A Computational Framework for Taxonomic Research: Diagnosing Body Shape within Fish Species Complexes

Yixin Chen[1], Henry L. Bart, Jr.[2], Shuqing Huang[3], and Huimin Chen[4]
[1]Department of Computer Science, [4]Department of Electrical Engineering
[1,4]University of New Orleans, New Orleans, LA 70148
[2]Tulane University Museum of Natural History, Belle Chasse, LA 70037
[3]Department of Electrical Engineering and Computer Science
[2,3]Tulane University, New Orleans, LA 70118
[1]yixin@cs.uno.edu, [2]hank@museum.tulane.edu, [3]shuang4@tulane.edu, [4]hchen2@uno.edu

Abstract

It is estimated that ninety percent of the world's species have yet to be discovered and described. The main reason for the slow pace of new species description is that the science of taxonomy, as traditionally practiced, can be very laborious. To formally describe a new species, taxonomists have to manually gather and analyze data from large numbers of specimens, often from broad geographic areas, and identify the smallest subset of external body characters that uniquely diagnoses the new species as distinct from all its known relatives. In this paper, we use an automated feature selection and classification approach to address the taxonomic impediment in new species discovery. The proposed computational framework can identify body shape characters that unite populations within species, as well as distinguishing among species. It also provides statistical "clues" for assisting taxonomists to identify new species or subspecies.

1. Introduction

Approximately 1.4 million species are known to science. However, estimates based on the rate of new species discovery place the total number of species on earth about 10–30 times of this number. Human population expansion and habitat destruction are causing extinctions of both known and yet to be discovered species. The accelerated pace of species decline has fueled the current *biodiversity crisis* [4], in which it is feared that many of the earth's species will be lost before they can be discovered and described. The job of discovering and describing new species falls on the taxonomists. The science of taxonomy has been suffering from dwindling numbers of experts over the past few decades [5]. Moreover, the pace of taxonomic research, as traditionally practiced, is very slow.

A family of software tools has been designed in recent years for gathering and analyzing data on shape variation from images of specimens [6, 8]. These software tools, referred to collectively as *geometric morphometrics* software, use homologous landmarks (points that are arguably related by evolutionary descent) along the body (Figure 1). The software allows superposition and alignment of landmarks from different specimens, adjustment for body size differences among specimens, and multivariate statistical analysis of derived shape variables. However, these analysis can only help taxonomists recognize overall shape differences. They may identify groups of specimens as distinct, but the derived variables are difficult to interpret in terms of particular aspects of body shape. Thus, geometric morphometric techniques cannot be used to uniquely diagnose one group of specimens as distinct from others.

We propose a computational framework for analyzing variation in body shape within species complexes (groups of closely related species or subspecies) in the sucker genus *Carpiodes*. The proposed approach automatically identifies an "optimal" set of body characters that unites populations within species, as well as distinguishing among species. It can provide statistical "clues" to species diagnosis, allowing taxonomists to recognize populations that would have been misdiagnosed by overall shape based techniques.

2. A Computational Framework

We start with a discussion of a taxonomic problem involving suckers of genus *Carpiodes*. However, our approach can be applied to a class of taxonomic problems. The genus *Carpiodes*, as currently recognized, comprises three widely distributed species: the river carp-sucker *Car-*

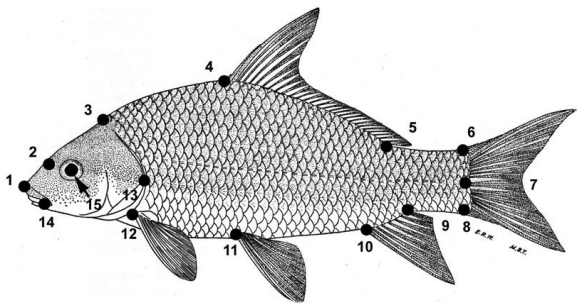

Figure 1. Digitized 15 homologous landmarks using TpsDIG Version 1.4 (© 2004 by F. James Rohlf).

piodes carpio (*C. carpio*); the quillback *Carpiodes cyprinus* (*C. cyprinus*), and the highfin carp-sucker *Carpiodes velifer* (*C. velifer*). Most taxonomists regard each of these species as complexes of multiple species in need of revision [7]. The goal of *taxonomic revision* in this case is to identify and formally describe the unrecognized species.

