) decluttering on distribution of variance across the red glass data set and performance of PCA and K-NN modeling.

Treatment	**Undecluttered**	**EPO(1)**	**EPO(6)**	**EPO(Full Rank)**
Variance of class means	0.0775 (79.3%)	0.0670 (95.7%)	0.0544 (98.7%)	0.0127 (99.9%)
Mean variance of replicates	0.0161 (20.7%)	0.00285 (4.3%)	0.000724 (1.3%)	0.0000229 (0.2%)
PCA	# PC: 7 Cum Var: 82.2%	# PC: 6 Cum Var: 80.6%	# PC: 5 Cum Var: 81.6%	# PC: 5 Cum Var: 79.8%
KNN Misclassified (9:3 split)	1NN: 1 Cal; 1 Pred 3NN: 5 Cal; 1 Pred	1NN: 2 Cal; 1 Pred 3NN: 7 Cal; 1 Pred	1NN: 1 Cal; 1 Pred 3NN: 5 Cal; 1 Pred	1NN: 0 Cal; 0 Pred 3NN: 0 Cal; 0 Pred

Construction and application of one class versus all other classes models for each of the three classes with test-set samples shows that this is, in truth, an easy classification problem for PLS-DA (**Table 6**). The models perform with 100% success using four factors without the need for EPO. Similarly, the models perform well after application of EPO with a small basis set of principal components (<6 PC) to declutter the data. However, when full-rank EPO is applied to declutter, most of the test-set samples are far enough from the mean of the training set samples that validation-set samples are deemed 'indeterminate' in assignment. By comparison, discretization of the data leads to a more parsimonious PLS-DA model with 100% accuracy. PLS-DA with discretization reduces the required complexity of the model from four factors to only one factor.

Table 5: PLS-DA results for classification of four red glass samples with different External Parameter Orthogonalization (EPO) and CAIM discretization strategies for spectral pretreatment.

Sample		**Undecluttered**	**EPO (1PC)**	**EPO (6PC)**	**EPO (Full Rank)**	**CAIM**
'G'	Factors:	4	4	4	4	1
	Correct	12	12	12	0	12
	Incorrect	0	0	0	0	0
	Not Classified	0	0	0	12	0
	False Positives	0	0	0	0	0
'J' and 'R'	Factors:	5	4	4	3	1
	Correct	24	24	24	1	24
	Incorrect	0	0	0	0	0
	Not Classified	0	0	0	23	0
	False Positives	1	1	0	0	0
'S'	Factors:	4	4	4	4	1
	Correct	12	12	11	2	12

	Incorrect	0	0	1	0	0
	Not Classified	0	0	0	10	0
	False Positives	0	0	0	0	0

The failure of PLS-DA to successfully classify the training set glasses with full-rank EPO highlights the problems associated with EPO. EPO is designed to reduce the intra-class variance, and Tables 2 and 5 show EPO performs well at this task. However, a reduction in inter-class variance may be an unintended consequence. In the ash data, the intra-class variance is reduced by three orders of magnitude while the inter-class variance is reduced by two orders of magnitude when progressing from un-decluttered data to full-rank EPO. By way of comparison, applying full-rank EPO to the glass data also reduces the intra-class variance by three orders of magnitude, but the inter-class variance is only reduced by a factor of 5. Much less inter-class variance is lost with EPO on the glass data than with EPO applied to the ash data. This would explain why EPO on the ash failed with fewer EPO factors than when EPO eventually fails on the glass. Given that full-rank EPO does fail when decluttering the glass data, a migration of the test set within the PC space of the decluttered training set space is evident in the glass data (**Figure 3**) as it is in the ash data (**Figure 2C**)

