Classification and regression trees to predict the species of an unidentified *Puntius* specimen

Thilan, A.W.L.P.[1], De Silva, M.P.K.S.K.[2] and Jayasekara, L.A.L.W.[1]

[1]*Department of Mathematics, University of Ruhuna, Matara, Sri Lanka.*

[2]*Department of Zoology, University of Ruhuna, Matara, Sri Lanka.*

✉ pubudu@maths.ruh.ac.lk

Abstract

Out of 82 freshwater fish species in Sri Lanka, the Genus *Puntius* represents 16 species (19.5%). However, ambiguities in taxonomic identification of different *Puntius* species remain as a well known research area. Hence, in this study Classification and Regression Trees (CART) and Random Forests analysis were carried out to identify and differentiate among *Puntius* species using their morphometric, meristic and coded variables.

Total of 316 specimens representing eight Sri Lankan *Puntius* species were collected at four different altitude ranges from five major river basins in Sri Lanka. Fifteen meristic characters, four coded variables and twenty three morphometric characters were recorded from each specimen. In the case of combining meristic and coded variables, the correct classification rate for model was 98% and the value of the *Kappa statistic* which is a chance-corrected measure of prediction was 0.982. In random forests analysis, the classification error rates based on the out of bag samples, averaged over many bootstrap samples, provide an unbiased estimate of prediction error for combination of meristic and coded variables was 0.95%. The overall correct classification rate of CART model to predict the species of an unidentified *Puntius* specimen using its morphometric measurements was 80% and the value of the *Kappa statistic* was 0.769. The corresponding unbiased estimate of prediction error from random forests for morphometric data was 14.24%.

In this study eight *Puntius* species were considered and *Number of transverse scales (tr)* and *Total length, (TL)* were the most important meristic and morphometric variable respectively for differentiating among those species.

Keywords: classification and regression trees, prediction, *Puntius* species, taxonomy, random forests

Introduction

Classification and Regression Trees (CART) have an intuitive representation, the resulting model is easy to understand and interpret by humans. The decision trees are nonparametric models, no intervention being required from the user, and thus they are very suited for exploratory knowledge discovery (Breiman *et al.*, 1984; Shinet *et al.*, 1993; Death and Fabricius, 2000; Vayssieres *et al.*, 2000; Hancock *et al.*, 2002). Accuracy of decision trees is comparable or superior to other models (Selker *et al.*, 1995; Smith *et al.*, 1997; Germanson *et al.*, 1998). Random Forest is an alternative approach to classification using classification trees.

Sri Lankan freshwater fish fauna consist of 82 species and out of that 16 species (19.5%) belong to the Genus *Puntius* (Hamilton, 1822). The Genus *Puntius* are members of the Family known as Cyprinidae (Table 1).

Among the 16 species of *Puntius* in Sri Lanka nine are endemic. Many species of *Puntius* are attractive as aquarium fish due to their beautiful coloration, striking body markings, general body shape and small size as well as the ease of rearing in home aquaria. Due to over exploitation, as a result of the aquarium trade and general habitat degradation, some *Puntius* species have become highly threatened and are prone to extinction (Table 1). Conservation of these species has become a critical issue and recognition of Sri Lanka as a global biodiversity hot spot has raised their conservation profile. Effective methods for species identification are required to assist their conservation. Identification of *Puntius* species is based currently on several characters that incorporate external morphology, morphometric and meristic characters (Deraniyagala, 1952; Munro, 1955; Jayaram, 1991; Pethiyagoda, 1991). Some morphological characters are overlapping among species. Some characters vary marginally among species. Some characters (e.g. osteological ones) are also difficult to obtain in a short time period and can damage the specimen. Descriptions of colour patterns and markings on the body may fade or may not be clearly seen in preserved specimens. In museum specimens, sorting of specimens with respect to their species have become troublesome. These problems have led to misidentification and taxonomic ambiguities among *Puntius* species. Therefore present study amid on to develop a methodology to accurately differentiate among *Puntius* species.

