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

termed the "taxonomic crisis" (Dayrat, 2005). Without this knowledge, the abundance of cryptic or unknown species might be under or overestimated. [1]

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. 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 [5]. A family of automated species identification systems has been designed in recent years for gathering and analyzing data from images of specimens [7] [11].

3 Material and Methods

3.1 Preprocessing

Raw data is highly susceptible to noise, missing values and inconsistency. The quality of data affects the classification results. In order to improve the classification results raw data is preprocessed. Preprocessing is one of the most critical steps for machine learning algorithms as well as the data mining process which deals the preparation and transformation of the initial dataset.

In this proposed system, the data preprocessing has been performed in two steps. The data sets will used by this system are mixed of nominal and continuous types. However, most of the machine learning algorithms is primarily oriented to handling discrete features. Therefore, each continuous attribute is partitioned into categories. In the second phase, the features and the samples have been analyzed for missing variables and records with appropriate mean values by using mean imputation.

3.2 Features Selection

Most of the feature subset selection (FSS) algorithms are not suitable for all dataset. There is rarely a good way to choose appropriate FSS algorithms for various type of dataset. Therefore, FSS algorithm automatic recommendation is very important and practically useful.

Selecting suitable attributes is also an important step for effective and efficient fish 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 main idea 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 paper, a combination based Feature Selection method is presented. It works as an iterative algorithm for feature selection, the Combination based

Feature Selection algorithm focus on optimizing the performance of the classifiers on unknown data. In this approach, feature vectors are first selected and then features are selected by using classification performance of classifiers and majority voting of the features as a criterion of feature subset selection.

3.2.1 The Proposed Combination Based Feature Selection Algorithm

The Combination based Feature Selection, described in detail in Fig. 1, is iterative in nature. It ranks each features vector according to its average accuracy value, and then selects the features vectors with the highest accuracy values

Find the optimal feature vector by evaluating the classification performance of classifiers for all possible feature vectors

1. Let D be the set of Fish data instances with m features, C be a set of predefined Classifiers
2. Let A be the feature set of each instance A={f1,f2,……fm}
3. Let k be the number of features and s be the feature vector with k features s={f1,f2,….fk}
4. Create S be the set of the feature vectors S = {s1,s2,….,sN} by using Combination Theory where $N={}^{m}C_{k}$

   // i.e m = 16, k = 4 , N=1840 length 4 vectors
5. Let S' be a set of selected feature vectors S'= ϕ
6. Optimal feature vector F= ϕ
7. For each classifiers c ϵ C
8. For each feature vector s ϵ S

   Measures the performance of each classifier c generated using D with s
9. If classification accuracy greater than the accuracy-threshold

   S'= S' ∪ s //select the relevant feature vectors
10. End for

11. For each feature fi in s in S'

    Compute voting(fj), the count of each feature in A that appeared in S', j=1,2,….m
12. voting(fj)=Σ I(fj=fi) //where I(.) is the indicator function that returns 1 if the argument is true and 0 otherwise.
13. End If
14. End for
15. Compute majority-voting (fi) =

$$= \frac{voting(fi)}{\sum_{i=1}^{m} voting(fi)}$$

16. End For
17. Select the relevant features from *A* by using majority- voting of each feature *fi* that exceeds the voting-threshold

    F= F ∪ {fi}
18. End For
19. Return F.

Fig. 1 Combination based Feature Selection Algorithm

3.2.2 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. Most of 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 considered two classification algorithms for evaluation of the relevant features subsets.

Naive Bayes Algorithm. Naïve Bayes is one of the most effective and efficient inductive learning algorithms. It has been widely used for data classification. As a classifier it learns from training data from the conditional probability of each attribute given the class label. Using Bayes rule to compute the probability of the classes given the particular instance of the attributes, prediction of the class is done by identifying the class with the highest posterior probability. Computation is made possible by making the assumption that all attributes are conditionally independent given the value of the class. Naïve Bayes as a standard classification method in machine learning stems partly because it is easy to program, its intuitive, it is fast to train and can easily deal with missing attributes [11]. For a Sample S with n levels of genus {g1, g2,….,gn} for the n features, the posterior probability that s belongs to class label c is

$$P(c|S) \sim \prod_{c \epsilon C} P(g_{1,} g_{2,} \dots, g_n|c)$$

Here $P(g_{1,} g_{2,} \dots, g_n|c)$ are conditional probabilities estimated from training samples.

