Investigating Complex LIBS Samples Through the Integration of Raman Spectroscopy and Advanced Machine Learning Methods

Sofia Pozsonyiova[1] , Prasoon K. Diwakar [2]

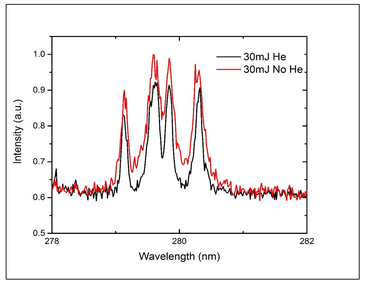
1. Macalester College, St. Paul, MN, 55105, USA,

2. South Dakota School of Mines and Technology, Rapid City, SD, 57701, USA

Introduction

Laser-induced breakdown spectroscopy (LIBS) is an optical emission spectroscopy technique which relies on large emission spectral data sets to understand and interpret spectral information. Dendrograms are commonly used for this task as they provide an appealing tree-based representation, however, there are a number of algorithmic issues that arise with this method. For example, when observations are partitioned into subgroups, 1 to *n*, dendrograms alone cannot tell which observations form each of the particular clusters.

Thus, this work refines and explores additional unsupervised learning methods to more effectively visualize and classify spectra data. These additional methods include the creation of a new method to be able to distinguish and directly extract each observation within a certain cluster in the dendrogram; as well as the addition of as well as the addition of K-Means methods in conjunction with Raman spectroscopy data for faster and more reliable sample identification, classification, and pattern recognition.



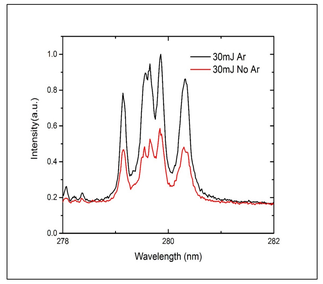


Figure 1. Raw Spectra of a sample element Mg, in Ar, He, N at Energy 30 mJ

Data

- Constituted of spectra obtained from Lead (Pb), Chromium (Cr), and Tin (Sn) samples that were ablated in the presence of various ambient gases Argon (Ar), Helium (He), and Nitrogen (N) at different pulse energies varying from 5 to 100 mJ.
- The second data set constituted of spectra obtained from various mining sites. These ore samples are still being identified.

Tools

- Open-Source Statistical Software R
- Data Wrangling Packages: CRAN, Dpylr, Gower, Tidyr , Caret
- Visualization Packages: Ggplot2, Xtable, Plotly

Visualizations

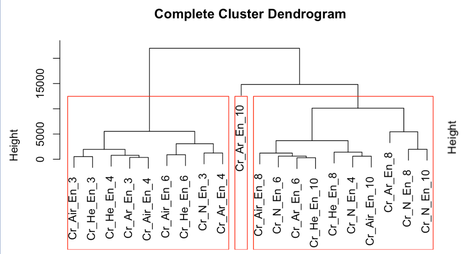


Figure 2: Visualization of a Complete Cr cluster dendrogram using Euclidean Distance

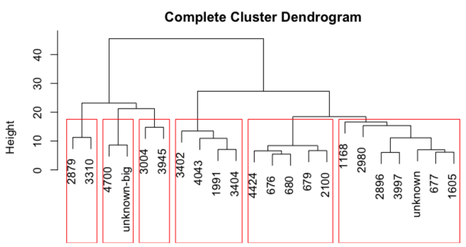


Figure 3: Visualization of future work...identifying unknown samples

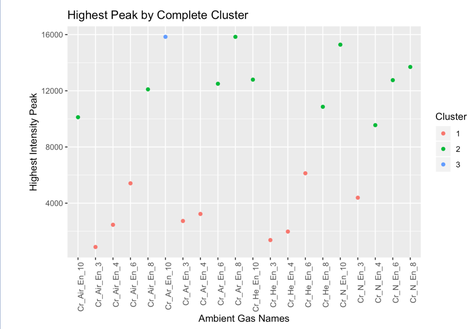


Figure 4: Ggplot visualization of a Complete Cr cluster spread using Euclidean Distance

