Anuário do Instituto de Geociências - UFRJ
ISSN 0101-9759 Vol. 29 - 1 / 2006 p. 153-162

FORAMS 2006

Comparison of CART and Discriminant Analysis of Morphometric Data in Foraminiferal Taxonomy

Pratul Kumar Saraswati[1] & Sanjeev V. Sabnis[2]

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

Received: 15/07/2006 Accepted: 15/08/2006

Abstract

Paleontologists use statistical methods for prediction and classification of taxa. Over the years, the statistical analyses of morphometric data are carried out under the assumption of multivariate normality. In an earlier study, three closely resembling species of a biostratigraphically important genus *Nummulites* were discriminated by multi-group discrimination. Two discriminant functions that used diameter and thickness of the tests and height and length of chambers in the final whorl accounted for nearly 100% discrimination. In this paper Classification and Regression Tree (CART), a non-parametric method, is used for classification and prediction of the same data set. In all 111 iterations of CART methodology are performed by splitting the data set of 55 observations into training, validation and test data sets in varying proportions. In the validation data sets 40% of the iterations are correctly classified and only one case of misclassification in 49% of the iterations is noted. As regards test data sets, nearly 70% contain no misclassification cases whereas in about 25% test data sets only one case of misclassification is found. The results suggest that the method is highly successful in assigning an individual to a particular species. The key variables on the basis of which tree models are built are combinations of thickness of the test (T), height of the chambers in the final whorl (HL) and diameter of the test (D). Both discriminant analysis and CART thus appear to be comparable in discriminating the three species. However, CART reduces the number of requisite variables without increasing the misclassification error. The method is very useful for professional geologists for quick identification of species.

Keywords: Morphometrics; Foraminifera; *Nummulites;*Discriminant Analysis; CART

1 Introduction

Taxonomists and evolutionists spend most time in understanding the size and shape of the biological forms. The classification and prediction naturally remains in the domain of these specialists. Quantitative methods are developed to bring objectivity in classification and discrimination. But their use has been limited possibly because of two reasons:

1) the time consuming task of morphometric data generation and
2) unfamiliarity of many micropaleontologists with statistical methods.

The unprecedented growth of both hardware and software has, however, made significant impact on morphological data handling.

The term morphometrics in reference to the biological forms denotes the shape and size of the organisms or their anatomical parts. Some authors use it interchangeably with numerical taxonomy even though the two terms are not identical. Taxonomy is among the several applications of morphometrics. A detailed discussion on application of morphological data in numerical taxonomy is given by Sneath and Sokal (1973). Blackith & Reyment (1971), Reyment *et al*. (1984) and Reyment (1991) explain statistical techniques in size and shape analysis of microfossils, particularly foraminifera. A range of multivariate statistical procedures is available to analyze morphological data. Essentially, these procedures involve two stages – classification and discrimination. Classification is done by cluster analysis and there are a number of possible methods to classify a set of objects (Everitt, 1980). The discriminant analysis is a process of discriminating two or more a priori defined groups by a linear combination of two or more variables. It is to be noted that this method assumes that the unknown specimen belongs to one of the populations used in the computation of discriminant function. Most of the multivariate statistical techniques are based on the assumption of multivariate normality of the data. The morphometric data generally deviate from normal distribution and, therefore, it violates the basic assumption of the method used. The reliability of prediction is questionable in such cases. Classification and Regression Tree (CART) is a non parametric method used for classification and prediction problems. This does not require normally distributed data and it is used by medical professionals in classifying patients into clinically important categories. Feldesman (2002) possibly for the first time used this technique for morphometric data of modern hominoids. In spite of several advantages of this statistical procedure it is yet to be used in foraminiferal taxonomy. In this paper we use this method to distinguish three species of *Nummulites* from western India and compare the results with more popular standard method of multigroup discriminant analysis.