Table 1: *Puntius* species of Sri Lanka and their status

Species	Index	Sample size	E	V	En	CE	A
Puntius amphibius	1	-					Common
Puntius asoka	2	-	+			+	Rare
Puntius bandula	3	-	+			+	Very rare
Puntius bimaculatus	4	55	+				Very common
Puntius chola	5	34					Common*
Puntius cumingii	6	-	+	+			Common
Puntius dorsalis	7	54					Common
Puntius singhala	8	-					Common
Puntius martenstyni	9	38	+			+	Rare
Puntius nigrofasciatus	10	27	+	+			Not yet rare
Puntius pleurotaenia	11	37	+	+			Common
Puntius sarana	12	40					Common
Puntius srilankensis	13	-	+		+		Very rare
Puntius ticto	14	-					Common
Puntius titteya	15	-	+	+			Common
Puntius vittatus	16	31					Common

E - endemic, V - Vulnerable, En - Endangered, CE - Critically Endangered, A - Abundance.
* - Uncommon in wet zone

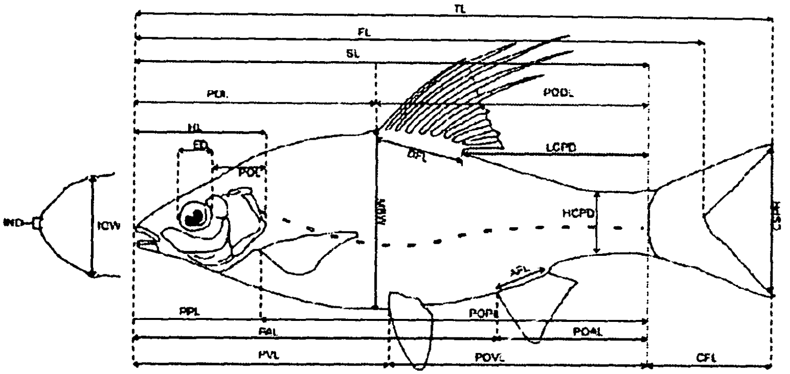
The objective of the current study is to use tree based classification techniques known as *Classification and Regression Trees* and *Random Forests* to accurately predict the species of an unidentified *Puntius* specimen using its morphometric, meristic and coded variables.

Materials and Methods

Data collection

A total of 316 specimens from eight described species of *Puntius* were sampled from five major river basins in Sri Lanka. For this study, only eight *Puntius* species were considered as they are abundantly found in fresh water bodies. Identification of specimens was carried out in the field to species level using external morphological characters (Pethiyagoda, 1991). Additional identification was undertaken in the laboratory using standard fish keys and guides (Deraniyagala, 1952; Munro, 1955; Jayaram, 1991; Pethiyagoda, 1991). A total of forty two (42) characters of which 23 represented morphometric measurements (Figure. 1), four represented coded variables and fifteen represented meristic traits (Tables 2 and 3) were scored from each specimens.

Figure 1: Morphometric characters measured in this study



Total length, TL; Standard length, SL; Fork length, FL; Maximum body depth, MBW; Head length, HL; Eye diameter, ED; Distance between pair of nostrils, IND; Inter orbital distance, IOW; Post orbital length, POL; Dorsal fin length, DFL; Pre dorsal length, PDL; Post dorsal length, PODL; Anal fin length, AFL; Pre anal length, PAL; Post anal length, POAL; Pre ventral length, PVL; Post ventral length, POVL; Pre pelvic length, PPL; Post pelvic length, POPL; Caudal fin length, CFL; Width of the caudal fin when fully spread, CSPR; Caudal peduncle height, HCPD; end of the dorsal fin to end of the caudal peduncle length, LCPD.

