A comparison of CART and discriminant analysis of morphometric data in foraminiferal taxonomy

Pratul Kumar Saraswati1 & Sanjeev V. Sabnis2

1Department of Earth Sciences, Indian Institute of Technology – Bombay, Powai, Mumbai 400076 India
2Department of Mathematics, Indian Institute of Technology – Bombay, Powai, Mumbai 400076 India
pratul@iitb.ac.in

Traditionally, taxonomists have relied on qualitative and dichotomous characters for foraminiferal classification and discrimination. The advantages of quantification of morphological characters have been demonstrated by many in resolving the problems of taxonomy and evolutionary history. The unprecedented growth of information technology has made a strong case for quantification of morphological data and automated recognition of foraminiferal species. The automated recognition is possible either by image analysis or by statistical methods. The most commonly used statistical procedure by paleontologists is discriminant analysis of morphological data. The technique was originally developed based on the assumption of multivariate normality of the data set. The morphometric data generally deviate from normal distribution and, therefore, it violates the basic assumption of the method used. Here we use an alternative, non-parametric method, called Classification and Regression Tree (CART), to distinguish three closely resembling species of Middle Eocene Nummulites from western India. The species include N. beaumonti, N. neglectus and N. stamineus. We measured diameter and thickness of the test and five other parameters in oriented sections of 55 specimens. The data was analyzed by CART using Statistical Analysis Software (XLMiner). Thirty three data sets were randomly selected as training data to construct the tree. The tree uses two variables, thickness (T) of the test and height of the chamber in the final whorl (HL), as important variables to distinguish the three species. In the validation-data only one of the sixteen specimens was misclassified, while all the six test-data were assigned to correct classes. The same morphometric data were also analyzed by multi-group discrimination. The diameter and thickness of the tests and height and length of chambers in the final whorl were found to be most suitable for discrimination. Two discriminant functions account for nearly 100% discrimination. Both, CART and multigroup discriminant
A comparison of CART and discriminant analysis of morphometric data in foraminiferal taxonomy
Pratul Kumar Saraswati & Sanjeev V. Sabnis

analysis thus appear to be equally efficient in discriminating the three species. However, CART reduces the number of requisite variables for prediction without increasing the misclassification error. The most important advantage of CART though is that the success of the technique in classification does not depend on the normality of the data. It is not easy to determine to what extent the deviation from normality would affect the performance of the technique and therefore the reliability of prediction. The applications of CART in medicine and anthropology have shown that the technique also handles the missing data very efficiently. We suggest CART as a better option for class assignment in foraminiferal taxonomy for several reasons:

1) it does not require normal distribution of variables;
2) it can also accept data-set with some missing variables;
3) it may reduce the number of variables required for prediction; and
4) it quickens species discrimination and, therefore useful for well-site geologist.