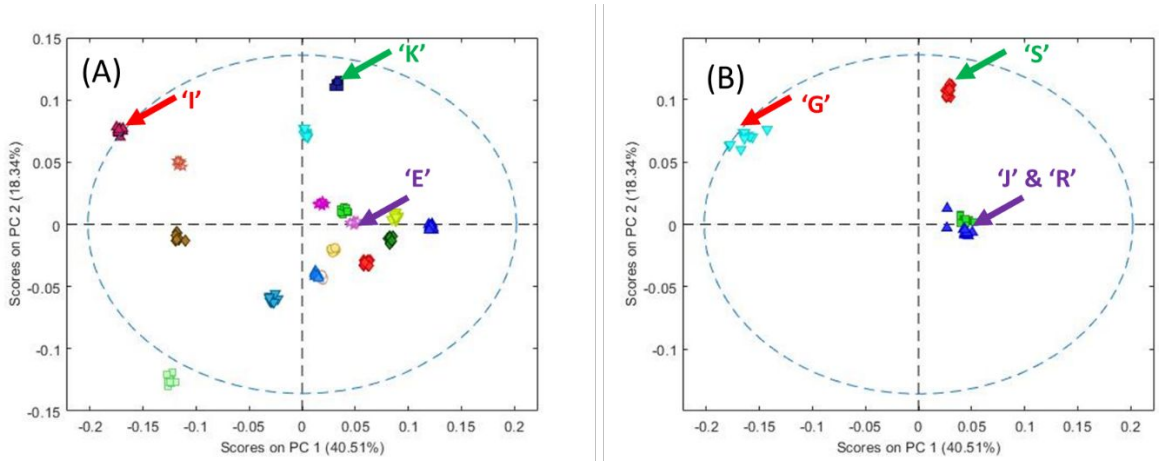

Figure 3: Score plot of the red glass training set (A) and test set (B) set in the two principal component (PC) space defined by the training set. The test set is comprised of 4 glass rods that are duplicated colors to glass rods in the training set (e.g., rods 'I' and 'G' are purported to be the same color by the manufacturer). Arrows are set to the exact same location in each plot and serve as a reference to visualize the slightly different locations of the training and test classes following application of full-rank External Parameter Orthogonalization (EPO).

Discretization both avoids the concern of optimizing the number of EPO factors and potentially offers a more parsimonious PLS-DA model. Parsimonious models generally perform better than more complicated models because the higher factors generally have a worse signal-to-noise ratio than the initial factors. PLS-DA, with or without EPO, required four factors to successfully classify any one of the glass colors. However, with discretization, only a 1-factor PLS-DA model is required (**Table 6**). Score plots

of the discretized data illuminate how discretization enhances the ability of PLS-DA to differentiate among classes (**Figure 4**). For each model, a set of discretization rules, D, is adopted to more optimally distinguish between the target class and all other classes for every variable. Discretization inherently declutters the data while separating the target class from most of the other observations. In the case for the glasses, a 7-PC model is still needed to describe all the variance within the training set following discretization, regardless of the target class. However, within the 7-PC space are clear planes of demarcation between the target class and other aggregated classes. For example, for the determination of class 'K'/'S', the one-dimensional demarking is best seen in a score plot of PC 5 versus PC 6 (**Figure 4A**). However, the separation is evident in other PCs also. Similar plots show the same effect for the other target classes. Recall that a unique set of discretization rules is determined for each target class based on the training set and then apply these same rules to future samples; hence, the discretization scheme inherently enhances the development of classification models. PLS-DA can exploit these differences with a simple 1-factor model.

