Proposed method to decide the appropriate feature set for fish classification tasks using Artificial Neural Network and Decision Tree

Mutasem khalil Alsmadi
Dept. Information science
University Kebangsaan Malaysia
Kuala Lumpur, Malaysia

Prof.Dr.Khairuddin Bin Omar
Dept. System Science and Management
University Kebangsaan Malaysia
Kuala Lumpur, Malaysia

Prof.Dr.Shahrul Azman Noah
Dept. System Science and Management
University Kebangsaan Malaysia
Kuala Lumpur, Malaysia

Summary
We presents in this paper a novel fish classification methodology based on a robust feature selection technique using Artificial Neural Network and Decision Tree. Unlike existing works for fish classification, which propose descriptors and do not analyze their individual impacts in the whole classification task, we propose a general set of features and their correspondent weights that should be used as a priori information by the classifier. In this sense, instead of studying techniques for improving the classifiers structure itself, we consider it as a "black box" and focus our research in the determination of which input information must bring a robust fish discrimination. The study area selected for our proposed method from department of fisheries Malaysia ministry of agricultural and Agro-based industry in putrajaya, Malaysia region currently, the database contains several hundreds of fishes. In the future, we shall enhance the capability of the decision tree and ratification neural network classifier to deal with more than 2,000 fish species, which is the total amount of fish species along the coast of Malaysia. Data acquired on 22th August, 2008, is used. The classification problem involved the identification of 350 types of image fishes; family ,Scientific Name , English name , local name, Habitat , poison fish and non-poison, based on set of extraction feature .The main contribution of this paper is enhancement recognize and classify fishes based on digital image. Both classification and recognition are based on feature extraction.

Key words:
Artificial Neural Network ,Decision Tree, multilayer-perceptron (MLP) ,Image Recognition, poison fish and non poison.

1. Introduction

Recognition and cataloging are the vital facets in this up-to-the-minute era of research & development, hence exploiting the accessible techniques in Artificial Intelligence (AI) and Data Mining (DM) to achieve optimal production levels, examination procedures, and enhancing methodologies in most fields principally in the agricultural domain.

Artificial neural networks are defined as computational models of nervous system. Significantly natural organisms do not only possess nervous system; in fact they also evolve genetic information stored in the nucleus of their cells (genotype). Furthermore, the nervous system as a whole is part of the phenotype which is derived from the genotype through a specific development process. The information specified in the genotype determines assorted aspects of the nervous system which are expressed as innate behavioral tendencies and predispositions to learn [7], acknowledges that when neural networks are viewed in the broader biological context of Artificial Life, they tend to be accompanied by genotypes and to become members of budding populations of networks in which genotypes are inherited from parents to offspring. Many researchers such as Holland, Schwefel, and Koza, have stated that Artificial Neural Networks are evolved by the utilization of evolutionary algorithms.

Moreover, there are several methods that can make the computer more intelligent and to give it enough intelligence to recognize and to understand the images that the user gives to it. One of this ways is using the Artificial Intelligence (AI) and Decision Tree (DT) Science. Using one of AI techniques such as Neural Network (NN) will help us in recognize and then classify the entered image which will give a big contribution in the agriculture domain especially in fish recognition and classification.

Nery *et al.* (2006) The Object classification problem lies at the core of the task of estimating the prevalence of each fish species. They mentioned about that this issue still has a problem with classification and identification of fish species, and the authors understand that any solution to the automatic classification of the fish should address the following issues as appropriate:

1) Arbitrary fish size and orientation; fish size and orientation are unknown a priori and can be totally arbitrary;

2) Feature variability; some features may present large differences among different fish species;

Manuscript received March 5, 2009
Manuscript revised March 20, 2009

3) Environmental changes; variations in illumination parameters, such as power and color and water characteristics, such as turbidity, temperature, not uncommon. The environment can be either outdoor or indoor;

4) Poor image quality; image acquisition process can be affected by noise from various sources as well as by distortions and aberrations in the optical system;

5) Segmentation failures; due to its inherent difficulty, segmentation may become unreliable or fail completely;

And the vast majority of research-based classification of fish points out that the basic problem in the classification of fish; they typically use small groups of features without previous thorough analysis of the individual impacts of each factor in the classification accuracy.