Table 2: Coded variables scored on *Puntius* species

Characters	Acronyms
Nature of lateral line** 1) Complete lateral line 2)Incomplete lateral line	nll
Position of mouth ** 1) sub terminal (When the mouth is opened it directs towards downward) 2) terminal (When the mouth is opened it directs front of the head and points forward)	pom
Nature of dorsal fin spines** 1) smooth 2) serrate	ndfs
Number of barbells 0) no barbells or a pair of rudimentary barbels, 1) one pair of barbells 2) two pairs of barbels	nb

**Characters were quantified as 0, 1 and 2 on a nominal scale and this number was used in the analysis

Linear measurements were made using venire calipers to the nearest 0.01 millimeter. A Stereo microscope (Wild M5A) and hand lens were used to determine meristic counts and to score coded variables. Coded characters (Table 2) were converted to a discrete form and also included with meristic characters in the analysis. All morphometric variables were standardized to remove the effect of individual size. In that case, *eye diameter (ED)* and *post orbital length (POL)* were divided by *head length (HL)* and all other variables shown in Figure. 1, were divided by *standard length (SL)* to remove the effect of individual size (Austin and Knott, 1996).

Table 3: Meristic measurements scored from *Puntius* species

Scale counts	
Characters	**Acronyms**
Number of lateral line scales	lls
Number of transverse scales*	tr
Pre dorsal scales-counted from the edge of the operculum to the beginning of the dorsal fin	prds
Post dorsal scales-counted from the end of the dorsal fin to the beginning of the caudal fin	psds
Dorsal fin scales-counted from the beginning of the dorsal fin to the end of the dorsal fin	dfsc
Scales around the caudal peduncle	cped
Fin ray counts	
Characters	**Acronyms**
Number of dorsal fin rays	dfr
Number of anal fin rays	afr
Number of pelvic fin rays	pfr
Number of caudal fin rays	cfr
Number of ventral fin rays	vfr
Fin spine counts	
Characters	**Acronyms**
Number of dorsal fin spines	dfs
Number of anal fin spines	afs
Number of pelvic fin spines	pfs
Number of ventral fin spines	vfs

*Numbers one to nine were used in the analysis. Transverse scales were divided into 09 categories according to the arrangement (1) 3.5/2.5; (2) 3.5/3; (3) 3.5/3.5; (4) 4.5/2.5; (5) 4.5/3; (6) 4.5/3.5; (7)5/3.5; (8) 5.5/2.5; (9) 5.5/3.5

Classification and regression trees

Classification and Regression Tree (CART) analysis that predicts group membership is a non-parametric procedure and it can handle missing data as well (Breiman *et al.*, 1984; Death and Fabricius, 2000; Karels *et al.*, 2004; Saraswati and Sabnis, 2006). In the process of making the classification tree, the parent node split into two child nodes and the process is repeated by treating each child node as a parent node (Breiman *et al.*, 1984; Shin *et al.*, 1993; Karels *et al.*, 2004; Saraswati and Sabnis, 2006). The decision rule at each split is based on a value of a single explanatory variable.

There are a few proposed measures of node impurity (Death and Fabricius, 2000; Karels *et al.*, 2004; Saraswati and Sabnis, 2006) and they indicate the amount of mixing classes among samples contained in the node. Impurity is largest when all classes are equally mixed together and become smallest when the node contains only one class. The *information index* is one such measure and defined as:

$$i(t) = -\sum p(j|t) \ln p(j|t)$$

where $p(j|t)$ = probability that a case is in class j given that it falls into node t; equal to the proportion of cases at node *t* in class *j* if priors are equal to class sizes, but the preferred method seems to be the *Gini index*, defined as:

$$i(t) = 1 - \sum_{j}' p^2(j\,t).$$

At each node *t*, the algorithm selects the split from the set of all possible splits that maximizes the reduction in overall tree impurity (Hancock *et al.*, 2002; Ishwara, 2007; Atabati *et al.*, 2009). Once a node is declared terminal, the observations in that node get classified to the class containing the highest probability of membership in that node; i.e., the class that minimizes the probability of misclassification. In general, CART analysis consists of three steps (Atabati *et al.*, 2009): (i) the maximal-tree building, (ii) the tree pruning, (iii) the optimal tree selection.

