

COMPARATIVE STUDY OF ARTIFICIAL INTELLIGENCE TECHNIQUES IN OILS CLASSIFICATION FROM BRAZILIAN SEDIMENTARY BASINS EMPHASIZING DECISION TREES

Ana Cristina da Silva SERRA¹, Carlos Siqueira BANDEIRA de MELLO² and Mário Duncan RANGEL²

1. COPPE/UFRJ, Laboratory for Computation Methods in Engineering-LAMCE, PO: 68552, 21949-900, Rio de Janeiro, Brazil. 2. PETROBRAS/CENPES/PDEXP/Geoquímica, R. Horácio Macedo, Cid. Univ., Ilha do Fundão, 21949-900, Rio de Janeiro, RJ, Brazil

Nowadays geochemists have a large variety of geochemical data available to provide information about the paleoenvironmental characterization of oils (PETERS *et al.*, 2005). Since interpretation of this information is usually made through a visual analysis of many geochemical data, a technique that could provide the classification of these different paleoenvironments using the most important parameters would be very useful in the process of oil characterization.

The objective of this study was to evaluate the performance of several artificial intelligence techniques like decision trees, decision rules and neural networks in the classification of oils from different origins, trying to optimize the process of data interpretation and oil classification. The results of these classification techniques were compared with the classifications provided by previous geochemical studies.

The database for this classification study consisted in 2924 oil samples with different origins (lacustrine, marine and mixed), degrees of thermal evolution, and distinct levels of biodegradation from different Brazilian sedimentary basins. The parameters used as input variables in the classification model were the results of the following analyses: bulk, gas chromatography, liquid chromatography, stable carbon isotopes and GCMS and GCMSMS (biomarker ratios). The variables mostly affected by biodegradation and thermal alteration processes of oils had to be removed, to provide generic model using variables mostly related to the oils origin.

Firstly the database was submitted to a preliminary exploratory study to evaluate inconsistencies in data and *outliers* (samples with abnormal behavior) through boxplots and clusters and correlations between variables using scatterplots. After that several classification artificial intelligence techniques (QUINLAN, 1993), decision rules (WITTEN & FRANK, 1999), and neural networks (HAYKIN, 1999) and with special emphasis in decision trees were employed to evaluate the accuracy in classification of oils from different origins.

Using these techniques, classification results have shown that by using only 22 geochemical parameters it is possible to obtain a good match with a previous classification

given by the specialist geochemists with accuracy above 90%. For the subtypes where the number of samples is less than 10 the accuracy was reduced. The samples classification was made in two levels: first to classify in the 3 principal types of oils (lacustrine, marine and mixed) and later the classification in subtypes (e.g. lacustrine saline, lacustrine freshwater, marine siliciclastic, marine evaporitic, etc.).

If compared with the other classification techniques employed in this study, decision tree presented the best results, especially because this method provides the most important parameters selected to distinguish the different classes and the interval of occurrence of each parameter for each class, contrasting with neural networks that only shows the % accuracy in the classification. In the figure 1 the decision tree used for the classification of mixed oils is show, where it is possible to examine all the parameters selected for the discrimination between the groups and its values. The parameters are ratios between C_{26} and C_{28} tricyclic terpanes (26/28 TRI) and ratio between C_{30} tetracyclic polyprenoids (21R + 21S) and C_{27} diasteranes (20R + 20S) (TPP).

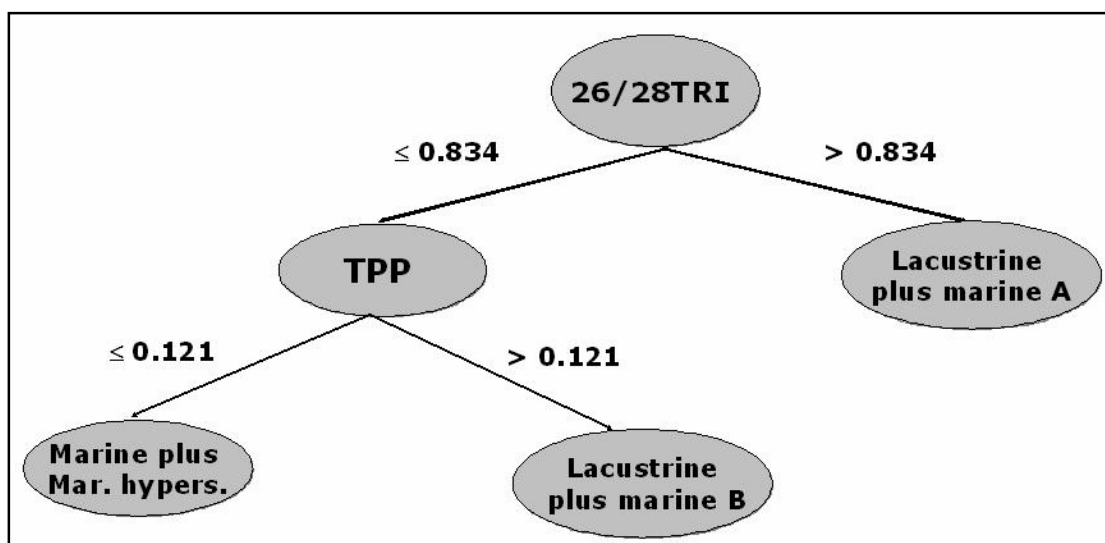


Figure 1. Decision tree obtained for the classification of mixed oils, showing the biomarker ratios used for the classification in the Lacustrine plus marine A, Lacustrine plus marine B and Marine plus marine hypersaline.

REFERENCES

- HAYKIN, S., 1999, *Neural Networks: A Comprehensive Foundation*, 2 ed., New Jersey, Prentice Hall.
- PETERS, K. E., MOLDOVAN, J.M., CLIFFORD, C., 2005, *The Biomarker Guide Volume 2: Biomarkers and Isotopes in Petroleum Exploration and Earth History*. 2^{ed.}, Cambridge University Press.
- QUINLAN, J.R., 1993, *C4.5: Programs for Machine Learning*, 1 ed., California, Morgan Kaufmann.
- WITTEN, I.H., FRANK, E., 1999, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, 1 ed., California, Morgan Kaufmann.