2. Artificial Neural network

An ANN is a form of artificial intelligence that imitates some functions of the human brain. The network comprises a large number of simple processing units linked by weighted connections according to a specified architecture. The knowledge of the network is stored in the strength of the weighted connections between units. Such networks can learn and generalize; they are parallel in nature (Aleksander and Morton, 1990; Hammerstrom, 1993a, 1993b; Bishop, 1995). ANN classifiers have been used in a wide range of applications in remote sensing including supervised classification (Benediktsoon et al., 1990; Kanellopoulos et al., 1992; Swinnen, 2001), and unsupervised classification (Baraldi and Parmiggiani, 1995; Hara et al., 1995; Sveinsson, 2001).

Artificial Neural Networks (ANN) , due to its useful properties such as: flexibility, nonparametric nature which gives independence of data distribution, plus ability to handle data acquired at different levels of measurement, precision, and -once trained-rapid data processing , ability to handle non-linear relations between feature and classes, less sensitive to some of the problems associated with conventional classifiers and it makes no assumptions about the nature and distribution of the data, highly parallel mechanism, excellent fault tolerance, adaptation and self-learning, has become increasingly developed and successfully used in character recognition [1,2, 3, 4, 5, 6]. The key power provided by such networks is that they admit fairly simple algorithms where the form of non-linearity that can be learned from the training data. The models are thus extremely powerful, have nice theoretical properties, and apply well to a vast array of real-world applications.

3. Neural Network

The multilayer-perceptron (MLP) model using the back propagation (BP) algorithm is one of the well-known neural network classifiers, which consists of sets of nodes arranged in multiple layers with connections only between node in the adjacent layers by weights. The layer, where the inputs information is presented, is known as the inputlayer. The layer where the processed information is retrieved is called the output layer. All layers between the input and output layers are known as hidden layers. A schematic of a MLP model layers are shown in Figure 1.

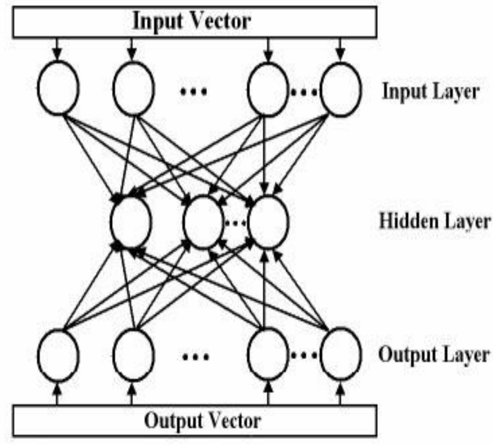

Figure 1: Schematic of a 3-layer MLP model.

For all nodes in the network, except the input layer nodes, the total input of each node is the sum of weighted outputs of the nodes in the previous layer. Each node is activated with the input to the node and the activation function of the node. In Figure 7, node computations are shown (Hagan et al., 1996).

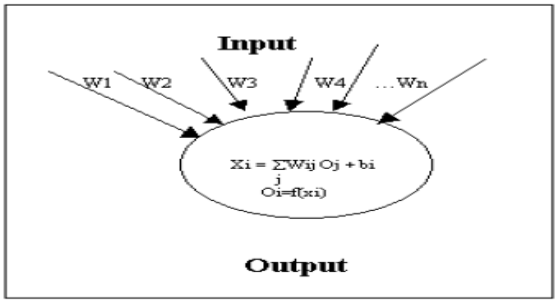

Figure 2: Node computations.

The input and output of node i (except for the input layer) in a MLP mode, according to the BP algorithm is :

Input : $X_i = \sum w_{ij} + B_i$ (1)
Output : $O_i = F(X_i)$ (2)

Where
Wij : the weight of the connection from node i to node j.

Bi : the numerical value called bias.

F : the activation function.

The sum in Eq.18 is over all nodes J in the previous layer. The output function is a nonlinear function, which allows a network to solve problems that a linear network cannot solve. In this study the sigmoid function given in Eq.20 is used to determine the output state, sigmoid function is used to simplify error calculations.