Maximal tree building

The tree which consists of all homogeneous child nodes or contains one or a user-defined minimal number of observations is called the maximal tree and the terminal nodes, represent the final groups formed by the tree (Yohannes and Hoddinott, 1999). This maximal tree will usually contain too many leaves and will over fit the learning data set, which will cause poor predictive abilities for new sample (Karels *et al.*, 2004).

The best splitter is defined as the variable that will minimize the impurity I of the two child nodes (Yohannes and Hoddinott, 1999; Ishwara, 2007). The goodness of a split is then defined as the impurity decrease between the parent node and its children:

$$\Delta i(s, t_p) = i_p(t_p) - P_L i(t_L) - P_R i(t_R)$$

where s is a candidate split, P_L and P_R are the fractions of observations of the parent node t_p that go into the child nodes t_L and t_R respectively. The best splitter is the one that will maximize

$$\Delta i(s, t_p).$$

Tree pruning

The selection of a smaller tree, derived from the maximal is then necessary for predictive purposes. The procedure of pruning generates a sequence of smaller trees, obtained by removing successively branches of the maximal tree.

Optimal tree selection

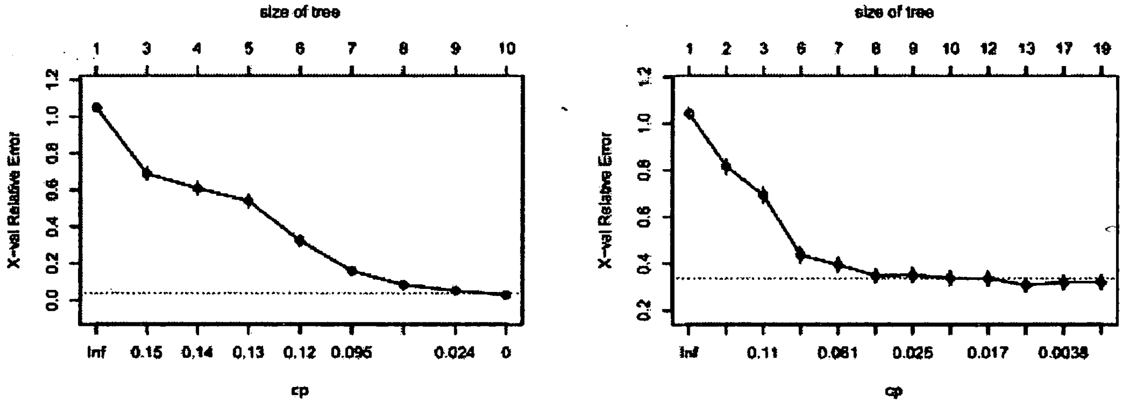
The cross-validation relative error plot (aka, the *cp plot*) provides a plot of the relative error against tree size. One logical choice of tree size is the one that produces the minimum relative error. Another option is to choose a slightly smaller tree, specifically the tree with a relative error that is within 1 standard errors of the minimum relative error. In the *cp plot*, this will be the first point that falls below the dotted line (Death and Fabricius, 2000; Atabati *et al.*, 2009; Varmuza and Filzmoser, 2009). The *cp plot* is based on V-fold cross-validation and it is a random process, so we might think that the results of a single V-fold cross-validation might not always give the best result (Death and Fabricius, 2000; Karels, *et al.*, 2004; Atabati *et al.*, 2009; Siroky, 2009). Therefore, logically we can do the V-fold cross-validation over many times and average the relative errors to get a better estimate of the cross-validation error and thus a smoother *cp plot.*

Random forests

Random Forests offer dramatic improvements in predictive accuracy and stability, but they do not have intuitive trees behind to interpret. Number of clever ways to visualize Random Forests that make them attractive methods not only for prediction but also for data description, model assessment and model improvement have now been suggested and developed (Siroky, 2009). Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (Liaw and Wiener, 2002; Siroky, 2009). In this case, a bootstrap sample with replacement of two-thirds of the observations is taken and a tree is built, but using a random subset of the variables at each node, and then the oob ("out of bag"), or hold-out observations, are submitted to the tree. The classification error rates based on the oob samples, averaged over many bootstrap samples, provides an unbiased estimate of prediction error (Liaw and Wiener, 2002; Prasad *et al.*, 2006; Ishwara, 2007; Siroky, 2009).

