Pattern Anal Applic (2010) 13:23–34
DOI 10.1007/s10044-009-0157-y

THEORETICAL ADVANCES

Joint feature selection and classification for taxonomic problems within fish species complexes

Yixin Chen · Shuqing Huang · Huimin Chen ·
Henry L. Bart Jr

Received: 19 January 2007 / Accepted: 16 May 2008 / Published online: 15 April 2009

Abstract It is estimated that 90% of the world's species are yet to be discovered and described. The main reason for the slow pace of new species description is that the science of taxonomy can be very laborious. To formally describe a new species, taxonomists have to manually gather and analyze data from large numbers of specimens and identify the smallest subset of external body characters that uniquely diagnose the new species as distinct from all its known relatives. In this paper, we present an automated feature selection and classification scheme using logistic regression with controlled false discovery rate to address the taxonomic research need impediment in new species discovery. Unlike traditional taxonomic practice, our scheme automatically selects body shape features from specimen samples with landmarks that unite populations within species, as well as distinguishing among species. It also provides probabilistic assessment of the classification accuracy using the selected features in identifying new species. We apply the scheme to a taxonomic problem involving species of suckers in the genus *Carpiodes*. The results confirm the necessity of feature selection for classifier design and provide additional insight on the suspicious specimens which have traditionally been misdiagnosed as *C. carpio* but are in fact more close to *C. cyprinus*. We also compare the classification accuracy of our scheme with several well-known machine learning algorithms without and with feature selection.

Keywords Feature selection · False discovery rate · Logistic regression · Taxonomy · Systematics

Y. Chen (✉)
Department of Computer and Information Science,
University of Mississippi, University, MS 38677, USA
e-mail: ychen@cs.olemiss.edu

S. Huang
General Dynamics, 1201 Elmwood Park Blvd., New Orleans,
LA 70123, USA
e-mail: Sophie.Huang@gdit.com

H. Chen
Department of Electrical Engineering,
University of New Orleans, New Orleans, LA 70148, USA
e-mail: hchen2@uno.edu

H. L. Bart Jr
Department of Ecology and Evolutionary Biology,
Tulane University, New Orleans, LA 70118, USA
e-mail: hank@museum.tulane.edu

H. L. Bart Jr
Tulane University Museum of Natural History, Belle Chasse,
LA 70037, USA

1 Originality and contributions

In describing new species of animals, taxonomists traditionally gathered data on numerous morphological characters from a large number of specimens to find the smallest subset of indicators diagnostic of the true taxonomic groups. In this paper, we proposed a computational framework for automatic identification of "representative" body shape characters using a joint feature selection and classification approach. We applied logistic regression classifier with feature selection by controlling the false discovery rate to a taxonomic problem in genus *Carpiodes*. The results look promising: the proposed method not only learned the classifier that well separated the three known species in *Carpiodes* using only a few body shape features, but also recognized "suspicious" specimens that could not be identified previously without the aid of DNA analysis. We also compared our results with those using other