Over the years since [7] was published, Henry L. Bart has examined shape and DNA sequence variation in all *Carpiodes* populations. An analyses of overall body shape based on a geometric morphometric technique grouped specimens from the Rio Grande, upper Colorado River, and other western Gulf Slope rivers with *C. carpio* specimens from the Mississippi River Basin. However, a surprising finding from the DNA sequence analysis was that the forms in Rio Grande and upper Colorado River system of Texas do not agree at all with *C. carpio*. Rather, they are closely related to *C. cyprinus*, which was not known to occur on the western Gulf Slope. Careful inspection of the specimens in the Rio Grande and upper Colorado River system reveals that they lack the protuberance ("nipple") on the lower lip, which is diagnostic of *C. carpio* and *C. velifer*. They also have a relatively large head and a long snout, characters seen only in *C. cyprinus*. However, specimens from these populations also have an elongate and slender body, and it is these characters that cause them to be erroneously classified as *C. carpio* based on overall body shape analysis.

It took Henry L. Bart three years of careful study of over 1000 *Carpiodes* specimens to determine that Rio Grande and upper Colorado River populations were misdiagnosed as *C. carpio*, and instead represented a new species related to *C. cyprinus*. The question we attempt to address next is: Can machine learning methods be applied to body shape analysis in a way that diagnoses taxonomic groups in genus *Carpiodes* more quickly and accurately?

In general, we are interested in the following taxonomic problem: given a collection of labeled specimens represented in a feature space, identify features and construct classifiers based on the selected features to distinguish among the known categories (or species). Without loss of generality, we focus on the digitized images of specimens with landmarks specified as in Figure 1. Let $LM_j \in \mathbb{R}^2$, $j = 1, \cdots, 15$, be the coordinates of landmarks on a specimen. The feature vector of the i-th specimen, $\mathbf{x}_i$, is given by $\mathbf{x}_i = \phi(LM_1, \cdots, LM_{15}) \in \mathbb{R}^d$ where the mapping $\phi(\cdot)$ transforms the coordinates of landmarks to shape characters, i.e., each feature corresponds to a particular shape character. These features (i.e., $\phi(\cdot)$) are specified by a domain expert. The classification of $\mathbf{x}_i$ is clearly a multi-class problem. We propose to use a tree structure to organize binary classifiers into a multi-class classifier. In this paper, a specimen is first classified as *C. velifer* against the other two species. If the specimen is not from *C. velifer*, it is further classified as *C. carpio* or *C. cyprinus*.

For a given collection of samples $\mathbf{x}_i$ with the corresponding labels $y_i \in \{-1, 1\}$, designing a binary classifier can be solved by any conventional supervised learning algorithm. However, the above mapping usually produces a large number of redundant or irrelevant features because the shape of specimens from closely related species may differ in only a very small number of characters. Feature subset selection is a well-researched topic in the areas of statistics, machine learning, and pattern recognition.The proposed approach is based on a joint feature selection and classification framework, which is a challenging model selection problem in general. So we provide two different wrapper models [1, 3] to see if they are consistent in selecting the useful features.

1. Logistic Regression Classifier with False Discovery Rate Controlled Feature Selection
   We consider the binary classification with the following model: $P(y_i = 1|\mathbf{x}_i, \mathbf{w}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}_i}}$ where the unknown parameter $\mathbf{w}$ needs to be estimated based on the maximum likelihood criterion. The feature subset selection problem can be viewed as multiple hypothesis testing and cast into the following formulation. Assume that the dimension of $\mathbf{w}$ is d. A hypothesis H_k describes the index set $\mathcal{I}_k \subseteq \{1, \cdots, d\}$ of the non-zero components of $\mathbf{w}$, i.e., the selected feature subset. Formally, we can write

   H_k: $w_i \neq 0$ if $i \in \mathcal{I}_k$, otherwise $w_i = 0,\ i = 1, \cdots, d$.