Results and Discussion

CART and Random Forest analysis were carried out using R statistical software with two additional libraries known as *rpart* and *randomForest* (Breiman 2001, 2002; Liaw and Wiener, 2002; Breiman and Cutler, 2010). In addition to that *biostats* and *cartware* which are not formal R libraries were also used. In this study, coded variables were considered together with meristic variables. The smoothed *cp plot* in Figure. 2(a) indicates that the optimal tree (Death and Fabricius, 2000; Atabati *et al.*, 2009; Varmuza and Filzmoser, 2009) should have 10 leaves for considered meristic and coded variables.

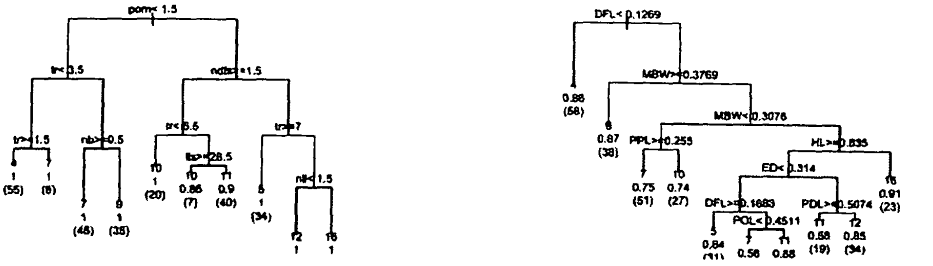


(a) Meristic and coded variables **(b) Morphometric variables**

Figure 2: The cp plot

The expected probability of classification into any group by chance is proportional to the group size. Therefore, with more dissimilar group sizes, a "chance-corrected" measure of prediction is important. The most popular such measure is *Cohen's Kappa statistic* (Cohen, 1960), which is appropriate when prior probabilities are assumed to be equal to sample sizes. The numbers printed below the terminal nodes (leaves) of the classification tree in Figure 3 are interpreted as follows.

It is also interested to know how each of the measured variables, including those that did not make it into the final tree, performs in terms of their ability to distinguish among groups. For this purpose, it is useful to compute a measure of variable importance (Karels *et al.*, 2004; Siroky, 2009). Here, variable importance is determined by calculating for each variable at each node the change in impurity (eg. gini index) attributable to the best surrogate split on that variable (Banerjee *et al.*, 2008). The importance of coded and meristic variables is listed in Table 4 and according to that *Number of transverse scales (tr)* is the most important meristic variable to predict the species of an unidentified *Puntius* specimen.



(a) Meristic and coded variables **(b) Morphometric variables**

Figure 3: Classification tree

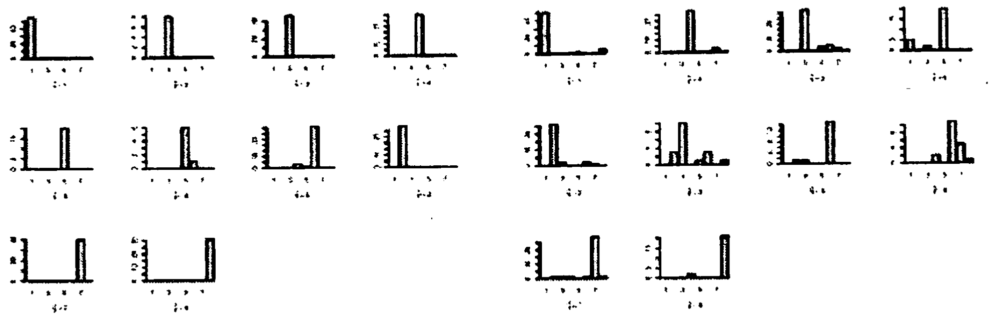
The topmost number is the predicted class of the node. All observations that end up in this node are predicted (or classified) to be the class value listed here. The second number gives the proportion of observations correctly classified in this node. The third number given parenthetically gives the number of observations in the node.

