Comparative Study of Attribute Selection for Morphological Identification of Fishes

Than Thida Hnin, Khin Thidar Lynn
University of Computer Studies, Mandalay
thanthidahnin@gmail.com, lynnthidar@gmail.com

Abstract

Taxonomy is the science of naming, describing and classifying organisms that includes all plants, animals and microorganisms of the world. Using morphological, behavioral, genetic information and biochemical observations, taxonomists identify and describe species into classification. 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 specimen and provide statistical clues for assisting taxonomists to identify accurate species or revision of misdiagnosed species. For this system, feature selection is an essential step to effectively reduce data dimensionality. This system first selects the best relevant features by using combination and the classification performance of two classifiers, Random Forest and Attributed Selected. And then correctly classifies the fish species and compares the accuracy of these two classifiers.

1. Introduction

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

Although automated species identification might be a good option to 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. An automated species identification system is a matter of a one-to many matching, which not only needs to match an individual specimen with one of a set of extremely similar species to one another, but also is necessary to be able to reject it as belonging to a species that is not part of this set. Accepting these difficulties, the aim of this study was to determine whether morphometric variation among fish species allows automated taxonomic identification of the species.

As the advances of efficient machine learning and data mining algorithms, the idea is to use different approaches for developing the fish identification system, rather than the ones used in previous 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. In addition, through automated parameter tuning, it is possible to grow powerful models that can successfully predict class affiliations of future observations.

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. Without this knowledge, the abundance of cryptic or unknown species might be under- or overestimated.

Meristic and morphometric characters are powerful taxonomic tools for measuring discreteness and relationships among fish species. For this reason, analysis of morphometric and meristic characters has

not been widely used by ichthyologists to differentiate between different species and among different populations within a species. [4] However, Morphological characteristics often are found insufficient for the identification of cryptic species. Several cryptic species of anurans display a high level of morphological similarities that often make them virtually impossible to distinguish on the basis of morphological parameters. As these species usually are misidentified or ignored because of their taxonomic complexity in both ecologically diversified regions.

An automated species identification system is a matter of a one-to many matching, which not only needs to match an individual specimen with one of a set of extremely similar species to one another, but also is necessary to be able to reject it as belonging to a species that is not part of this set (Gaston and O'Neill, 2004) and patterns variation among fish species allows automated taxonomic identification of the species. [4]

A family of automated species identification systems has been designed in recent years for gathering and analyzing data from images of specimens [5] [9]. However, many of the taxonomic characteristics cannot be observed in a photograph.

The aim of this system is to use different machine learning techniques for developing the automated species identification system rather than the traditional taxonomic fish identification. To achieve efficient gene selection from thousands of candidate genes that can contribute in identifying fish, this work aims at developing a system utilizing efficient features selection and classification techniques and provide automated fish identification system for Myanmar.

3. Material and Methods

3.1. Preprocessing

The pre-processing has been performed in two stages. The data sets will used by this system are mixed of nominal and continuous types. Therefore, each numeric attribute needs to be discretized into intervals by first. In the second phase, the features and the samples have been analyzed for missing variables and records with appropriate mean values or null values. Based on the importance of data the missing features or the samples have been removed. Renaming and transformation has been performed for few attributes. Duplicate values are eliminated by ignore.

3.2. Features Selection

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.

Selecting suitable attributes is also an important step for effective and efficient classification. Many potential attributes may be used in the fish classification such as measurements and scale counts of body parts, and it can be done by the feature selection process. The purpose of feature selection is to determine the most relevant and the least amount of data representation of the specimen in order to minimize the within-class pattern variability, whilst, enhancing the between-class pattern variability.

In this system, features selection is performed using combination theory and two supervised classifiers, Random Forest and Attributes Selected. This combination based features selection method selects features using classification performance of classifiers as a criterion of feature subset selection. Dataset used in this system contains 16 features. These are Mouth, Teeth, Barbels, Snout, Operculum, Eye, Head, Predorsal scales, Dorsal Fin, Pelvic figs, Pectoral fins, Anal fin, Caudal fin, Dorsal fin spines, Adipose fin and Lateral line. Rather than using all 16 features for this task, we worked with varying the features, increasing in increments of 5 up to the maximum possible features for each data set in terms of combination such as (Mouth, Teeth, Barbels, Snout and Operculum), (Mouth, Teeth, Barbels, Snout, Eye), (Mouth, Teeth, Barbels, Snout, Head) etc. . By varying features in this way, we are able to gain some insight into each method's ability to capture the best features in a data set.