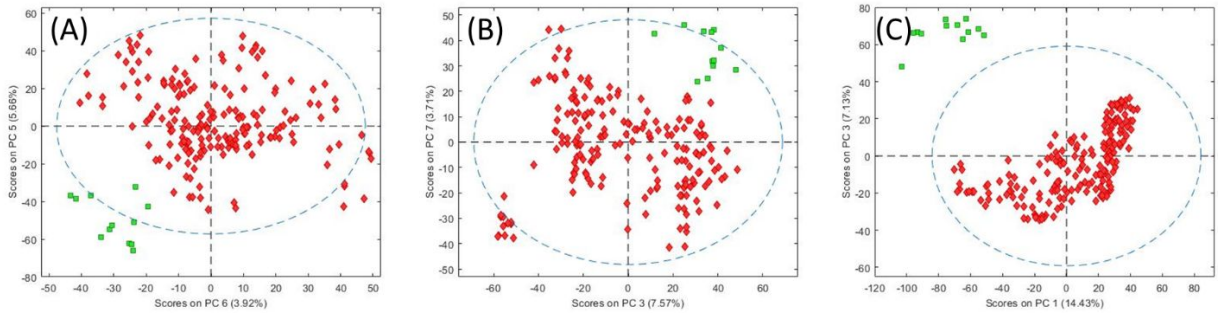

Figure 4: Score plots for the red glass training set following discretization in preparation for one class versus all other classes PLS-DA model for training set glasses 'K' (A), 'E' (B), and 'I' (C). These models correspond to test glasses 'S', 'J' and 'R', and 'G', respectively. The target training glass is green while the other aggregated classes are red. While a 7 PC model is needed two describe all the systematic variance in the discretized, the 2-PC score plot that best shows separation is presented here.

Discretization can also be applied to enhance PLS-DA models in decision trees. For example, a classification problem may be better approached by a series of two-way classifications in a hierarchical decision tree as opposed to a set of one class versus all other classes models. In these situations, each nominal class would be assigned to one of two separate super groups based on their proximities in a higher dimension space. Discretization rules would be optimized to differentiate among the super groups for each variable. In the case of the glass data, there is a natural break splitting the 17 classes into a 7-class super group and a 10-class super group. These two groups of classes are significantly better resolved following discretization (**Figure 5**). In fact, the potential for a second binary split, differentiating among each group in the upper and lower halves of the plot, is evident. This would further divide the 10-class super group (red) into a 6-class and a 4-class super group, respectively. Similarly, the 7-class super group could be further split into a 4-class and 3-class group.

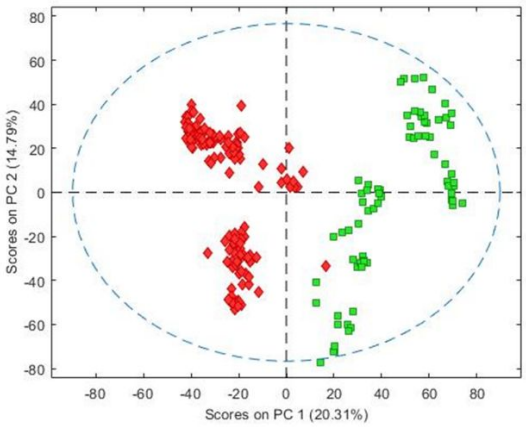

Figure 5: Score plot following discretization to split the 17-class problem into a 7-class group (green) and a 10-class group (red) as the first step of a potential hierarchical decision tree approach.

4.3. Classification of Dalbergia

Analysis of the *Dalbergia* data set by PLS DA following supervised discretization yielded a sensitivity of 0.98 and a selectivity of 0.98 (**Table 6**). This performance is comparable to that observed from PLS-DA using external parameter orthogonalization where the sensitivity was also 0.98 but the selectivity was 0.99. Each of these strategies resulted in one sample from the three prediction sets that was misclassified. With discretization, one sample from Class 5 was miss-assigned to Class 3, whereas with EPO one sample from Class 3 was assigned to Class 7. However, with discretization, samples were ambiguously assigned in six separate incidences. For example, in the validation set, three samples from Class 1 were both accurately assigned to Class 1, but also ambiguously identified as members of Class 4, Class 6, and Class 9.

Table 6: Performance of PLS-DA with CAIM discretization on the classification of *Dalbergia* samples analyzed by LIBS.