The classification tree in Figure 3(a) consists of meristic and coded variables. In there, the first split is on a coded variable, known as *Position of mouth, (pom).* If *pom* < 1.5, follow the tree to the left (true conditions always go left), otherwise (i.e., the false condition) follow the tree to the right (Hancock *et al.*, 2002; Saraswati and Sabnis, 2006). Following the tree to the left and again if *Number of transverse scales, (tr)* < 3.5 and also if *tr* ≥ 1.5, we find that the prediction is "*Puntius bimaculatus*"(i.e. grouping variable = 4), otherwise prediction is "*Puntius dorsalis*"(i.e. grouping variable = 7) and it is correct for 100%. In this way going along the tree (Rovlias and Kotsou, 2004), we can predict the species of an unidentified *Puntius* specimen using its measured meristic and coded variables. The overall correct classification rate of CART model to predict the species of an unidentified *Puntius* specimen using meristic and coded variables together was 98% and the value of the corresponding *Kappa statistic* was 0.982.

The smoothed *cp plot* is in Figure. 2(b) suggest that 10 leaves tree as the optimal tree (Death and Fabricius, 2000; Atabati *et al.*, 2009; Varmuza and Filzmoser, 2009) for morphometric data. According to the classification tree is in Figure. 3(b) the first split is on a morphometric variable, known as *Dorsal fin length, (DFL).* By going along that tree, we can predict the species of an unidentified *Puntius* specimen using its measured morphometric variables. In that case, the overall correct classification rate of CART model to predict the species of an unidentified *Puntius* specimen was 80% and the value of the corresponding *Kappa statistic* was 0.769.

Once we have settled on a pruned tree, we can produce a bar plot of class membership at each terminal node as shown in Figure. 4(a) and 4(b) respectively for combined (meristic and coded) and morphometric data.

The bar plots give the number of observations in each class in each terminal node. A single bar plot is produced for each terminal node. They represent the distribution of the response for the objects within each node (Atabati *et al.*, 2009). It is also an informative view of the misclassification and clearly indicates the suitability of coded and meristic variables together for the intended prediction than using morphometric variables.



(a) Meristic and coded variables **(b) Morphometric variables**

Figure 4: The distribution of the response for the objects within each node

The importance of the measured morphometric variables is listed in Table 5 and according to that the most important morphometric variable for such a prediction is *Total length, (TL).*

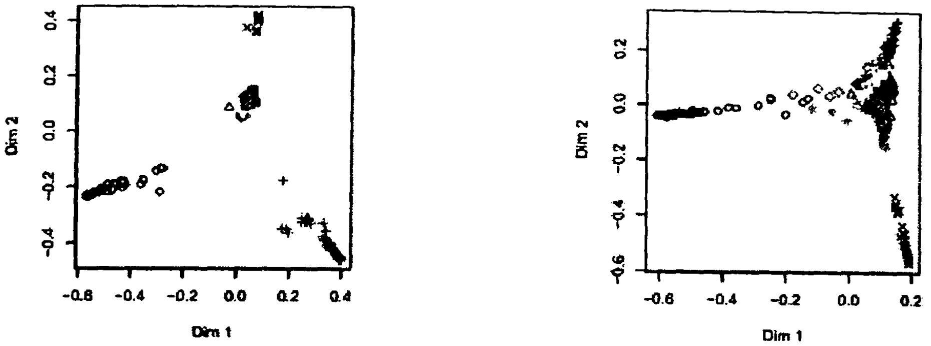
Table 4: Importance of meristic and coded variables