   We apply the FDR technique to estimate $\mathcal{I}_k$. The procedure starts with the test statistic $T_1, \cdots, T_d$ based on the element-wise estimate $\hat{w}_1, \cdots, \hat{w}_d$. Each test statistic T_i is associated with a p-value, π_i, indicating its credibility when $w_i = 0$. For any user specified level $q \in (0, 1)$, the FDR is controlled by performing the following steps.

   - Order the p-values such that $\pi_{(1)} \leq \cdots \leq \pi_{(d)}$.
   - Compute the index $k = \max\left\{i | \pi_{(i)} \leq \frac{i}{d}q\right\}$.
   - Reject all hypotheses $w_{(j)} = 0$ for $1 \leq j \leq k$ where $w_{(j)}$ corresponds to the ordered p-value $\pi_{(j)}$. If no such k exists, then $\mathbf{w} = 0$.

Table 1. Features describing shape characters. Non-shape related variation has been removed from LM_i, the landmark coordinates.

x_1	The distance between the tip of the snout and the naris, computed as the distance between LM_1 and LM_2.
x_2	The slope of the line connecting the tip of the snout and the naris, computed as the angle between the vertical axis and the line connecting LM_1 and LM_2.
x_3	The distance between the naris and the back of the mouth, computed as the distance between LM_2 and LM_{14}.
x_4	The slope of the line connecting the naris and the back of the mouth, computed as the angle between the vertical axis and the line connecting LM_2 and LM_{14}.
x_5	The size of head in proportion of the size of the body, computed as the area of the head polygon (vertices defined in sequence by LM_1, LM_2, LM_3, LM_{13}, LM_{12}, and LM_{14}) divided by the area of the body polygon (vertices defined in sequence by LM_3, LM_4, LM_5, LM_6, LM_7, LM_8, LM_9, LM_{10}, LM_{11}, LM_{12}, and LM_{13})
x_6	The length of the head in proportion of the length of the body, computed as the distance between LM_1 and LM_{13} divided by the distance between LM_{13} and LM_7.
x_7	The distance between LM_7 and LM_8.
x_8	The sum of the distance between LM_3 and LM_{13}, the distance between LM_{12} and LM_{13}, and the distance between LM_1 and LM_{13} divided by the distance between LM_{13} and LM_7.
x_9	The distance between the naris and the tip of the snout in proportion to the distance between the naris and the eye, computed as the distance between LM_1 and LM_2 divided by the distance between LM_2 and LM_{15}
x_{10}	The distance between LM_4 and LM_{11} divided by the distance between LM_{13} and LM_7.
x_{11}	The distance between LM_3 and LM_4 divided by the distance between LM_{13} and LM_7.
x_{12}	The angle between the vertical axis and the line connecting LM_{10} and LM_5.

Once the subset $\hat{\mathcal{I}}_k$ is determined, the logistic regression should be recomputed using only the selected input features.

2. 1-norm Support Vector Machines
   Joint feature subset selection and classification can also be formulated as a regularized optimization problem. Consider the problem of finding a linear classifier $y = \text{sign}\left(\mathbf{w}^T\mathbf{x} + b\right)$ where $\mathbf{w}$ and b are model parameters. The SVM approach constructs classifiers based on hyperplanes by minimizing a regularized training error $\lambda R[\cdot] + error$ where $R[\cdot]$ is a regularization operator, λ is called the regularization parameter, and $error$ is commonly defined through a hinge loss function $\xi = \max\{1 - y(\mathbf{w}^T\mathbf{x} + b), 0\}$. When an optimal solution $\mathbf{w}$ is obtained, the magnitude of its component w_k indicates the significance of the effect of the k-th feature on the classifier. Those features corresponding to a non-zero w_k are selected and used in the classifier.