Class	Model	PCs	Validation Set 1	Validation Set 2	Validation Set 3	Sensitivity	Selectivity
1	1 vs (2 & 4)[a]	2	1,1	1,1	1,1	1.00	1.00
2	2 vs (1 & 4)[a]	4	2,2	2,2	2,2,4	1.00	0.98
3	3 vs 5 [b]	2	3,3	3,3,5	3,3	1.00	0.98
4	4 vs (1 & 2)[a]	3	1,4,4	4,4	1,4,4	1.00	0.96
5	3 vs 5 [b]	2	5,5	5	5.5	0.83	1.00
6	6 vs 9 vs 1-5 [c]	3	6,6	6,6	1,6,6	1.00	0.98
7	7 vs 8 vs all [c]	3	7,7	7,7	7,7	1.00	1.00
8	7 vs 8 vs all [c]	3	8,8	8,8	8,8	1.00	1.00
9	6 vs 9 vs 1-5 [c]	3	9,9	2,9,9	1,9,9	1.00	0.96
	TOTAL:					**0.98**	**0.98**

a. Flat classification
b. Binary classification
c. One-versus-all others classification

A four-level hierarchical decision tree was constructed to best classify the nine species of exotic hardwoods (**Figure 6A**). This tree was constructed to perform as many classifications as possible at each level. For example, at the first level the decision is made between classifying a sample as belonging to Class 7, Class 8, or the set of all other classes. Because overall sensitivity of a particular classification is the multiplicative factor of sensitivities at every prior decision node, classification is generally viewed to be more reliable at the top of the tree than at the bottom, and that the net sensitivity decreases as the decisions move down the tree. Consequently, a flat classifier was used at each node provided that it performed as well or better than a series of two-way classifiers. This was the case at first two levels of the decision tree. However, at the third level a flat classifier could not distinguish between classes one, two, and four or between classes three and five; the best model could only differentiate between these two groups of classes. Differentiation among each of these groups was then performed on lowest, final level of the decision tree.