Variable	Importance	Variable	Importance	Variable	Importance
tr	100.00	cepd	37.26	nll	22.69
lls	65.98	dfcs	37.26	dfr	21.25
nb	64.08	psds	35.03	pom	6.11
dfs	45.72	prds	31.04	cfr	4.35
ndfs	44.21	pfr	29.01	vfs	3.01

Table 5: Importance of morphometric variables

Variable	Importance	Variable	Importance	Variable	Importance
TL	100.00	AFL	21.15	FL	14.09
DFL	38.71	MBW	20.83	.	.
SI	29.28	POL	20.28	.	.
HL	22.69	PDL	20.19	HCPD	5.28
PAL	22.08	ED	17.42	IND	2.40

In random forests the unbiased estimate (Liaw and Wiener, 2002; Prasad *et al.*, 2006; Ishwara, 2007) of prediction errors are 0.95% and 14.24% for combined and morphometric data respectively. It clearly indicates the suitability of coded and meristic variables together for the intended prediction than using morphometric variables. The random forest also computes a matrix of proximity measures among the input. The proximity matrix is in fact a similarity matrix, with pair wise similarities between samples defined by how often they end up in the same terminal node in the random forest (Breiman and Cutler, 2010). The proximities give an intrinsic measure of similarities between observations. However, the most useful property of proximities is that form Euclidean distances in a high dimensional space, they can be projected down onto a low dimensional space using principal coordinates analysis as shown in Figure 5(a) and 5(b) for combined and morphometric data respectively. In the case of using meristic and coded variables together, we can see a good separation of species as forming separated clusters for them. This gives an informative view of the data (Breiman, 2002; Liaw and Wiener, 2002) and indicates the suitability of meristic and coded variable together for the intended prediction.



(a) Meristic and coded variables (b) Morphometric variables

Figure 5: Multi-dimensional scaling plot of proximity matrix

Conclusion

The generated trees were evaluated and applied for the prediction of the species of an unidentified *Puntius* specimen. The results have shown that this methodology has good prediction power for this purpose. The application of CART to meristic (with coded variables) and morphometric data have shown that the CART analysis is able to perform a better prediction using meristic (with coded variables) data than morphometric data in terms of prediction accuracy. Moreover, the output of rules sets from the CART analysis can provide useful insight into the relationships between the response and the predictor variables and the relative importance of predictor variables.

In conclusion, CART technique is a visually useful way to predict the species of an unidentified *Puntius* specimen. A tree diagram, illustrating the meristic (with coded variables) or morphometric variables, provides some threshold values that split the specimens into subgroups according to their species. In this study, the Genus *Puntius* was used as a text model and the accuracy of result motivate us to extend the use of this novel approach to predict species of an unidentified specimen belong to any species.

References

Atabati, M., Zarei, K. and Abdinasab, E. (2009). Classification and regression tree analysis for molecular descriptor selection and binding affinities prediction of imida-zobenzodiazepines in quantitative structure-activity relationship studies. *Bulletin of Korean Chemical Society*, 30(11), 2717-2722.

Austin, C.M. and Knott, B. (1996). Systematics of the freshwater crayfish genus *Cherax Erichson* (Decapoda: Parastacidae) in south-western Australia: electrophoretic, morphological and habitat variation. *Australian Journal of Zoology*, 44, 223-258.

Banerjee, A.K., Arora, N. and Murty, U.S.N. (2008). Classification and Regression Tree (CART) Analysis for Deriving Variable Importance of Parameters Influencing Average Flexibility of CaMK Kinase Family, *Electronic Journal of Biology*, 4(1), 27-33.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and Regression Trees*. Chapman and Hall, New York.

Breiman, L. (2001). Random Forests, *Machine Learning*, 45(1), 5-32.

Breiman, L. (2002). Manual on Setting Up, Using and Understanding Random Forests V3.1.

