Asian Journal of Scientific Research

ISSN 1992-1454

science alert
http://www.scialert.net

ANSInet
an open access publisher
http://ansinet.com

OPEN ACCESS Asian Journal of Scientific Research

ISSN 1992-1454
DOI: 10.3923/ajsr.2016.266.272

Case Report
Computer Aided Taxonomy: A Case Study on the Automated Identification of Invasive Ladybirds in the UK

M.Z. Ayob and M.K.A. Kadir

British Malaysian Institute, Universiti Kuala Lumpur, Bt. 8, Jalan Sungai Pusu, 53100 Gombak, Selangor, Malaysia

Abstract
Feature selection and minimization are some of the early steps in a pattern recognition system. Computer aided taxonomy consisting of decision tree and human interaction is proposed as an intermediate process in the identification of invasive ladybirds in the UK. The proposed methodology have been compared with learning-based system such as multilayer perceptron (MLP). The J48 has been able to reduce the span of features. Using J48 decision tree along with MLP shows that the decision tree and human interaction together forms a constructive element for improving the identification of ladybird species with black spot colours.

Key words: Computer aided taxonomy, species identification, decision tree, multilayer perceptron, human interaction

Received: June 16, 2016 Accepted: August 29, 2016 Published: November 15, 2016

Citation: M.Z. Ayob and M.K.A. Kadir, 2016. Computer aided taxonomy: A case study on the automated identification of invasive ladybirds in the UK. Asian J. Sci. Res., 9: 266-272.

Corresponding Author: M.Z. Ayob, British Malaysian Institute, Universiti Kuala Lumpur, Bt. 8, Jalan Sungai Pusu, 53100 Gombak, Selangor, Malaysia

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

The principle of identification involves the process of comparing a representation of an individual specimen with taxa[1,2]. In contradiction to typical dichotomous key, Automated Species Identification (ASI) involves the use of a computer to aid species identification. Inputs may range from image sensors such as camera or sounds. Applications range from identification of quarantine fungal pest *Tilletia indica*[3], Lepidoptera[4-7]. Image-based identification has also been extended on Hymenoptera, for example on braconid wasps[8-10], honeybees[11-14], solitary bees[15], ichnumonid wasps[16], parasitic wasps[17] and leafhoppers[18], bumblebees[19], moth[20] and spider[21]. In botany, one of the pioneering study by Clark[22,23] of the University of Surrey aims for the identification of mature specimens taken from the crown of the tree. The system uses characters obtained from cultivated species of the genus *Tilia*, commonly known as lime trees. Clark showed that a systematic methodology in applying character and measurement data into MLP results in effectively tuned system parameters, which could be useful for non-experts to use for plant identification. Results for species identifications were shown in term of confusion matrix. The identification performance of the MLP was improved by 16% after the inclusion of minimal geographic information, represented in term of 3 geographic characters in binary code. A study published by Wu *et al.*[23] showed image processing and Probabilistic Neural Network (PNN) been applied to build general purpose automated leaf recognition for plant classification[24]. They were able to derive 12 leaf features from 5 basic geometric features, which were extracted after the implementation of image processing techniques on selected plant leaves. In this study, an attempt to use J48 decision tree along with multilayer perceptron (MLP) employing back propagation algorithm has been performed.

MATERIALS AND METHODS

J48 decision tree: The main concern in the study is the feature selection and minimization of features. The WEKA has been used for system development and testing. The WEKA is a machine learning tool, which stands for 'Waikato Environment for Knowledge Analysis[25]. It was developed by researchers in the University of Waikato, New Zealand[26]. The WEKA contains a collection of machine learning programs developed in JAVA to facilitate data mining tasks, such as training and testing artificial neural networks, decision trees and statistical visualizations. Classifiers included in WEKA are Bayes, RBF functions, Support Vector Machine (SVM), multilayer perceptron (MLP), Learning Vector Quantisation (LVQ), J48 decision tree and many more.

