G8 Team Project: Mineral Grains Identification using Data Science

Literature survey contribution

Peter Millitz (23088298)

1. Chen, Y., & Zhao, Q. (2021). Mineral exploration targeting by combination of recursive indicator elimination with the ℓ2-regularization logistic regression based on geochemical data. *Ore Geology Reviews*, *135*, 104213. https://doi.org/https://doi.org/10.1016/j.oregeorev.2021.104213

   This paper discusses a particular challenge in geochemical exploration involving the selection of indicator elements in mineral exploration and introduces a method called Recursive Indicator Elimination (RIE), which combines a recursive elimination process with Machine Learning (ML) techniques. The RIE method seeks to identify an optimal subset of geochemical elements that are closely related to mineral deposits. This optimal subset is subsequently used to develop an ML model, specifically, a logistic regression model using ℓ2-regularization, for effective mineral exploration targeting. The training data used to build the logistic regression model was heavily imbalanced.

   The study compares the RIE method with other approaches like Area Under the Curve (AUC) and Principal Component Analysis (PCA) in combination with logistic regression. The results show that the RIE combined with logistic regression yields the best performance in terms of mineral exploration targeting.

2. Yu, D., Lee, S. J., Lee, W. J., Kim, S. C., Lim, J., & Kwon, S. W. (2015). Classification of spectral data using fused lasso logistic regression. *Chemometrics and Intelligent Laboratory Systems*, *142*, 70-77. https://doi.org/https://doi.org/10.1016/j.chemolab.2015.01.006

   This paper presents a method for analysing high-dimensional spectral data frequently used in biological and medical research. The authors' main goal was to identify differently expressed peaks between two groups and create an efficient classifier using a technique called fused lasso logistic regression (FLLR).

   They showed that the FLLR has a grouping property on regression coefficients, which simultaneously selects a group of highly correlated variables together. They argued that both the sparsity and the grouping property of the FLLR can overcome some of the main challenges in the analysis of the spectral data. For example, shifts and misalignment of peaks caused by factors like instrumental instability or experimental differences which can reduce the accuracy of statistical analysis. Additionally, spectral data is often characterized by high dimensionality and low sample size.

   The spectral data studied in this paper had three key characteristics: the dimension of covariates was much larger than the number of samples (high-dimensionality), sparsity (many coefficients were zero) and sequential ordering of covariates. The authors demonstrated that the fused lasso regression technique was well-suited to these characteristics and provided benefits such as data-dependent binning of covariates and resolving issues related to misalignment of samples.