2 Materials

The data for the present case study refer to three species of middle Eocene *Nummulites* from Kutch (India). The species are *N. beaumonti*, *N. neglectus* and *N. stamineus* that have close morphological resemblance and therefore debated by earlier workers as to their taxonomic status (Samanta *et al.*, 1990). Saraswati & Patra (2000) used statistical methods to resolve the debate and discriminate the three species (Figure 1). They selected fifteen specimens of *N. stamineus* and twenty specimens each of *N. beaumonti* and *N. neglectus* to measure the diameter and thickness of the tests, length and height of chambers in the first and last whorls and thickness of the marginal cord in oriented equatorial sections. A statistical summary of the morphometric data for the three species is given in Table 1. The same data set in the present study is analyzed by the CART method using XLMiner software (www.xlminer.com). The details of the method are given below.

Species	Stat.	D	T	LI	HI	LL	HL	M
N. beaumonti	Min	3.80	1.24	0.03	0.03	0.26	0.25	0.13
	Max	11.80	3.90	0.05	0.05	0.42	0.33	0.24
	Mean	7.06	2.45	0.03	0.05	0.35	0.29	0.20
	Std Dev	2.14	0.61	0.01	0.01	0.03	0.02	0.02
N. neglectus	Min	11.20	3.42	0.03	0.04	0.42	0.40	0.20
	Max	23.30	7.12	0.06	0.07	0.60	0.57	0.27
	Mean	17.31	5.31	0.04	0.05	0.51	0.50	0.24
	Std Dev	2.49	0.78	0.01	0.01	0.05	0.04	0.02
N. stamineus	Min	8.70	4.42	0.03	0.04	0.25	0.24	0.24
	Max	12.60	5.18	0.06	0.06	0.33	0.30	0.30
	Mean	10.95	4.73	0.04	0.05	0.30	0.26	0.26
	Std Dev	1.06	0.20	0.01	0.01	0.02	0.01	0.02

Table 1 Statistical summary of the morphometric parameters of the three species of *Nummulites*, all measurements in mm (after Saraswati & Patra, 2000).

3 Methodology of CART

CART methodology deals with the classification problem. It is technically known as binary recursive partioning. The process is binary as parent nodes are split in to exactly two daughter nodes and recursive because process can be repeated treating each child node as a parent node. In CART, a binary tree like structure is drawn such that all nodes in the same layer constitute a partition of root node and partition becomes finer as layer gets deeper and deeper. The root node contains the learning sample. The entire construction of a tree involves around three steps:

1) Selection of a splitting criterion at each node.
2) The decision when to declare a node terminal or continue splitting it.
3) Assignment of a class to each terminal node.

Normally splits are performed by putting condition on the coordinates of the measurement vector $\underline{X} = (X_{1,} X_{2,\ldots,} X_n)$. At each variable it finds the best split. Then it compares n best splits and selects the best of the best. To split a node, CART always asks questions that have YES or NO answers, as "Is length ≤ 5.2?" or "Is diameter > 1.8?" Those cases in the node answering YES go to the left descendant node (t_L) and, rest go to right descendant node (t_R). The nature of the split depends upon nature of variables. For each ordered variable x_n, the set of questions includes all questions of the form $\{Is\ x_n \leq c\,?\}$ for all c $\in \mathbb{R}$, and, if x_n is categorical taking values in the set $\{b_1, b_2, \mathrm{L}\ , b_p\}$, say, then the set of questions includes all questions of the form $\{Is\ x_n\ \hat{\mathrm{I}}\ \ S\,?\}$ as S ranges over all subsets of $\{b_1, b_2, \mathrm{L}\ , b_p\}$. It may further be noted that the total number of distinct splits that correspond to all the variables in the data is always finite.

The choice of the best split is based on the impurity function ϕ such that:

1) ϕ is a non-negative function having concave shape,
2) for any p in $(0,1), f(p) = f(1 -\ p)$ and $f(0) = f(1) < f(p)$.