The expert system was designed based on J48 decision tree. The J48 is an open source Java implementation of the C4.5 algorithm. The C4.5 was actually derived from ID3. Both are Ross Quinlan's algorithms for generating classification models, better known as decision trees[25,27,28]. It contains a hierarchy of branches and leaves stemming from a root. When a classification is required, a decision tree uses its hierarchical and recursive nature to make decisions at each node (Fig. 1).

Imagine there are 10 samples each for the two dummy classes 'C5' and 'C7'. The most important is to determine, which attribute or feature to place at the root (top most node). The decision tree calculates the values of entropy before and after a node. For a binary split, entropy and information gain are given as:

$$\text{Entropy} = - p(a)\times\log_2 (p(a))-p(b)\times\log_2(p(b)) \quad (1)$$

$$\text{Information gain} = \text{Entropy before-entropy after} \quad (2)$$

Hall *et al.*[25] uses the term 'Information value' instead of entropy. The information gain for each candidate attribute is evaluated at each node and the attribute with the highest information gain is selected. To determine information gain for arbitrary attributes A (Fig. 2) and B (Fig. 3).

The Information Gain (IG) will be chosen from the attribute with the highest value:

$$\text{Information gain} = M0-M12$$

Or:

$$\text{Information gain} = M0-M34$$

To classify an unknown instance, the tree is traversed based on the values tested in successive nodes. If an attribute value is not nominal, the tree will form 2 subsets or branch. The branching depends on which subset the value lies in the decision tree. In the case of ladybird identification, the attributes are numeric. At a node, the number is checked if it is greater or smaller than a constant. This constant is the split criterion, where binary split occurs. Notice the 2 numbers at some leaves in Fig. 1 (the last nodes). The 1st number represents the total number of instances reaching the leaf. The 2nd is the number of those instances which are misclassified. Hence, the rule based part of the system aims at embedding structured human expertise into algorithmic form[29].

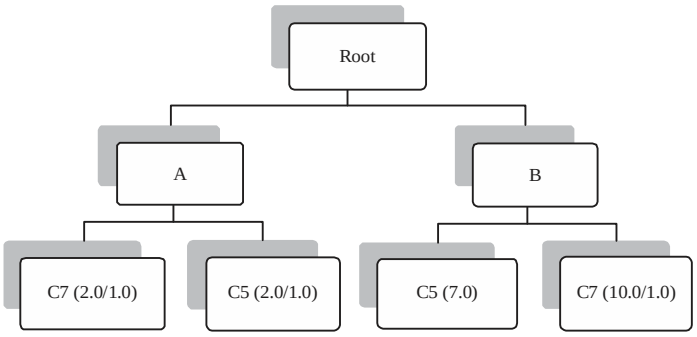

Fig. 1: Example decision tree

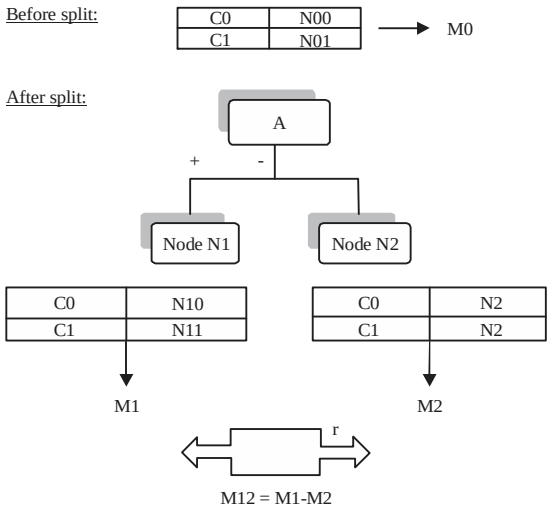

Fig. 2: Determine entropy for attribute 'A'