Breiman, L. and Cutler, A. (2010). Breiman and Cutler's random forests for classification and regression, Version 4.5-35.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Deraniyagala, P.E.P. (1952). *A coloured atlas of some vertebrates from Ceylon. Volume 1: Fishes*. The Ceylon Government Press, Ceylon.150 pp. 34pls.

Germanson, T.P., Lanzino, G. and Kongable, G.L. (1998). Risk classification after aneurysmal sub-arachnoid hemorrhage, *Surgical Neurology*, 49: 155-163.

Hamilton, F. (1822). *An account of the fishes found in river Ganges and its branches*, Archibald Constable, Edinburgh; Hurst, Robinson and Co., London. 40 pp., 39 pls,

Hancock, T.P., Coomans D.H. and Everingham, Y.L. (2002). Supervised Hierarchical Clustering Using CART, Department of Mathematics and Statistics, James Cook Univer-sity, Townsville, Queensland, Australia, 4811.

Ishwara, H. (2007). Variable importance in binary regression trees and forests, *Electronic Journal of statistics*, 1:519-537.

Jayaram, K.C. (1991). *Revision of the genus Puntius Hamilton from the Indian region (Pisces: Cypriniformes, Cyprinidae, Cyprininae)*. Zoological Survey of India, Calcutta, 178 p.

Karels, T.J., Bryant, A.A. and Hik, D.S. (2004). Comparision of discriminant function and classification tree analyses for age classification of marmots, *Oikos*, 105:575-587.

Liaw, A. and Wiener, M. (2002). Classification and regression by random forest, *R News*, 2/3, 18-22.

Moreno, P.J., Logan, B. and Raj, B. (2001). *A boosting approach for confidence scoring*, Cambridge Research Laboratory, Compaq Computer Corporation, Cambridge MA 02142-1612.

Munro, I.S.R. (1955). *The marine and freshwater fishes of Ceylon*. Department of External Affairs, Canbara, Australia. 349 p.

Pethiyagoda, R. (1991). *Freshwater fishes of Sri Lanka*. Wildlife Heritage Trust of Sri Lanka, Colombo. 362 p.

Prasad, A.M., Iverson, L.R. and Liaw, A. (2006). Newer classification and regression tree techniques: bagging and random forests for ecological prediction, *Ecosystems*, 9, 181-199.

Rovlias, A. and Kotsou, S. (2004). Classification and regression tree for prediction of outcome after severe head injury using simple clinical and laboratory variables, *Jouranal of Neurotrauma*, 21, 886-893.

Saraswati, P.K. and Sabnis, S.V. (2006). Comparison of CART and discriminant analysis of morphometric data in foraminiferal taxonomy, *Anurio do Instituto de Geocincias*, 29(1), 153-162.

Selker, H.P., Griffith, J.L., Patil, S., Long, W.J. and D'Agostino, R.B. (1995). A comparison of performance of mathematical predictive methods for medical diagnosis: identifying acute cardiac ischemia among emergency department patients. *Journal of Investigative Medicine*, 43, 468-476.

Shin, W.H., Lee, B.S, Lee, Y.K. and Lee, J.S. (1993). Speech/ non-speech classification using multiple features for robust endpoint detection, Information Technology Lab, LG Corporate Institute of Technology.

Siroky, D.S. (2009). Navigating random forests and related advances in algorithmic modeling, *Statistics Surveys*, 3, 147-163.

Smith, S.J., Iverson, S.J. and Bowen, W.D. (1997). Fatty acid signatures and classification trees: new tools for investigating the foraging ecology of seals. *Canadian Journal Fisheries and Aquatic Sciences*, 54, 1377-1386.

Varmuza, K. and Filzmoser, P. (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*: Taylor and Francis Group, LLC.

Vayssieres, M.P., Plant, R.E. and Allen-Diaz, B.H. (2000). Classification trees: An alternative non-parametric approach for predicting species distributions. *Journal of Vegetation Science*, 11: 679-694.

Yohannes, Y. and Hoddinott, J. (1999). *Classification and Regression Trees: An introduction*. International Food Policy Research Institute.