The goodness of a feature, or feature subset is evaluated by comparing the performance of the two learning classifiers applied on the selected subset that come from combination.

By comparing these two classification results, this system selected the10 features Head, Mouth, Teeth, Dorsal fin spines, Caudal fin, Snout, Eye, Anal fin, Predorsal Scales, Barbels as the best features subset. To classify the unknown species, these features are used to construct the training data in our proposed

identification system. It leads to the most accurate classification results on fish datasets.

This system also used Correlation based Feature Selection (CFS) in first step of Attributed Selected classifier. CFS evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. Correlation coefficient is used to estimate correlation between subset of attributes and class, as well as inter-correlations between the features. Relevance of a group of features grows with the correlation between features and classes, and decreases with growing inter-correlation. CFS is also used to determine the best feature subset and is usually combined with search strategies such as forward selection, backward elimination, bi-directional search, best-first search and genetic search. In this system, CFS is combined with best-first search. The following equation is used for CFS.

$$r_{zc} = \frac{k r_{zi}}{\sqrt{k + k(k-1) r_{ii}}} \qquad (1)$$

Where r_{zc} is the correlation between the summed feature subsets and the class variable, k is the number of subset features, r_{zi} is the average of the correlations between the subset features and the class variable, and r_{ii} is the average inter-correlation between subset features [3]. CFS can select the 7 features such as Head, Mouth, Teeth, Dorsal fin spines, Predorsal Scales, Eye and Anal fin. The number of features selected by CFS is less but the classification result based on these features subset is lower accuracy than combination based method.

3.3. Classification Algorithms

With the exponential growth in the amount of data that is being generated in recent years, there is a pressing need for applying machine learning algorithms to large data sets. Machine Learning algorithms are powerful tools not only for classification but also for the features selection. These algorithms can get higher classification accuracy for datasets from a wide variety of bioinformatics applications with complex behavior as well as various application areas. This system considers several classification algorithms and regression algorithm for classification and prediction of the species.

3.3.1. Random Forest Algorithm

Random Forest (RF) has been widely used for multi-label classification [8]. It is operated by constructing decision tree structure by the training examples. One of the popular algorithms is tree bagging, in which the training process includes repeatedly selecting a bootstrap sample of the training set and fitting the trees to them. After the training process, the label decision is made either on the majority of the votes or a weighted combination from individual trees.

Pseudo code for Random Forest (RF) is shown in Figure 1.

Input: D: training sample
a: number of input instance to be used to generate classification tree
T: total number of classification trees in random forest
OT: Classification Output from each tree T
1) OT is empty
2) for i=1 to T
3) Db = Form random sample subsets after selecting 2/3rd instances randomly from D
/* For every tree this sample would be randomly selected*/
4) Cb = Build classification trees using random subsets Db
5) Validate the classifier Cb using remaining 1/3rd instances //Refer Step 3.
6) OT=store classification outputs of classification trees
7) next i
8) Apply voting mechanism to derive output ORT of the Random forest (ensemble of classification trees)
9) return ORT

Figure 1. Random Forest Algorithm

3.3.2. Attributes Selected Algorithm

This classifier can provide the automatically feature selection and classification procedure. It has two main functions (1) evaluation and (2) classification. The first step of this algorithm uses CFS to search feature subsets according to the degree of redundancy among the features. The aims to find the subsets of features that are individually highly correlated with the class but have low inter-correlation. To determine the best feature subset, this step is usually combined with best-first search strategies. The second step is the classification by using the result features subset.

Pseudo code for Attributed Selected (AS) is shown in Figure 2.