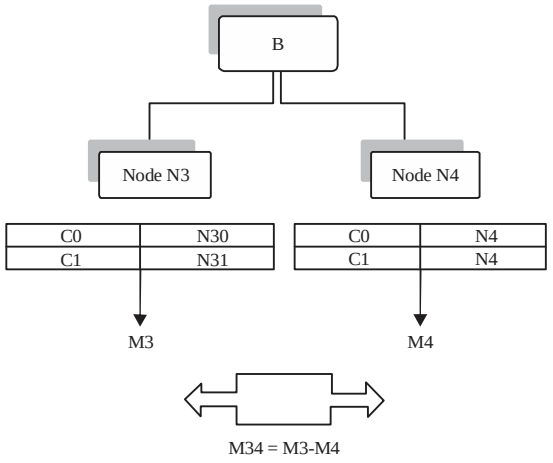

Fig. 3: Determine entropy for attribute 'B'

In short, the decision tree simplifies the solution when looking for which feature to use in a particular identification. It makes automated identification easier by reducing the number of features and shorten identification time[30].

Multilayer perceptron: A multilayer perceptron (MLP) neural network consists of numerous units of perceptrons with 1 or more hidden layers. A perceptron consists of a single neuron with adjustable synaptic weights and a hard limiter[31]. The weighted sum of the inputs is applied to the hard limiter. The input signals are propagated in a forward direction on a layer-by-layer basis. Neurons in the hidden layer function to detect the features, because the weights of the neurons represent the features hidden in the input pattern. The perceptron gives out +1 if the input is positive, while giving -1 if the input is negative. Therefore, the perceptron behaves as a simple classifier. In other words n-dimensional space is divided by a hyper plane into two decision regions.

Central to the operations of a MLP neural network is the feed forward and backpropagation algorithm. Feed forward operation study by introducing input to the hidden neuron, firing up neurons and calculating errors. This is normally done during training stage.

Training is done by presenting examples of the input and output relationship to the neural network. The connection weights will be adjusted in order to minimize an error function between the historical outputs and the outputs predicted by the neural network. Back propagation itself means adjusting weights in hidden layers by propagating errors back towards the input layer. By doing so the changes in input weight and output weight per neuron are calculated[32]. In order to perform classification hence identification, a neural network algorithm has to discriminate taxa by constructing decision boundaries. The boundaries are constructed between example patterns of known taxa in n-dimensional space. A simple 2-dimensional feature space is shown in Fig. 4.

It is desirable to see the relationship between the outputs of a neural network with a decision tree through WEKA simulation, as derived in Eq. 1. The datasets and simulation results are presented in the next section.

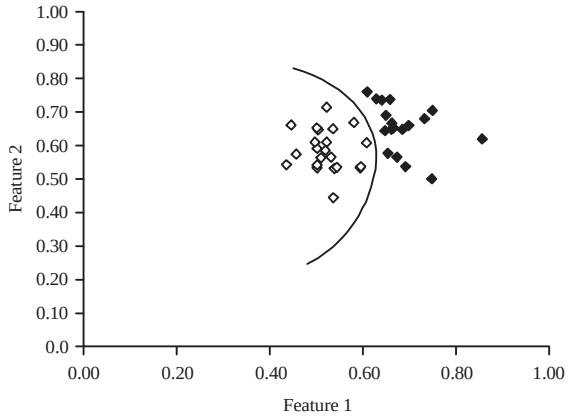

Fig. 4: Two dimensional feature space and non-linear decision boundary

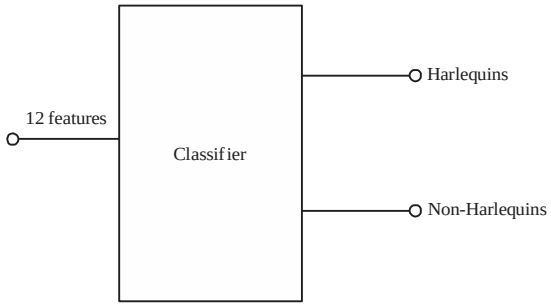

Fig. 5: Training scheme for sorting Harlequins from non-Harlequins

Harlequin ladybirds identification in UK: There are 12 features extracted from images containing an unknown ladybird. As explained in previous, these features shall be fed into a classifier. A scheme is shown in Fig. 5 with an aim to perform pre-sorting between Harlequins and non-Harlequins.

Training and test setup: The tests are conducted by grouping the ladybird images based on spot colour, specifically narrowed down to images containing ladybirds with black-spot colours (Fig. 6). The species involved are given in Table 1.