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.

4 The Proposed Automated Species Identification System

In this system, we use two classifiers to choose the main morphometric features of the fish species. These fish datasets are identified by the researcher of Mandalay University. Supervised machine learning classifiers such as Naïve Bayes (NB) and Attributed Selected (AS) that have been used many applications in bioinformatics

are applied in feature selection. 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 orders, families, genera and species can be discriminated based on the morphological characters (attributes).

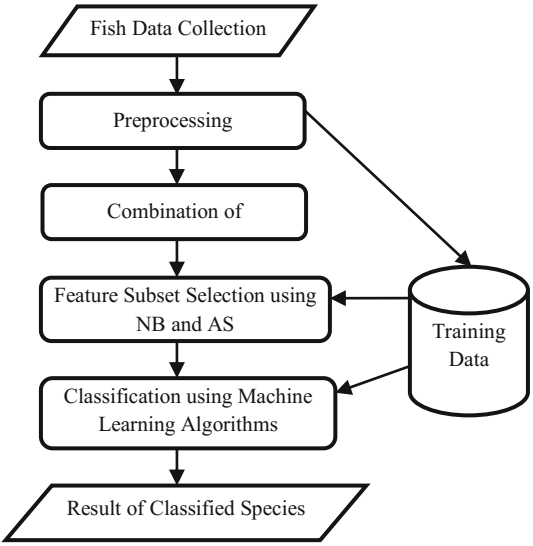

Fig. 2 System Flow Diagram of the system

4.1 Dataset Description

Recent years, some researchers (taxonomists) from Mandalay University attempt to build the taxonomic fish identification system. However there are some inherent problems occur in the features selection and classification. The first important limitation in the dataset is the huge number of records. And researchers could not create the efficient database for these data sets. Researchers worked manually on documents and could not effectively identify 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 Feature Evaluation

The parameters used to fine tune the feature selection are accuracy-threshold (at) and voting-threshold (vt).

Only the accuracy and frequency measures which exceed these predefined thresholds are considered for evaluation. On testing with various values for these parameters we found the system gave optimum performance for the values of at=0.9 and vt=0.043.

4.3 Experiment Results

This system performs the feature subset selection by using different classifiers and combination theory. In this work, we used the fish datasets of 1516 instances belonging to 20 classes. Each instance contains 16 attributes. These attributes are Mouth, Teeth, Barbels, Snout, Operculum, Eye, Head, Predorsal Scales, Dorsal Fin, Pelvic Fins, Pectoral Fins, Anal Fin, Caudal Fin, Dorsal Fin Spines, Adipose Fin and Lateral Line. Three popular classification algorithms including J48 decision tree-inducing algorithm, Multilayer Perceptron (MLP) and Support Vector Machine (SVM) are tested with different number of features on Fish dataset. For the purpose of this experimentation, Weka, Data Mining open source machine learning software is used [5]. Parameter settings were changed only when a significant improvement in performance based on preliminary experimentation was obtained. For the MLP learner, hidden layers were changed to 3 to define a network with one hidden layer containing three nodes. For SVM, the complexity constant c was changed from 1.0 to 3.0.

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 with k=4. The results along with the experimentation of different methods are compared based on accuracy, F-measure, area under the ROC, average precision and True Positive rate, False Positive rate and Recall. Table 1 summarizes the classifiers' performance on the test set and the number of features selected in each of the experiments. The accuracy levels are the fraction of correctly classified test set instances. The result shows that for most of the classifiers tested in this experiment, the features selected by proposed method work better in accuracy than the features selected by Correlation based Feature Selection method. Especially, the proposed method can effectively reduce the irrelevant features and successfully compete with an existing feature selection method. It can select the 5 relevant features, Mouth, Teeth, Snout, Head and Dorsal fin spines.