Input: D Training Instances
Intermediate Output: best attributes subset B obtained by the Output of Correlation-Based Feature Selection (CFS),
Db = Training Instances with best attributes
Output: OT, Classification Output for the Attributed Selected classifier
BuildClassifier (Db)
1) Check for base cases
2) For each attribute *a*
3) Find the normalized information gain from splitting on *a*
4) Let bestAttrib be the attribute with the highest normalized information gain
5) Create a decision node that splits on bestAttrib
6) Recursive on the subsets obtained by splitting on bestAttrib, and add those nodes as children of node
7) Compute output OT of the complete data set

Figure 2. Attributes Selected Algorithm

4. The Proposed Automated Species Identification System

In this system, we use two classifiers to accurately identify the fish species from India and adjacent countries [14]. These fish datasets are identified by the researcher of Mandalay University. Supervised machine learning classifiers such as Random Forest [8] and Attributed Selected that have been used many applications in bioinformatics are applied in this system. To maximize the performance of the classifiers on previously unseen data, and reducing training data, combination based feature selection is used.

By using classifiers, the main morphological characters (attributes) are discriminated between the orders, families, genera and species. The proposed system attempts to apply the strength of Attribute Selected and Random Forest. There are mainly two reasons for selecting this methodology instead of traditional methods or other pattern recognition techniques.

Firstly, systems based on the use of traditional statistical methods, such as discriminant analysis or principal component analysis, have now been largely abandoned, mainly because they make restrictive assumptions about the statistical nature of the data, such as assuming linearity. Secondly, Random Forest will overcome the problem of over fitting. In training data, they are less sensitive to outlier data and parameters can be set easily and therefore, eliminates the need for pruning the trees. Random Forest not only keeps the benefits achieved by the Decision Trees but through the use of bagging on samples, its voting scheme through which decision is made and a random subsets of variables, it most of the time achieves better results than Decision Trees.[6]

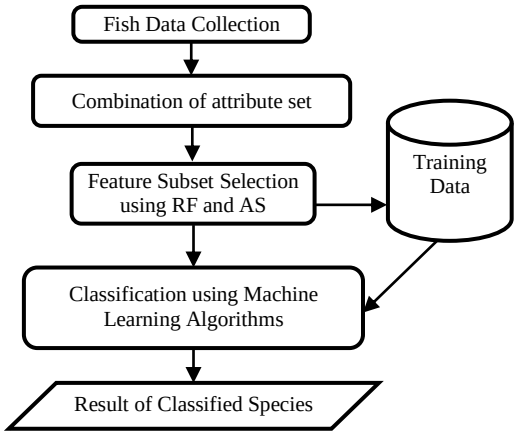

Figure 3. System Flow Diagram of the System

4.1. Dataset Description

Recent years, some researchers (taxonomists) from Mandalay University used the fish dataset from [14] to build the taxonomic fish identification system. However there are some inherent problems occur in the features extraction and classification. The first important limitation in the dataset is the huge number of redundant records. And they could not create the efficient database for these data sets. They worked manually on documents and could not be effectively identified because of the lack of methods and huge amount of data sets. The proposed system has to create fish database and use this database for determining the main taxonomic features of fish that are promoting divergence among closely related species.

4.2. Experiment Results

This system performs the feature subset selection by using different classifiers and combination theory. The Combination based Features Selection is iterative in nature. It ranks each features subset according to its average accuracy value, and then selects the features subsets with the highest accuracy values. In this work, we used the fish datasets of 1516 instances belonging to 20 classes. Each instance contains 16 attributes. The evaluation of the best features subsets of this system required comparing performance over all possible subsets of features on each classifier. This estimation considered both large and small features sets to estimate the performance of classifiers for the task of feature selection. And we assumed that the size of significant interactions between different combinations of features is much

smaller than 16 features, we limited ourselves to evaluating the performance from combination sampled from the whole set of features, with being a bound on the combination size 5. We performed 10 fold cross validation to test the efficiency of the model built during the training phase. The results along with the experimentation of different methods are compared based on accuracy, F-measure, area under the ROC, average precision, True Positive rate, False Positive rate, Recall, kappa statistics and squared error. Some of the comparison results are shown in Table 1.

By using these results, we can extract the most important features subset with the accuracy more than 90%. To develop the automated fish identification system with relevant features, this system will classify and identify the fish species by using different machine learning techniques based on the important features subset which are chosen from this experiment.

Table 1. Accuracy Comparison of Machine learning algorithms using Fish Dataset

Classifier	Detection Accuracy (%)	Precision	Recall	TP	FP
Random Forest	99.6%	0.998	0.996	0.996	0.005
Attribute Selected	95.7%	0.959	0.957	0.957	0.011

5. Conclusion

In this paper, combination based features subset selection method is used. This system also used the evaluators and search methods for the effective feature subset selection. The filter based features subset selection methods achieved the features subset with less features. However, the result of classification accuracy based on these features subsets are lower than the result based on the features subsets which are chosen by the combination of 5 up to maximum possible features tests.