RESULTS AND DISCUSSION

Results are presented in the form of confusion matrix to show level of accuracy for the tests. The reader is referred to some extension of the metrics derived from the confusion matrix in Bradley's and Fawcett's study[28,33,34]. The metrics are:

Table 1: Assignment of class labels for black-spotted ladybirds

Species	Class label
A. punctata (2)	A2
C. punctata (3)	C5
C. punctata (7)	C7
H. axyridis f. *succinea*	H3

Table 2: Metrics for MLP (E4 and H1)

Class	TP rate	FP rate	Precision	Recall	AUC
E4	1	0.05	0.952	1	0.975
H1	0.95	0	1	0.95	0.975
Weighted average	0.975	0.025	0.976	0.975	0.975

Table 3: Metrics for J48 decision tree (E4 and H1)

Class	TP rate	FP rate	Precision	Recall	AUC
E4	0.925	0.075	0.925	0.925	0.959
H1	0.925	0.075	0.925	0.925	0.959
Weighted average	0.925	0.075	0.925	0.925	0.959

Table 4: Identification metrics for combination of J48 and MLP (E4 and H1)

Class	TP rate	FP rate	Precision	Recall	AUC
E4	1	0.075	0.93	1	0.963
H1	0.925	0	1	0.925	0.963
Weighted average	0.963	0.038	0.965	0.963	0.963

Table 5: Summary of results

Classes	Classifier(s)	Accuracy based on features (%)
A2H1	MLP	80
	J48	94
	MLP+J48	98
C5H1	MLP	80
	J48	98
	MLP+J48	100
C7H1	MLP	100
	J48	98
	MLP+J48	100

$$\text{Sensitivity} = \text{Recall} = \text{TP rate} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Specificity} = \frac{TN}{FP+TN}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

Metrics for MLP (E4 and H1) and for J48 decision tree (E4 and HI) are given in Table 2 and 3, respectively. Table 4 shows identification metrics for combination of J48 and MLP (E4 and H1). Classifiers of different classes are given in Table 5.

Test of significance: In order to measure improvements and validate the results, the researchers used z-test as the test