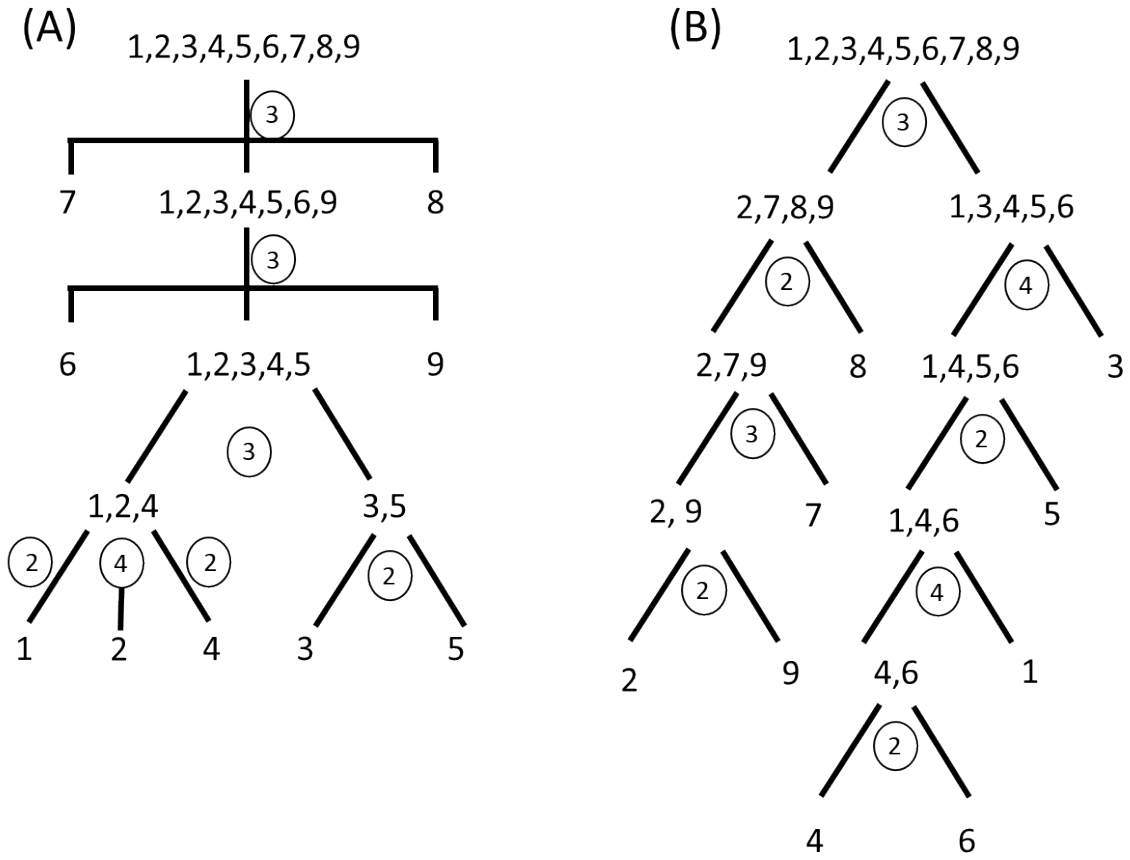

Figure 6: Hierarchical decision trees used for speciation of nine *Dalbergia* classes using PLS-DA with (A) discretization to remove the effect of clutter and (B) external parameter orthogonalization (EPO) to remove the effects of sample clutter. The numbers inside each circle indicate the number of factors in the PLS-DA model at each decision node.

The analysis of LIBS spectra from wood samples proved to be a particularly challenging application for discretization. The LIBS spectra have two sources of variance that together serve to frustrate the discretization algorithm. First, the overall efficiency of collecting a LIBS spectrum varies greatly from location to location due to differences sample density and moisture content. Thus, the LIBS signal of two different classes at a particular wavelength that may be well separated in intensity under ideal circumstances, may become confounded as the overall LIBS intensity varies. Such a problem could be corrected by normalizing each spectrum to unity. Spectral normalization was attempted and not proven beneficial for this application. There appears to be a second source of non-probative variance in the LIBS spectra that originates from either surface contamination or the history of the wood samples. This is the type of variance that is appropriately removed by external parameter orthogonalization and would justify why EPO worked better for this data set, in general, than did discretization.

Comparing the hierarchical decision tree optimized from the discretized data (**Figure 6A**) to the hierarchical decision tree optimized for EPO-based clutter removal in the previous study[27] (**Figure 6B**) provides valuable insight into the roles that both discretization and EPO can play in developing the 'best' hierarchical model for any given application; a truly optimized method would rely on discretization, EPO, or other methods as appropriate to construct compact decision tree possible. For example, discretization and EPO each worked relatively better than the other for separation of different classes from the bulk of the data. Discretization rapidly resolved Class 7 and Class 8 but struggled with resolving Class 3 from Class 5. On the other hand, EPO was able to resolve Class 3 from Class 5 with better success higher up the decision tree than with discretization, whereas Class 7 and Class 8 were better resolved farther down the decision tree than with discretization.

The ability to rely on the different strengths of each strategy to reduce clutter and improve classification is particularly beneficial to the ultimate goal of the *Dalbergia* classification project. *Dalbergia* is an endangered exotic hardwood that is subject to the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES)[29] that restricts logging, export, and import of different rosewood species. Consequently, multiple federal and international law enforcement agencies are interested in building a database and model for rapid determination of CITES compliance. Handheld LIBS is one of the methods under consideration for this role. To assess compliance, a model does not need to unambiguously determine the identity of a suspected *Dalbergia* log or sample. Instead, all that is needed is to determine whether the actual species of the exotic wood agrees with this species specified on the manifest. To best accomplish this, a separate model for each species in the library could be optimized for sensitivity and selectivity using the available tools as needed.