Our experimental results suggested that our combination based feature selection methodology can be successfully used to significantly improve the accuracy of fish classification systems. These results successfully demonstrated the value of applying combination theory concepts to feature selection. By using this feature selection method, we can extract the 10 common features with highest performance of both Random Forest and Attributes Selected classifiers. We used these best features for constructing the training examples with good generalization capability to correctly classify the class label of instance it has never seen before.

The key idea of this system is to reduce the time spent on the taxonomic identification of fishes and to provide a tool for accurate classification. For future work, further classifications are required to observe the feature subset selection in gene expression and to identify the fish species accurately and automatically based on different machine learning techniques.

References

[1] Ahmed Majid Taha, Aida Mustapha, and Soong-Der Chen, "Naive Bayes-Guided Bat Algorithm for Feature Selection", Hindawi Publishing Corporation, The Scientific World Journal, Volume 2013, Article ID 325973.

[2] Anne-Laure Boulesteix ,Silke Janitza ,Jochen Kruppa, Inke R. K¨onig, "Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics", July 25th 2012.

[3] Asha Gowda Karegowda, A. S. Manjunath, M.A.Jayaram, "Comparative Study of Attribute Selection Using Grain Ratio and Correlation Based Feature Selection ", International Journal of Information Technology and Knowledge Management , Volume 2, No. 2, July-December 2010, pp. 271-277.

[4] C. Guisande, A. Manjarrés-Hernández, P. Pelayo-Villamil, C. Granado-Lorencio, I. Riveiro, A. Acuna, E. Prieto-Piraquive, E. Janeiro, J.M. Matías, C. Patti, B. Patti, S. Mazzola S. Jiménez, V. Duque, F. Salmerón, "IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques", Fisheries Research102 , 2010, pp. 240–247.

[5] Huimin Chen Henry L. Bart, Jr. Shuqing Huang, "Integrated Feature Selection and Clustering for Taxonomic Problems within Fish Species Complexes", Journal of Multimedia, Vol. 3, No. 3, July 2008.

[6] I naki Inza, Borja Calvo, Rub´ en Arma˜ nanzas, Endika Bengoetxea, Pedro Larra˜ naga, and Jos´ eA.Lozano, "Machine Learning: An Indispensable Tool in Bioinformatics".

[7] Jehad Ali, Rehanullah Khan, Na sir Ahmad, Imran Maqsood, "Random Forests and Decision Trees", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 5, No 3, September 2012, ISSN (Online): 1694-0814.

[8] L. Breiman.,"Random Forests. Machine learning", 45(1):5–32, 2001.

[9] Mutasem Khalil Alsmadi, Khairuddin Bin Omar, Shahrul Azman Noah, Ibrahim Almarashdeh, "Fish Recognition Based On Robust Features Extraction From Color Texture Measurements Using Back-Propagation Classifier", Journal of Theoretical and Applied Information Technology, 2010.

[10] Ranjit Abraham , Jay B. Simha and S. Sitharama Iyengar, “Effective Discretization and Hybrid feature selection using Naïve Bayesian classifier for Medical datamining”, International Journal of Computational Intelligence Research, ISSN 0974-1259 Vol.5, No.2, 2009, pp. 116–129.

[11] Ron Kohavi, “Scaling up the Accuracy of Naïve-Bayes Classifier: Decision Tree-Hybrid”.

[12] S. M. Kamruzzaman, Farhana Haider and Ahmed Ryadh Hasan, “Text Classification using Association Rule with a Hybrid Concept of Naive Bayes Classifier and Genetic Algorithm”,

[13] Tarun Rao & T.V. Rajinikanth, “A Hybrid Random Forest based Support Vector Machine Classification supplemented by boosting”, Global Journal of Computer Science and Technology, C Software and Data Engineering, Volume 14 Issue 1 Version 1.0, 2014.

[14] Talwar, P.K. and Jhingran, A.G., *Inland Fishes of India and adjacent Countries*, Volume I, II, Oxford and IBH Publishing Co.Ltd. Calcutta, pp. 1-1158.