   The regularization operator in standard SVMs is the squared 2-norm of the weight vector $\mathbf{w}$, which formulates SVMs as quadratic programs (QP). Solving QPs is typically computationally more expensive than solving linear programs (LPs). SVMs can be transformed into LPs as in [3]. This is achieved by regularizing with a sparse-favoring norm, i.e., the 1-norm of $\mathbf{w}$.

3. Experimental Results

The specimens were collected from Tulane Museum of Natural History (128 *C. carpio*, 297 *C. cyprinus*, and 172 *C. velifer*). There are 53 specimens that were collected from the upper Colorado River in Texas and Rio Grande. They were traditionally recognized as *C. carpio*, yet recent DNA evidence suggests a contradictory conclusion. So we view these 53 specimens as "suspicious" populations. We computed 12 features, $x_1, \cdots, x_{12}$, for each specimen using the 15 landmarks. The description of each feature is given in Table 1. We also intentionally introduced 9 randomly generated features, x_{13}–x_{21}, to test the efficacy of the feature selection algorithm.

Table 2. Separating *C. velifer* from *C. carpio* and *C. cyprinus*. The last column shows the percentage of specimens from Colorado River in Texas and Rio Grande classified as *C. velifer*.

Algorithm	Features	Training Error	Test
LRC-FDR	x_9, x_{10}	5.5%	0%
1-norm SVM	x_{10}, x_{11}	11.7%	0%

The 53 "suspicious" specimens are test data with the remaining specimens as training data. The classification results for *C. velifer* against *C. carpio* and *C. cyprinus* are summarized in Table 2. LRC-FDR selected two features, x_9 and x_{10}, with training error 5.5%. 1-norm SVM chose x_{10} and x_{11}, with training error 11.7%. We also implemented four popularly used classifiers, namely, linear discriminant analysis (LDA), Bayesian classifier with Gaussian mixture model (Bayes), SVM with linear kernel, and SVM with Gaussian kernel. The results are shown in Table 3. It is interesting to observe that the test results for all the designed classifiers are quite consistent: none of the 53 suspicious specimens was classified as *C. velifer*. This suggests that the specimens from Colorado River in Texas and Rio Grande do not belong to *C. velifer*.

Next, we designed classifiers to further distinguish *C. carpio* and *C. cyprinus*. The results are given in Table 4. LRC-FDR selected x_7 and x_2, while 1-norm SVM chose

Table 3. Separating *C. velifer* from *C. carpio* and *C. cyprinus* using selected features. The last column shows the percentage of specimens from Colorado River in Texas and Rio Grande classified as *C. velifer*.

Algorithm	Features	Training Error	Test
LDA	x_9, x_{10}	7.4%	0%
	x_{10}, x_{11}	8.5%	0%
Bayes	x_9, x_{10}	5.1%	0%
	x_{10}, x_{11}	7.4%	0%
SVM: linear kernel	x_9, x_{10}	6.2%	0%
	x_{10}, x_{11}	8.7%	0%
SVM: Gaussian kernel	x_9, x_{10}	6.2%	0%
	x_{10}, x_{11}	8.2%	0%

Table 4. Separating *C. carpio* from *C. cyprinus*. The last column shows the percentage of specimens from Colorado River in Texas and Rio Grande classified as *C. carpio*.

Algorithm	Features	Training Error	Test
LRC-FDR	x_2, x_7	8.0%	77.4%
1-norm SVM	x_4, x_7	14.6%	43.4%

Table 5. Separating *C. carpio* from *C. cyprinus* using the selected features. The last column shows the percentage of specimens from Colorado River in Texas and Rio Grande classified as *C. carpio*.