If any given impurity function ϕ, the impurity measure of any node '*t*' is defined as

$$i(t) = f\ \left(p(1/t), p(2/t, \mathrm{L}\ , p(J/t)\right)$$

where $p\left(\frac{j}{t}\right)$ is the probability that a case is in class j given that it falls into *t*.

Most commonly used impurity functions are:

1) Gini's index,

2) entropy function index and they are respectively defined by

$$i(t) = \sum_{i \neq j} p(i/t)p(j/t),$$

and

$$i(t) = \sum_{j} p(j/t)\log[p(j/t)].$$

If a split '*s*' of a node *t* sends a proportion p_k of the data in *t* to t_R and proportion p_L to t_L, then define the decrease in impurity due to split *s* at node *t* as

$$\Delta i(s,t) = i(t) - \left[p_R i(t_R) + p_L i(t_L) \right].$$

This $\Delta i(s,t)$ measures the degree of reduction in impurity by going from parent node to daughter node. Thus among all possible splits at node *t* , split $s*$ is chosen which gives largest decrease in impurity, i.e.,

$$\Delta i(s^*,t) = \max_{s \subset S} \Delta i(s,t)$$

Now to terminate the tree growing, the following method is adopted: Suppose we have done some splitting and arrived at a current set of terminal nodes. The set of splits used, together with the order in which they were used, determines what we call a binary tree T.

Now define tree impurity as

$$I(T) = \sum_{t \in \tilde{T}} I(t) = \sum_{t \in \tilde{T}} i(t).p(t),$$

where $\tilde{T}$ is the current set of terminal nodes.

When any node $t \in T$ is split into t_L, left daughter node, and, t_R, right daughter node, the new tree T' has impurity

$$I(T') = \sum_{T-\{t\}} I(t) + I(t_L) + I(t_R).$$

Hence the decrease in tree impurity,

$$I(T) - I(T') = I(t) - I(t_L) - I(t_R).$$

This depends only on the node t and splits s. Therefore maximizing the decrease in tree impurity by splits on t is equivalent to maximizing the expression

$$\begin{aligned}\Delta I(s,t) &= I(t) - I(t_L) - I(t_R)\\ &= i(t)p(t) - i(t_L)p(t_L) - i(t_R).p(t_R)\\ &= \left[i(t) - i(t_L).p_L - i(t_R).p_R\right].p(t)\\ &= \Delta i(s,t).p(t)\end{aligned}$$

as $$p_L = \frac{p(t_L)}{p(t)}, \quad p_R = \frac{p(t_R)}{p(t)}.$$

Since $\Delta I(s,t)$ differs from $\Delta i(s,t)$ by factor p(t), the same split s* maximizes both expressions. Thus, the split selection procedure can be thought of as a repeated attempt to minimize overall tree impurity. For a threshold $\beta > 0$, a node t is declared terminal node if

$$\max_{s \in S} \Delta I(s,t) < \beta.$$

After getting terminal nodes, class character of a terminal node is determined by the Plurality Rule, namely,

$$if \quad p(j_0 / t) = \max_j p(j/t),$$

then 't' is designated as a class j_0 terminal node. If the maximum is attained for two or more different classes, then the class is assigned arbitrarily.

4 Results and Discussion

In all 111 iterations of CART methodology are performed by splitting the data set of 55 observations into training, validation and test data sets in varying proportions. The findings of this analysis are summarized in Tables 2 to 4. It is evident from Table 2 that the key variables on the basis of which majority of tree models have been built are combinations of T (thickness of the test), HL (height of the chamber in the final whorl) and D (diameter of the test). One such tree model is shown in Figures 2 and 3 for training data set and validation data set respectively. As far as validation data sets are concerned, in about 40% of them all the cases are classified correctly, while 49% contains only one case of misclassification. In regard to test data sets, nearly 70% contains no misclassification cases, whereas in about 25% of the cases only one case of misclassification is found.

Key variables in models	No. of modelsout of 111	Proportions of models out of 111
T, HL	51	49.04
T, D	30	28.85
T, D, HL	10	9.61

Table 2 Model Summary.