Table 1 Comparison of Accuracy levels and number of features selected using the different Classifiers

Classifiers	16 Features (without Features Selection)		(CFS with 9 Features)		(Proposed with 5 Features)	
	Accuracy	Time(Sec)	Accuracy	Time(Sec)	Accuracy	Time(Sec)
MLP(ANN)	79.1304%	10.64	79.1304%	6.84	82.6087 %	4.61
SMO(SVM)	100%	3.58	99.1304%	3.6	99.1304%	2.73
J48	92.1739%	0.07	92.1739%	0.05	93.0435%	0.03

5 Conclusion

In this paper, combination based features selection method is used. This method achieved not only the features subset with less features but also the result of classification accuracy. It can also be compared with the existing correlation based features selection method. The experimental results showed that the combination based feature selection methodology can define the curse of dimensionality to enhance the prediction performance, reduce measurement and storage requirements and also reduce training and prediction times. These results successfully demonstrated the value of applying combination theory concepts and majority voting method to feature selection. By using this feature selection method, we can extract the 5 common features. We used these best features for constructing the training examples with good generalization capability to correctly classify the unknown class label of instance.

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. Hernández-Serna, A., Jiménez-Segura, L.F.: Automatic identification of species with neural networks. Hernández-Serna and Jimenéz-Segura. PeerJ (2014). doi:10.7717/peerj.563
2. Taha, A.M., Mustapha, A., Chen, S.-D.: Naive Bayes-Guided Bat Algorithm for Feature Selection. Hindawi Publishing Corporation. The Scientific World Journal, vol. 2013, Article ID 325973
3. Boulesteix, A.-L., Janitza, S., Kruppa, J., König, I.R.: Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics, July 25, 2012
4. Karegowda, A.G., Manjunath, A.S., Jayaram, M.A.: Comparative Study of Attribute Selection Using Grain Ratio and Correlation Based Feature Selection. International Journal of Information Technology and Knowledge Management **2**(2), 271–277 (2010)
5. Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuna, A., Prieto-Piraquive, E., Janeiro, E., Matías, J.M., Patti, C., Patti, B., Mazzola, S., Jiménez, S., Duque, V., Salmerón, F.: IPez: An expert system for the taxonomic identification of fishes based on machine learning techniques. Fisheries Research **102**, 240–247 (2010)
6. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.: The weka data mining software: an update. ACM SIGKDD Explorations Newsletter **11**(1), 10–18 (2009)
7. Chen, H., Bart Jr., H.L., Huang, S.: Integrated Feature Selection and Clustering for Taxonomic Problems within Fish Species Complexes. Journal of Multimedia **3**(3), July 2008

8. Inza, I., Calvo, B., Armañanzas, R., Bengoetxea, E., Larrañaga, P., Lozano, J.A.: Machine Learning: An Indispensable Tool in Bioinformatics
9. Ali, J., Khan, R., Ahmad, N., Maqsood, I.: Random Forests and Decision Trees. IJCSI International Journal of Computer Science Issues **9**(5), No 3, 1694–0814, September 2012
10. Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (2001)
11. Alsmadi, M.K., Omar, K.B., Noah, S.A., Almarashdeh, I.: Fish Recognition Based On Robust Features Extraction From Color Texture Measurements Using Back-Propagation Classifier. Journal of Theoretical and Applied Information Technology (2010)
12. Abraham, R., Simha, J.B., Iyengar, S.S.: Effective Discretization and Hybrid feature selection using Naïve Bayesian classifier for Medical data mining. International Journal of Computational Intelligence Research **5**(2), 116–129 (2009). ISSN 0974-1259
13. Kohavi, R.: Scaling up the Accuracy of Naïve-Bayes Classifier: Decision Tree-Hybrid
14. Kamruzzaman, S.M., Haider, F., Hasan, A.R.: Text Classification using Association Rule with a Hybrid Concept of Naive Bayes Classifier and Genetic Algorithm
15. Rao, T., Rajinikanth, T.V.: A Hybrid Random Forest based Support Vector Machine Classification supplemented by boosting. Global Journal of Computer Science and Technology, C Software and Data Engineering **14**(1), Version 1.0 (2014)
16. Talwar, P.K., Jhingran, A.G.: Inland Fishes of India and adjacent Countries, vol. I, II. Oxford and IBH Publishing Co. Ltd., Calcutta, pp. 1–1158