Algorithm	Features	Training Error	Test
LDA	x_2, x_7	11.5%	81.1%
	x_4, x_7	8.9%	54.7%
Bayes	x_2, x_7	9.2%	56.6%
	x_4, x_7	9.2%	35.8%
SVM: linear kernel	x_2, x_7	9.2%	35.8%
	x_4, x_7	7.1%	56.6%
SVM: Gaussian kernel	x_2, x_7	8.5%	32.1%
	x_4, x_7	7.8%	54.7%

x_4 and x_7. Using the selected features, four other classifiers were constructed. The results are listed in Table 5. The test results in Tables 4–5 demonstrate significant variation: the percentage of the suspicious specimens classified as *C. carpio* varies from 32.1% to 81.1% for different classifiers. Although *C. carpio* can be distinguished, with a reasonable accuracy, from *C. cyprinus* with the selected features, it is difficult to categorize the specimens from Colorado River in Texas and Rio Grande to either *C. carpio* or *C. cyprinus* using the same classifiers. This can be viewed as an indication of a possible *new species*. It is very interesting that this clue reveals what has been obtained from the contradictory conclusions between the overall shape analysis and the DNA analysis.

We did extensive experiments to see whether feature selection is indispensable. The results for separating *C. velifer* from *C. carpio* and *C. cyprinus* using all 21 features were very similar to those listed in Table 3, except that the training errors were smaller. However, the classification of *C. carpio* and *C. cyprinus* using all 21 features generated significantly different test results. All four classifiers identify the majority of specimens from Colorado River in Texas and Rio Grande as *C. carpio*, which contradicts the conclusion drawn from Table 5. A similar observation was obtained even if the 9 random features were removed.

An interesting question arises: which results should we trust? We argue that feature selection is indispensable for the following reasons. From a taxonomic viewpoint, it is desirable to use a small number of body shape characters to describe a species as distinct from its known relatives. The feature selection procedure can identify those "most" diagnostic features. From a machine learning viewpoint, constraining the number of selected features is an effective way to avoid overfitting.

4. Acknowledgments

This work was supported in part by grants from the Louisiana Board of Regents (LBOR0077NR00C to YC), US National Science Foundation (DEB-0237013 to HLB), and The Research Institute for Children.

References

[1] F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone, "Adapting to Unknown Sparsity by Controlling the False Discovery Rate," *Annals of Statistics*, 2005, to appear.

[2] D. C. Adams, F. J. Rohlf, and D. E. Slice, "Geometric Morphometrics: Ten Years of Progress Following the 'Revolution'," *Ital. J. Zool.*, 71:5-16, 2004.

[3] J. Bi, K. P. Bennett, M. Embrechts, C. Breneman, and M. Song, "Dimensionality Reduction via Sparse Support Vector Machines," *Journal of Machine Learning Research*, 3:1229-1243, 2003.

[4] S. L. Pimm and J. H. Lawton, "Ecology–Planning for Biodiversity," *Science*, 279:2068-2069, 1998.

[5] J. E. Rodman and J. H. Cody, "The Taxonomic Impediment Overcome: NSF's Partnerships for Enhancing Expertise in Taxonomy (PEET) as a Model," *Systematic Biology*, 52:428-435, 2003.

[6] F. J. Rohlf and F. L. Bookstein, *Proceedings of the Michigan Morphometrics Workshop*, Special Publ. No. 2, The University of Michigan Museum of Zoology, 1990.

[7] R. D. Suttkus and H. L. Bart, Jr., "A Preliminary Analysis of the River Carpsucker, Carpiodes Carpio, in the Southern Portion of its Range," *In: L. Lozano (ed.) Libro jubilar en honor al Dr. Salvador Contreras Balderas*, Universidad Autonoma de Nuevo Leon, Monterrey Mexico, pp. 209-221, 2002.

[8] M. Zelditch, D. Swiderski, D. Sheets, and W. Fink, *Geometric Morphometrics for Biologists: a Primer*, Elsevier Academic Press: London, 2004.