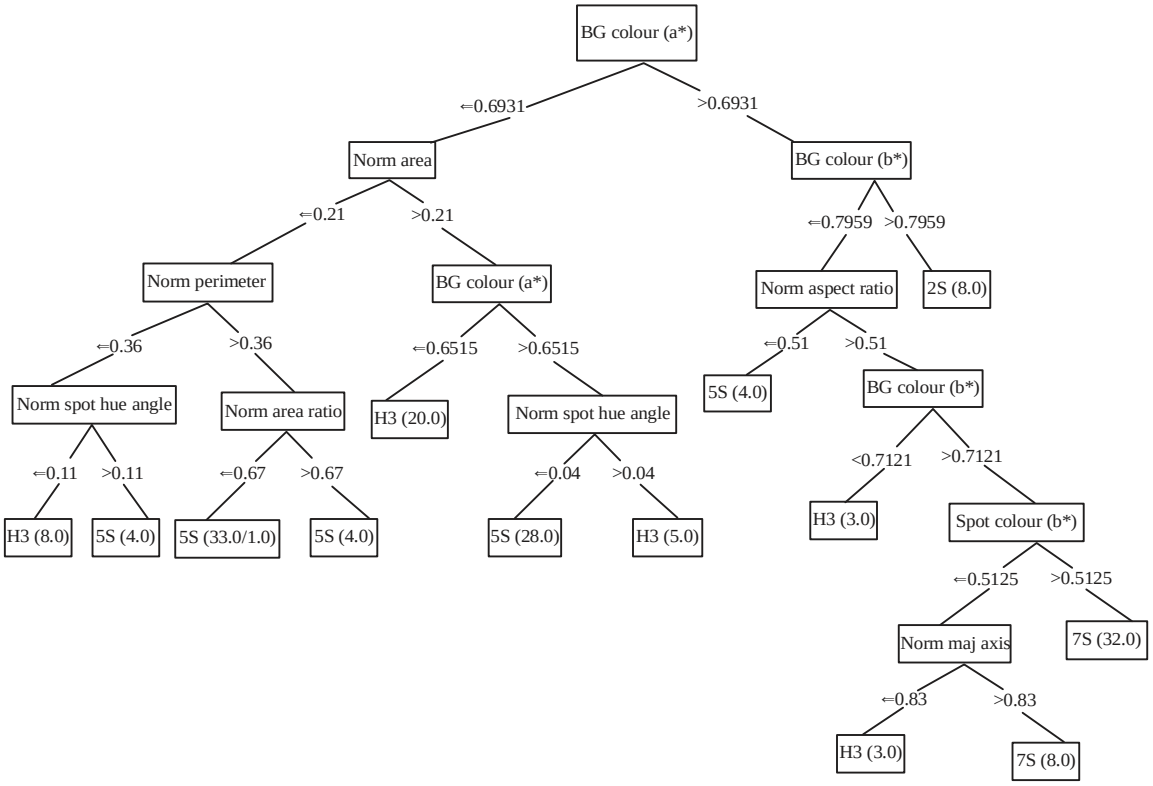

Fig. 6: Decision tree for black-spotted ladybird group

statistics. Here the researchers assumed the population distribution was a standard normal distribution. In term of the features, selected single-feature has been obtained from J48 decision tree test. For others, more than one feature is obtained. Some examples are given in Table 2 and 3. The test of significance will first consider the null hypothesis and the alternative hypothesis[35]. Null hypothesis is denoted as H_0, while the alternative hypothesis is called H_1.

The procedure to carry out the hypothesis test is outlined below:

Step 1: Set up the hypothesis
Step 2: Calculate test statistic, S
Step 3: Determine the critical value, C
Step 4: Check if S is less than or equal to C

If this condition is satisfied, reject the alternative hypothesis.

An example calculation on the procedure is explained using the ladybird scenario. Suppose data is obtained from a population consisting of two ladybird species, the two-spot ladybird (*C. punctata* 2) and *H. axyridis* f. *spectabilis*, each of which containing 50 samples. From earlier test using J48 decision tree, it is agreeable that the useful feature is spot colour (b*). The z-test used the mean value of some samples from the population of the feature, which in this case the mean of spot colour (b*) is used (Table 6). Following the procedure:

Table 6: Features obtained after J48 operations for four species

Features	A2	C5	C7	E4
Spot area				X
Spot perimeter				
Spot max axis length				X
Spot min axis length				
Spot area ratio				X
Spot aspect ratio				
Spot colour a*				
Spot colour b*	X			
Spot hue angle				
Elytra colour a*		X	X	
Elytra colour b*				
Elytra hue angle				

Step 1: Set up the hypothesis: The alternative hypothesis H_1 states that the mean value of spot colour (b*) is significantly different from the population mean. The null hypothesis H_0 will assume otherwise, meaning that there is no significant difference between the spot colour of the 2 species therefore they are the same speciess

Step 2: Calculate test statistic (S): This figure shows how much standard deviation units the samples are from the mean. Twenty random samples are taken from the population. In this study standard deviation σ of the population is 0.02376:

$$\text{Standard Error (SE)} = \sigma/\text{sqrt}(n) = 0.005313$$

$$\text{Test statistic (S)} = \frac{\text{Mean (2-spot)-Mean (other)}}{\text{SE} = \text{mod (-38.3719)}}$$

Step 3: Determine the critical value (C): Use normal distribution table, a two-tailed test and a 5% level of significance will give approximately C = 2.0

Step 4: Check if S is less than or equal to C: Since, S is larger than C, the alternative hypothesis is accepted and the null hypothesis is rejected. On this basis, the samples of ladybirds are significantly different from the expected value, which means they are not the same ladybird species

CONCLUSION

Inter-species separation of ladybirds can be performed using geometrical features or colour features. Due to the 'Curse of dimensionality', there is inherent limitation in the No. of features to arrive at a solution. In this study, J48 decision tree algorithm has indicated which feature is best for classification for the given training set, rather than using all features. This method has saved resources (time and labour). The root node on top of the tree shows the best feature and other nodes show features, which are arranged in descending order of importance. The values appearing between the nodes show the level of contribution and they are useful for generating rules. Results from tests on black-spotted ladybird images show significant improvement compared to MLP. This is an elegant solution for systems which typically involves human-computer interactions, for instance, the automation of dichotomous key for typical species identification.

