Fish Classification Based on Robust Features Selection Using Machine Learning Techniques

Than Thida Hnin and Khin Thidar Lynn

Abstract The taxonomic identification of fishes is a time-consuming process and making errors is indispensable for those who are not specialists. This system proposes an automated species identification system to identify taxonomic characters of species based on specimens. It also provides statistical clues for assisting taxonomists to identify accurate species or review misdiagnosed species. For this system, feature selection is an essential step to effectively reduce data dimensionality. By using combination theory, this system creates the set of attribute pairs to construct the training dataset. And then each attribute pair in training dataset is tested by using two classifiers. Based on the accuracy result of each classifier on attribute pairs and the majority voting of each feature in these attribute pairs, this system selects the most relevant feature set. Finally, this system applied three supervised classifiers to verify the effectiveness of selected features subset.

Keywords Combination theory · Taxonomy · Identification · Fishes

1 Introduction

The pace of new species discovery and description would speed up significantly cause of multimedia and machine learning techniques could be developed. To automatically identify diagnostic features of specimens, simply choose between two alternatives at each step based on the presence or absence of a particular feature, the number of scales or the range of ratios between body measurements.

Recently, taxonomists have been searching for more efficient methods to meet species identification requirements, such as developing digital image processing and pattern recognition techniques. Researchers already used the pattern recognition techniques for insects, plants, spiders, and plankton. These approaches can be extended for fish. Although automated species identification might be a good option to

T.T. Hnin(✉) · K.T. Lynn
University of Computer Studies, Mandalay, Myanmar (Burma)
e-mail: {thanthidahnin,lynnthidar}@gmail.com

T.T. Zin et al. (eds.), *Genetic and Evolutionary Computing*,
Advances in Intelligent Systems and Computing 387,
DOI: 10.1007/978-3-319-23204-1_24

the burden of routine fish taxonomic identification, there is not an automated taxonomic identification system for fishes based on specimen. In fact, automated species identification based on specimen has not become widely employed in any discipline of the biology. One of the explanations for why automated identifications have not become the norm for routine identifications is that such an approach is too difficult. The aim of this study is to determine whether morphometric variation among fish species allows automated taxonomic identification of the species. The key idea is to use the efficient machine learning algorithms for developing the fish identification system, rather than the ones used in traditional automated species identification systems. Machine learning algorithms are popular tools for classifying observations. These algorithms can attain high classification accuracy for datasets from a wide variety of applications and with complex behavior.

The success of applying machine learning methods to real-world problems depends on many factors. One such factor is the quality of available data. The more the collected data contain irrelevant or redundant information, or contain noisy and unreliable information, the more difficult for any machine learning algorithm to discover or obtain acceptable and practicable results. Feature subset selection is the process of identifying and removing as much of the irrelevant and redundant information as possible. Regardless of whether a learner attempts to select features itself, or ignores the issue, feature selection prior to learning has obvious merits.

Feature selection refers to the problem of selecting features that are relevant to predicting a target value for each instance in a dataset. Feature selection has several potential benefits: defying the curse of dimensionality to enhance the prediction performance, reducing measurement and storage requirements and reducing training and prediction times. This paper focuses on an attempt to maximize the performance of a classifier on unknown data and to recast the problem of feature selection in the context of taxonomic fish identification. To achieve efficient species identification, the proposed system can contribute in developing a system utilizing efficient features selection and classification techniques and provide automated fish identification system for Myanmar.

2 Related Work

Ecological interactions of fish assemblages in the pelagic environment can be partially determined by their larval distributions and recruitment to adult populations. The identification fish is essential for current studies on the distribution and reproductive strategies of pelagic fishes [2]. Thus, the assessment of biodiversity and its implication in the management of vulnerable marine ecosystems requires an accurate taxonomic identification of fishes. High levels of global biodiversity and a limited number of taxonomists represent significant challenges to the future of biological study and conservation. The main problem is that almost all taxonomic information exists in languages and formats not easily understood or shared without a high level of specialized knowledge and vocabularies. Thus, taxonomic knowledge is localized within limited geographical areas and among a limited number of taxonomists. This lack of accessibility of taxonomic knowledge to the general public has been