No. of cases misclassified	No. of iterationsout of 111	Proportions of iterations out of 111
0	44	39.64
1	54	48.65
Between 2 & 6	13	11.71

Table 3 Summarized results for Validation Samples.

No. of cases misclassified	No. of iterationsout of 111	Proportions of iterations out of 111
0	78	70.27
1	28	25.23
Between 2 & 5	05	4.50

Table 4 Summarized results for Test Samples.

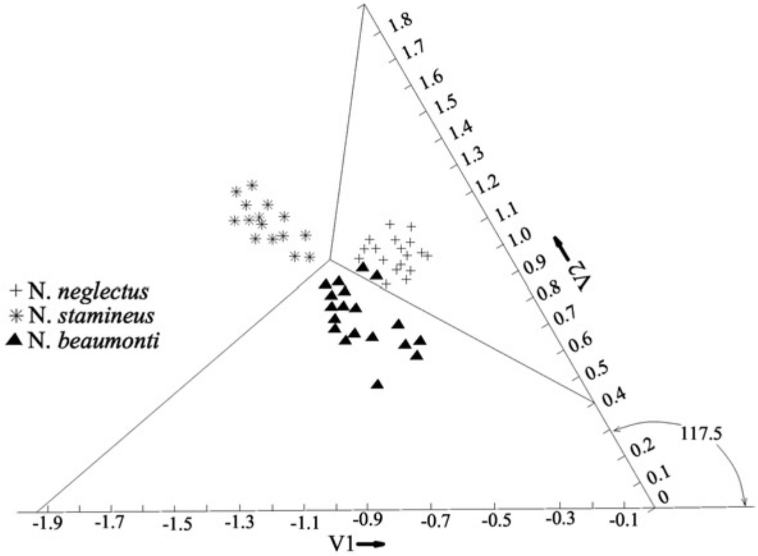

Figure1 Multigroup discriminant plot of *N. beaumonti, N. neglectus* and *N. stamineus* (after Saraswati and Patra, 2000).

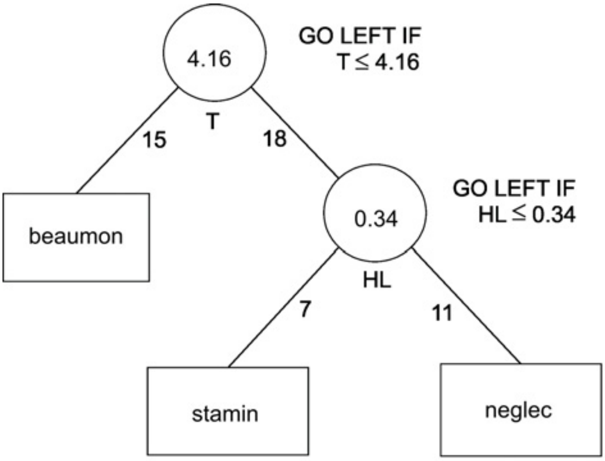

Figure 2 Classification Tree using training data set.

Saraswati & Patra (2000) used multigroup discriminant analysis for the same data set. The data matrix was subjected to log transformation in order to make observations to be multivariate normal and to scale down the strong influence of high magnitude variables. The following two discriminant functions (V_1 and V_2; V_1^V_2 = 117) as defined below discriminate the three species:

V_1 = 0.06 D – 0.14 T + 0.21 LL + 0.97 HL

V_2 = 0.11 D + 0.87 T – 0.39 LL – 0.28 HL

where D and T are diameter and thickness of the test respectively and LL and HL are length and height respectively of the chambers in the last whorl.

In this procedure two specimens of *N. beaumonti* are misclassified as *N. neglectus* (Figure 3). It may be noted that in both the statistical analyses some of the specimens of *N. beaumonti* are misclassified as *N. neglectus.* This is possibly due to closer morphological similarity of *N. neglectus* with *N. beaumonti* than between *N. neglectus* and *N. stamineus.* The observation of Samanta *et al.* (1990) that some of the illustrations of *N. beaumonti* by Sen Gupta (1965) are more akin to *N. neglectus* also supports that the two species have close resemblance.