ACKNOWLEDGMENT

The team would like to thank Majlis Amanah Rakyat (MARA) and UniKL for sponsoring the research studies.

REFERENCES

1. Dallwitz, M.J., 1980. A general system for coding taxonomic descriptions. Taxon, 29: 41-46.
2. Lebbe, J. and R. Vignes, 1998. State of the art in computer-aided identification in biology. Oceanis, 24: 305-317.
3. Chesmore, D., T. Bernard, A.J. Inman and R.J. Bowyer, 2003. Image analysis for the identification of the quarantine pest *Tilletia indica*. OEPP/EPPO Bullet., 33: 495-499.
4. Chesmore, E.D. and G. Monkman, 1994. Automated analysis of variation in Lepidoptera. Entomologist, 113: 171-182.
5. Watson, A.T., M.A. O'Neill and I.J. Kitching, 2004. Automated identification of live moths (Macrolepidoptera) using Digital Automated Identification System (DAISY). Syst. Biodivers., 1: 287-300.
6. White, R.J. and L. Winokur, 2003. Quantitative description and discrimination of butterfly wing patterns using moment invariant analysis. Bull. Entomol. Res., 93: 361-376.
7. Kipling, M.L. and D. Chesmore, 2005. Automated recognition of moth species using image processing and artificial neural networks. Proceedings of the Royal Entomological Society National Meeting, September 12-14, 2005, University of Sussex, Royal Entomological Society.
8. Weeks, P.J.D., I.D. Gauld, K.J. Gaston and M.A. O'Neill, 1997. Automating the identification of insects: A new solution to an old problem. Bull. Entomol. Res., 87: 203-211.
9. Weeks, P.J.D., M.A. O'Neill, K.J. Gaston and I.D. Gauld, 1999. Automating insect identification: Exploring the limitations of a prototype system. J. Applied Entomol., 123: 1-8.
10. Gauld, I.D., M.A. O'Neill and K.J. Gaston, 2000. Driving Miss Daisy: The Performance of an Automated Insect Identification System. In: Hymenoptera: Evolution, Biodiversity and Biological Control, Austin, A. and M. Dowton (Eds.). CSIRO Publishing, Collingwood, Victoria, ISBN: 9780643099104, pp: 303-312.
11. Daly, H.V., K. Hoelmer, P. Norman and T. Allen, 1982. Computer-assisted measurement and identification of honey bees (Hymenoptera: Apidae). Ann. Entomol. Soc. Am., 75: 591-594.
12. Schroder, S., W. Drescher, V. Steinhage and B. Kastenholz, 1995. An automated method for the identification of bee species (Hymenoptera: Apoidea). Proceedings of the International Symposium on Conserving Europe's Bees, April 6-7, 1995, London, UK.
13. Steinhage, V., B. Kastenholz, S. Schroder and W. Drescher, 1997. A Hierarchical Approach to Classify Solitary Bees Based on Image Analysis. In: Mustererkennung 1997, Paulus, E. and F.M. Wahl (Eds.). Springer, Berlin, Germany, ISBN: 978-3-540-63426-3, pp: 419-426.
14. Steinhage, V., S. Schroder, K. Lampe and A.B. Cremers, 2007. Automated Extraction and Analysis of Morphological Features for Species Identification. In: Automated Taxon Identification in Systematics: Theory, Approaches and Applications, MacLeod, N. (Ed.). CRC Press, London, ISBN: 9781420008074, pp: 115-130.