Conclusions

Supervised discretization, such as performed by CAIM, provides a reliable alternative to External Parameter Orthogonalization (EPO) for decluttering multivariate chemical sensor data. With EPO the number of factors in the model warrant careful consideration; too many can lead to prediction biases from overfitting the decluttering step. Because supervised discretization by the CAIM algorithm has no user adjustable parameters, CAIM is less prone to overfitting than EPO and more amenable to automated implementation. The one caveat to implementing CAIM is the method seems to work best with simple models where there are only two or three classes to be discerned. For more complicated

multiclass models, this leads to better performance of hierarchical decision trees than with flat classifiers.

Acknowledgements

The authors thank NSF CHE 2003839 and NSF CHE-2003867 for support of this project. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

References

(1) Yang, Y.; Webb, G. I.; Wu, X. Discretization Methods. In *Data Mining and Knowledge Discovery Handbook*; Springer US, 2009; pp 101–116. https://doi.org/10.1007/978-0-387-09823-4_6.

(2) Liu, H.; Hussain, F.; Tan, C. L.; Dash, M. Discretization: An Enabling Technique. *Data Min Knowl Discov* **2002**, *6* (4). https://doi.org/10.1023/A:1016304305535.

(3) Mizianty, M. J.; Kurgan, L. A.; Ogiela, M. R. Discretization as the Enabling Technique for the Nave Bayes and Semi-Nave Bayes-Based Classification. *Knowledge Engineering Review* **2010**, *25* (4). https://doi.org/10.1017/S0269888910000329.

(4) Thaiphan, R.; Phetkaew, T. Comparative Analysis of Discretization Algorithms on Decision Tree. In *Proceedings - 17th IEEE/ACIS International Conference on Computer and Information Science, ICIS 2018*; 2018. https://doi.org/10.1109/ICIS.2018.8466449.

(5) Dougherty, J.; Kohavi, R.; Sahami, M. Supervised and Unsupervised Discretization of Continuous Features. In *Machine Learning Proceedings 1995*; 1995. https://doi.org/10.1016/b978-1-55860-377-6.50032-3.

(6) Lavangnananda, K.; Chattanachot, S. Study of Discretization Methods in Classification. In *2017 9th International Conference on Knowledge and Smart Technology: Crunching Information of Everything, KST 2017*; 2017. https://doi.org/10.1109/KST.2017.7886082.

(7) Wu, X. A Bayesian Discretizer for Real-Valued Attributes. *Comput J* **1996**, *39* (8), 688–691. https://doi.org/10.1093/comjnl/39.8.688.

(8) Kumar, A.; Zhang, D. Biometric Recognition Using Entropy-Based Discretization. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*; 2007; Vol. 2. https://doi.org/10.1109/ICASSP.2007.366188.

(9) Kerber, R. Chimerge: Discretization of Numeric Attributes. In *Proceedings Tenth National Conference on Artificial Intelligence*; 1992.

(10) Tay, F. E. H.; Shen, L. A Modified Chi2 Algorithm for Discretization. *IEEE Trans Knowl Data Eng* **2002**, *14* (3). https://doi.org/10.1109/TKDE.2002.1000349.

(11) Sriwanna, K.; Puntumapon, K.; Waiyamai, K. An Enhanced Class-Attribute Interdependence Maximization Discretization Algorithm. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; 2012; Vol. 7713 LNAI. https://doi.org/10.1007/978-3-642-35527-1_39.

(12) Li, M.; Deng, S.; Feng, S.; Fan, J. An Effective Discretization Based on Class-Attribute Coherence Maximization. *Pattern Recognit Lett* **2011**, *32* (15). https://doi.org/10.1016/j.patrec.2011.08.008.

(13) Kurgan, L.; Cios, K. J. Discretization Algorithm That Uses Class-Attribute Interdependence Maximization. *Proc. of the 2001 International Conference on Artificial Intelligence (ICAI-2001)* **2001**.