$F(X_i)=1/(1+e(-x_i))$ (3)

Back-propagation (BP) learning algorithm is designed to reduce an error between the actual output and the desired output of the network in a gradient descent manner. The summed squared error (SSE) is defined as:

$SSE = ½(\sum p \sum I\ O_{pi} – T_{pi})$ (4)

Where p indexes the all-training patterns and i indexes the output nodes of the network. Opi and Tpi denote the actual output and the desired output of node, respectively when the input vector p is applied to the network (Hagan et al., 1996).

A set of representative input and output patterns is selected to train the network. The connection weight Wij is adjusted when each input pattern is presented. All the patterns are repeatedly presented to the network until the SSE function is minimized and the network "learns" the input patterns. Applications of the gradient descent method yield the following iterative weight update rule:

$Dw_{ij}(n+1) = h(d_iO_i) + aDw_{ij}(n)$ (5)

Where:
D: the learning factor.

a: the momentum factor.

di: the node error, the output of node I is then given as.

$d_i = (t_i-O_i)O_i(1-O_i)$ (6)
The node error at an arbitrary hidden node is
$\delta_i = O_i(1-O_j)\sum \delta_k W_{ki}$ (7)

For details about BP algorithm refer to Hagan et al.(1996).

4.Clustering Algorithms

In the last years, clustering algorithms have been finding a large variety of applications in the last years.. Among the others it is possible mention: optical character recognition, image classification, and medical applications object recognition and document analysis.

Clustering can be defined as a process that organizes objects into groups or clusters, whose members are similar in some way. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” from the objects belonging to other clusters. The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute best criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit his needs. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection) (Cordella et al,.2005).

5. Decision Trees

In the usual approach to classification, a common set of feature is used jointly in single decision step. An alternative approach is to use a multistage or sequential hierarchical decision scheme. The basic idea involved in any multistage approach is to break up a complex decision into a union of simpler decisions, hoping the final solution obtained in this way would resemble the intended desired solution. Hierarchical classifiers are a special type of multistage classifier that allows rejection of class labels at intermediate stage.

Classification trees offer an effective implementation of such hierarchical classifiers. Indeed, classification trees have become increasingly important due to their conceptual simplicity and computational efficiency. A decision tree classifier has a simple form which can be compactly stored and that efficiently classifies new data. Decision tree classifiers can perform automatic feature selection and complexity reduction, and their tree structure provides easily understandable and interpretable information regarding the predictive or generalization ability of the classification.

To construct a classification tree by heuristic approach, it is assumed that a data set consisting of feature vectors and their corresponding class labels are available. The features are identified based on problem specific knowledge. The decision tree is then constructed by recursively partitioning a data set into purer, more homogenous subsets on the basis of a set of tests applied to one or more attribute values at each branch or node in the tree. This procedure involves three steps: splitting nodes, determining which nodes are terminal nodes, and assigning class label to terminal nodes. The assignment of class labels to terminal nodes is straightforward: labels are assigned based on a majority vote or a weighted vote when it is assumed that certain classes are more likely than others.

6. The proposed method of recognition fish image

In the classification process performed by pattern recognition system three different operations can be distinguished: preprocessing, feature extraction and classification. The main steps in the figure 3 below which will be included in this paper in the Fish recognition system related to this proposed classifier will be as follows:

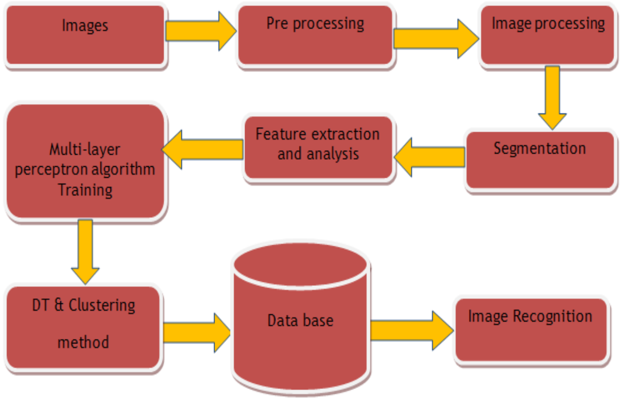

Figure 3: Fish recognition model

7. Conclusion

In this paper,, a novel approach to fish recognition based on artificial neural network and decision tree to the an optimal recognition of fish image , this proposed method is expected to has the capabilities to recognize and classify the given fish image based on its extracted features by utilizing the use of decision tree and clustering methods with back-propagation algorithm to determine the fish type. And to enhance the classification of the fish based on size measurements, shape measurements, color signatures and texture measurements, of the fish's patterns. The proposed method It will contribute a lot in the agriculture domain in Malaysia and Scientists in the same field, which they can utilize it to do new researches in investigating and exploring fishes and Marine world. Moreover, researchers, students, and amateurs will benefit from it in their own research.