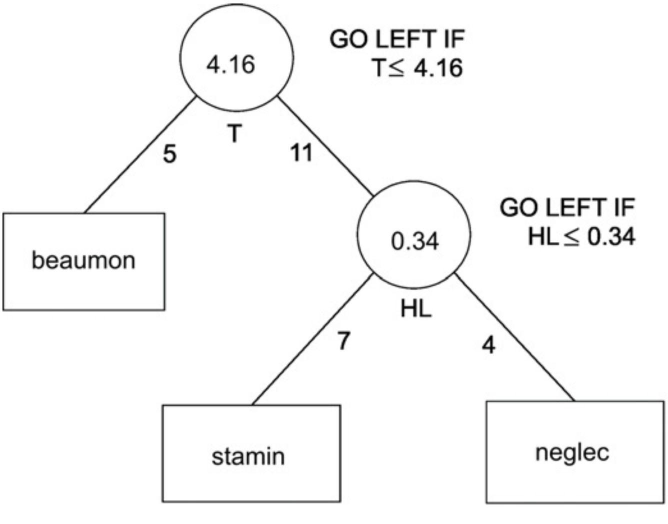

Figure 3 Classification Tree using validation data set.

Our study suggests that the multigroup discriminant analysis and CART are comparable in discriminating the three species. However, CART requires a fewer number of variables for classification and has low misclassification error. The method is very useful for professional geologists for quick identification of species. Once the classification tree for common and age-diagnostic species of a sedimentary basin is constructed, it can be used even by semi-specialists to identify the species. The applications of CART in medicine and anthropology have shown that this technique also handles the missing data very efficiently (Feldesman, 2002). This gives an added advantage to its use in the study of fossil foraminifera where some of the morphological features may be badly preserved or some of the measurements may not be possible in all the oriented sections despite the best efforts of the researcher. In view of these advantages we suggest CART as a better option for class assignment in foraminiferal taxonomy.

5 Acknowledgements

Thanks to Dr Norman McLeod for his invitation to write this paper and to the Director IIT Bombay for providing fund to one of us (PKS) to present it at the International Conference FORAMS 2006.

6 References

Blackith, R.E. & Reyment, R.A. 1971. Multivariate Morphometrics. Academic Press, London. 412 p.

Everitt, B. 1980. Cluster Analysis. 2nd Ed., Halsted Press, New York. 135 p.

Feldesman, M.R. 2002. Classification Tree as an Alternative to Linear Discriminant Analysis. *American Journal of Physical Anthropology, 119:* 257-275.

Reyment, R.A. 1971. Multivariate normality in morphometric analysis. *Mathematical Geology,* 3: 357- 368.

Reyment, R.A. 1991. Multidimensional Palaeobiology. Pergamon Press, Oxford. 377 p.

Reyment, R.A.; Blackith, R.E. & Campbell, N.A. 1984. *Multivariate Morphometrics*. II Ed. Academic Press. 233 p.

Samanta, B.K.; Bandopadhyay, K.P. & Lahiri, A. 1990. The occurrence of *Nummulites* Lamarck (Foraminiferida) in the Middle Eocene Harudi Formation and Fulra Limestone of Cutch, Gujarat, western India. *Bull. Geol. Min. Met. Soc. India, 55*: 1-66.

Saraswati, P.K. & Patra, P.K. 2000. Multivariate analysis of three closely resembling species of *Nummulites* from Middle Eocene of India. *Rev. de Micropal., 43*: 319-326.

Sen Gupta, B.K. 1965. Morphology of some key species of *Nummulites* from the Indian Eocene. *Jour. Paleont., 39*: 86-96.

Sneath, P.H.A. & Sokal, R.R. 1973. *Numerical Taxonomy*. W.H. Freemann and Co., San Francisco, 573 p.