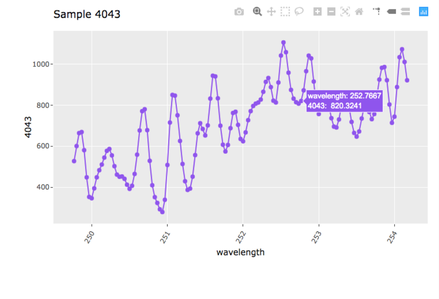


Figure 5: Implementation of Plotly and ggplot to improve spectra visualization

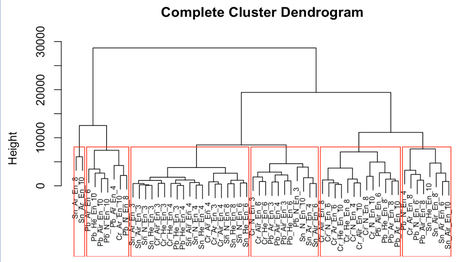


Figure 6: Ggplot visualization of a Complete cluster spread of the entire data set

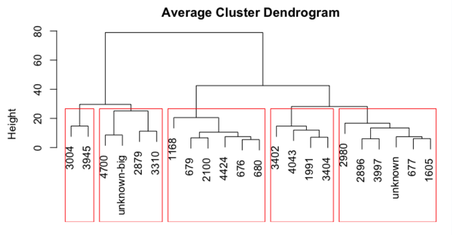


Figure 7: Ggplot visualization of an Average cluster spread of our future work data set

Observations

- The addition of gases such as argon (Ar) and Helium (He) to LIBS samples can have drastic effect on the spectra signals.
- Argon gas consistently showed an amplification in signal intensity due to increase in temperature of plasma, while Helium effect was dependent on the sample being tested. Nitrogen effect was in between Ar and He.
- In our ore data set we see that there are two main groups that are forming and within those groups are subsets. This could indicate in similarities such as region, sample elements, etc…

Discussion

- Effects were adequately captured by classification approach
- Clustering analysis can serve as an important first step to understanding the possible limitations and capabilities of an obtained spectra data set
- With the implementation of R package Dplyr we are now able to see which observations and distinct values constitute each cluster within the dendrogram
- Choice of dissimilarity measure, standardization, linkage-type, and height cut will all result in different patterns, therefore, parameters should be chosen wisely

	Intensity_Peak_High	Cluster	Name
1	880.88	1	Cr_Air_En_3
2	1373.50	1	Cr_He_En_3
3	2732.25	1	Cr_Ar_En_3
4	4389.81	1	Cr_N_En_3
5	2461.62	1	Cr_Air_En_4
6	1977.75	1	Cr_He_En_4
7	3232.75	1	Cr_Ar_En_4
8	5414.19	1	Cr_Air_En_6
9	6127.75	1	Cr_He_En_6
10	9554.88	2	Cr_N_En_4
11	12501.06	2	Cr_Ar_En_6
12	12758.19	2	Cr_N_En_6
13	12102.81	2	Cr_Air_En_8
14	10861.94	2	Cr_He_En_8
15	15841.19	2	Cr_Ar_En_8
16	13698.50	2	Cr_N_En_8
17	10120.44	2	Cr_Air_En_10
18	12794.19	2	Cr_He_En_10
19	15285.25	2	Cr_N_En_10
20	15843.56	3	Cr_Ar_En_10

Table 1: Cr Complete Cluster Breakdown

	Intensity_Peak_High	Cluster	Name
1	880.88	1	Cr_Air_En_3
2	1373.50	1	Cr_He_En_3
3	2732.25	1	Cr_Ar_En_3
4	4389.81	1	Cr_N_En_3
5	2461.62	1	Cr_Air_En_4
6	1977.75	1	Cr_He_En_4
7	3232.75	1	Cr_Ar_En_4
8	5414.19	1	Cr_Air_En_6
9	6127.75	1	Cr_He_En_6
10	9554.88	2	Cr_N_En_4
11	12501.06	2	Cr_Ar_En_6
12	12758.19	2	Cr_N_En_6
13	12102.81	2	Cr_Air_En_8
14	10861.94	2	Cr_He_En_8
15	13698.50	2	Cr_N_En_8
16	10120.44	2	Cr_Air_En_10
17	12794.19	2	Cr_He_En_10
18	15285.25	2	Cr_N_En_10
19	15841.19	3	Cr_Ar_En_8
20	15843.56	3	Cr_Ar_En_10

Table 2: Cr Complete Gower Cluster Breakdown

Figure 8 & 9 : Data wrangled table depicting each observation from each of the *k = 3* clusters

Limitations

- No real way of assessing the validity of unsupervised learning methods
- Techniques have not yet been designed or tailored specifically to spectral application
- Large amount of subjectivity that is dependent on research and data set

Future Work

- Applying this method to unknown ore samples to gain a better understanding of our initial data
- Incorporate additional variables such as elemental properties to get a better understanding of spectral similarity
- Create an R Package to tailor unsupervised learning methods to spectral data