References

[1] Parisi. D. Artificial Life and Higher Level Cognition, Brain and Cognition, Volume 34, Issue 1, June 1997, Pages 160-184, 2002.

[2] Hammerstrom, D., "Neural Networks at Work," IEEE Spectrum, 30(6), 26-32 (1993 a).

[3] Aleksander, I.; Morton, H., "An Introduction to Neural Computing," (London:Chapman and Hall) 1990.

[4] Bishop, C. M., "Neural Networks for Pattern Recognition," (Oxford: Clarendon Press) 1995.

[5] Benediktsson, J. A.; Swain, P. H.; Ersoy, O. K.; Hong, D., "Neural Network Approaches Versus Statistical Methods in Classification of Multisource Remote Sensing Data," IEEE Transactions Geosci. Remote Sensing, vol.28, pp. 540-551, July 1990.

[6] Kanellopoulos, I.; Varfis, A.; Wilkinson, G. C.; Megier, J., "Land-cover Discrimination in SPOT HRV Imagery Using an Artificial Neural Network – a 20-class Experiment," International Journal of Remote Sensing, 13,1992, pp 917-924. Swinnen, E.; Eerens, H.; Lissens, G.; Canters, F., "Sub-pixel Land-cover Classification with SPOT-VEGETATION Imagery," IEEE Geoscience andRemote Sensing Symposium, 2001, Volume: 1, 9-13 July 2001.

[7] Baraldi, A.; Parmiggiani, F., "A Neural Network for Unsupervised Categorisationof Multivalued Input Patterns: An Application to Satellite Image Clustering," IEEE Transactions on Geoscience and Remote Sensing, 33,305-316 (1995).

[8] Hara, Y.; Atkins, R. G.; Shin, R. T.; Kong, J. A.; Yeuh, S. H.; Kwok, R., Application of Neural Networks for Sea Ice Classification in Polarimetric SAR images," IEEE Trnasactions on Geoscience and Remote Sensing, 33, 704-748 (1995).

[9] Sveinsson, J.R.; Ulfarsson, M.O.; Benediktsson, J.A., "Cluster-Based Feature Extraction And Data Fusion In The Wavelet Domain," IEEE Geoscience and Remote Sensing Symposium, 2001, Volume: 2, 9-13 July 2001.

[10] L.P. Cordella.; C. De Stefano.," An Approach to Pattern Recognition by Evolutionary Computation" Novembre 2005.

[11] M. Pal .; P.M. Mather.," DECISION TREE BASED CLASSIFICATION OF REMOTELY SENSED DATA "22ndAsian Conference on Remote Sensing, Singapore, 5 - 9 November 2001.

[12] Anderson J A .An Introduction to Neural Networks.MIT Press, Cambridge, MA,1995.

[13] Chauvin Y, Rumelhart D E (eds.). "Backpropagation: Theory, Architectures, and Applications". Erlbaum, Mahwah, NJ,1995.

[14] Freund, Y. and Schapire, R. E. Large margin classification using the perceptron algorithm. In Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT' 98). ACM Press,1998.

MUTASEM KHALIL SARI AL SMADI received his BS degree in Software engineering in 2006 from Philadelphia University, Jordan, his MS degree in intelligent system in 2007 from University Utara Malaysia, Malaysia; currently he is doing PhD in Intelligent System in University Kebangsaan Malaysia in Malaysia.

Khairuddin Omar is an Associate Professor in Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Malaysia (e-mail: ko@ftsm.ukm.my). His research interests includes Arabic/Jawi Optical Text Recognition, Artificial Intelligence for Pattern Recognition & Islamic Information System.

Shahrul Azman Noah is currently an associate professor at the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM). He received his MSc and PhD in Information Studies from the University of Sheffiled, UK in 1994 and 1998 respectively. His research interests include information retrieval, knowledge representation, and semantic technology. He has published numerous papers related to these areas. He currently leads the Knowledge Technology research group at UKM.