(14) Kurgan, L. A.; Cios, K. J. CAIM Discretization Algorithm. *IEEE Trans Knowl Data Eng* **2004**, *16* (2). https://doi.org/10.1109/TKDE.2004.1269594.

(15) Wold, S.; Antti, H.; Lindgren, F.; Ohman, J. Orthogonal Signal Correction of Near-infrared Spectra. *Chemometrics Intell. Lab. Syst.* **1998**, *44*, 175–185.

(16) Westerhuis, J. A.; de Jong, S.; Smilde, A. K. Direct Orthogonal Signal Correction. *Chemometrics Intell. Lab. Syst.* **2001**, *56*, 13–25.

(17) Svensson, O.; Kourti, T.; MacGregor, J. F. An Investigation of Orthogonal Signal Correction Algorithms and Their Characteristics. *J Chemom* **2002**, *16* (4), 176–188. https://doi.org/10.1002/cem.700.

(18) Andries, E.; Kalivas, J. H. Interrelationships between Generalized Tikhonov Regularization, Generalized Net Analyte Signal, and Generalized Least Squares for Desensitizing a Multivariate Calibration to Interferences. *J Chemom* **2013**, *27* (5), 126–140. https://doi.org/10.1002/cem.2501.

(19) Roger, J.-M.; Chauchard, F.; Bellon-Maurel, V. EPO–PLS External Parameter Orthogonalisation of PLS Application to Temperature-Independent Measurement of Sugar Content of Intact Fruits. *Chemometrics and Intelligent Laboratory Systems* **2003**, *66* (2), 191–204. https://doi.org/10.1016/S0169-7439(03)00051-0.

(20) Amirvaresi, A.; Parastar, H. External Parameter Orthogonalization-Support Vector Machine for Processing of Attenuated Total Reflectance-Mid-Infrared Spectra: A Solution for Saffron Authenticity Problem. *Anal Chim Acta* **2021**, *1154*, 338308. https://doi.org/10.1016/j.aca.2021.338308.

(21) Martins, H.; Naes, T. *Multivariate Calibration*; Wiley, 1989.

(22) Fayyad, U. M.; Irani, K. B. On the Handling of Continuous-Valued Attributes in Decision Tree Generation. *Mach Learn* **1992**, *8* (1), 87–102. https://doi.org/10.1007/BF00994007.

(23) Cios, K. J.; Pedrycz, W.; Swiniarski, R. W. *Data Mining Methods for Knowledge Discovery*; Springer, 1998.

(24) Wong, A. K. C.; Liu, T. S. Typicality, Diversity, and Feature Pattern of an Ensemble. *IEEE Transactions on Computers* **1975**, *C–24* (2), 158–181. https://doi.org/10.1109/T-C.1975.224183.

(25) Eilers, P. H. C.; Boelens, H. F. M. *Baseline Correction with Asymmetric Least Squares Smoothing*; 2005.

(26) Savitzky, A.; Golay, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal Chem* **1964**, *36* (8). https://doi.org/10.1021/ac60214a047.

(27) Celani, C. P.; Lancaster, C. A.; Jordan, J. A.; Espinoza, E. O.; Booksh, K. S. Assessing Utility of Handheld Laser Induced Breakdown Spectroscopy as a Means of *Dalbergia* Speciation. *Analyst* **2019**, *144* (17), 5117–5126. https://doi.org/10.1039/C9AN00984A.

(28) de Boves Harrington, P. Statistical Validation of Classification and Calibration Models Using Bootstrapped Latin Partitions. *TrAC Trends in Analytical Chemistry* **2006**, *25* (11), 1112–1124. https://doi.org/10.1016/j.trac.2006.10.010.

(29) *Convention on International Trade in Endangered Species of Wild Fauna and Flora. The CITES Appendices*. https://www.cites.org/eng/app/index.php.