15. Roth, V., A. Pogoda, V. Steinhage and S. Schroder, 1999. Integrating feature-based and pixel-based classification for the automated identification of solitary bees. Proceedings of the Annual Convention of the German Society for Pattern Recognition, September 15-17, 1999, Bonn, pp: 120-129.
16. Yu, D.S., E.G. Kokko, J.R. Barron, G.B. Schaalje and B.E. Gowen, 1992. Identification of ichneumonid wasps using image analysis of wings. Syst. Entomol., 17: 389-395.
17. Angel, P.N., 1999. Multiscale image analysis for the automated localisation of taxonomic landmark points and the identification of speceis of parasitic wasp. Ph.D. Thesis, University of Glamorgan.
18. Dietrich, C.H. and C.D. Pooley, 1994. Automated identification of leafhoppers (Homoptera: Cicadellidae: *Draeculacephala* Ball). Ann. Entomol. Soc. Am., 87: 412-423.
19. Dai, J., 2006. Automated identification of insect taxa using structural image processing. Ph.D. Thesis, Department of Electronics, University of York, England.
20. Mayo, M. and A. Watson, 2006. Automatic species identification of live moths. Proceedings of the 26th SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, December 11-13, 2006, Cambridge, UK., pp: 195-202.
21. Russell, K.N., M.N. Do, J.C. Huff and N.I. Platnick, 2007. Introducing SPIDA-Web: Wavelets, Neural Networks and Internet Accessibility in an Image-Based Automated Identification System. In: Automated Taxon Identification in Systematics: Theory, Approaches and Applications, MacLeod, N. (Ed.). CRC Press, London, ISBN: 9781420008074, pp: 131-152.
22. Clark, J.Y., 2004. Identification of botanical specimens using artificial neural networks. Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, October 7-8, 2004, La Jolla, CA., USA., pp: 87-94.
23. Clark, J.Y., 2007. Plant Identification from Characters and Measurements Using Artificial Neural Networks. In: Automated Taxon Identification in Systematics: Theory, Approaches and Applications, MacLeod, N. (Ed.). CRC Press, London, ISBN: 9781420008074, pp: 207-224.
24. Wu, S.G., F.S. Bao, E.Y. Xu, Y.X. Wang, Y.F. Chang and Q.L. Xiang, 2007. A leaf recognition algorithm for plant classification using probabilistic neural network. Proceedings of the IEEE International Symposium on Signal Processing and Information Technology, December 15-18, 2007, Giza, Egypt, pp: 11-16.
25. Witten, I.H. and E. Frank, 2005. Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edn., Morgan Kaufman, San Francisco, CA., USA., ISBN-13: 9780080477022, Pages: 560.
26. Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten, 2009. The WEKA data mining software: An update. SIGKDD Explorat. Newslett, 11: 10-18.
27. Quinlan, J.R., 1996. Improved use of continuous attributes in C4.5. J. Artif. Intell. Res., 4: 77-90.
28. Omid, M., 2011. Design of an expert system for sorting pistachio nuts through decision tree and fuzzy logic classifier. Expert Syst. Applic., 38: 4339-4347.
29. Kecman, V., 2001. Learning and Soft Computing: Support Vector Machines, Neural Networks and Fuzzy Logic Models. 1st Edn., The MIT Press, Cambridge, MA, USA., ISBN-13: 9780262112550, Pages: 541.
30. Ayob, M.Z. and E.D. Chesmore, 2012. Hybrid feature extractor for Harlequin ladybird identification using color images. Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, May 9-12, 2012, San Diego, CA., USA., pp: 214-221.
31. Negnevitsky, M., 2005. Artificial Intelligence: A Guide to Intelligent Systems. 2nd Edn., Addison Wesley, London, UK., ISBN-13: 9780321204660, pp: 165-217.
32. Lang, R.I.W., 2007. Neural Networks in Brief. In: Automated Taxon Identification in Systematics: Theory, Approaches and Applications, MacLeod, N. (Ed.). CRC Press, London, ISBN: 9781420008074, pp: 47-68.
33. Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn., 30: 1145-1159.
34. Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett., 27: 861-874.
35. Graham, G.A., 2010. Understand Statistics. Hodder & Stoughton, London, UK., ISBN-13: 9781444105049, Pages: 384.