Ecological Informatics 20 (2014) 1–12

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Evaluating the performance of artificial neural networks for the classification of freshwater benthic macroinvertebrates

Henry Joutsijoki [a,*,1], Kristian Meissner [b], Moncef Gabbouj [c], Serkan Kiranyaz [c], Jenni Raitoharju [c], Johanna Ärje [d,1], Salme Kärkkäinen [d], Ville Tirronen [e], Tuomas Turpeinen [f], Martti Juhola [a]

[a] University of Tampere, School of Information Sciences, Kanslerinrinne 1, FI-33014 Tampere, Finland
[b] Finnish Environment Institute, Freshwater Centre, FI-40500 Jyväskylä, Finland
[c] Tampere University of Technology, Department of Signal Processing, P.O. Box 553, FI-33101 Tampere, Finland
[d] University of Jyväskylä, Department of Mathematics and Statistics, FI-40014 Jyväskylä, Finland
[e] University of Jyväskylä, Department of Mathematical Information Technology, FI-40014 Jyväskylä, Finland
[f] University of Jyväskylä, Department of Physics, FI-40014 Jyväskylä, Finland

ARTICLE INFO

Article history:
Received 1 November 2013
Received in revised form 14 January 2014
Accepted 20 January 2014
Available online 25 January 2014

Keywords:
Benthic macroinvertebrates
Artificial neural networks
Multi-Layer Perceptron
Radial Basis Function network
Probabilistic neural network
Classification

ABSTRACT

Macroinvertebrates form an important functional component of aquatic ecosystems. Their ability to indicate various types of anthropogenic stressors is widely recognized which has made them an integral component of freshwater biomonitoring. The use of macroinvertebrates in biomonitoring is dependent on manual taxa identification which is currently a time-consuming and cost-intensive process conducted by highly trained taxonomical experts. Automated taxa identification of macroinvertebrates is a relatively recent research development. Previous studies have displayed great potential for solutions to this demanding data mining application. In this research we have a collection of 1350 images from eight different macroinvertebrate taxa and the aim is to examine the suitability of artificial neural networks (ANNs) for automated taxa identification of macroinvertebrates. More specifically, the focus is drawn on different training algorithms of Multi-Layer Perceptron (MLP), probabilistic neural network (PNN) and Radial Basis Function network (RBFN). We performed thorough experimental tests and we tested altogether 13 training algorithms for MLPs. The best classification accuracy of MLPs, 95.3%, was obtained by two conjugate gradient backpropagation variations and scaled conjugate gradient backpropagation. For PNN 92.8% and for RBFN 95.7% accuracies were achieved. The results show how important a proper choice of ANN is in order to obtain high accuracy in the automated taxa identification of macroinvertebrates and the obtained model can outperform the level of identification which is made by a taxonomist.

1. Introduction

Although water covers the majority of the surface of our planet, only 2.5% of all water resources are fresh water (Oki and Kanae, 2006) making it a resource of utmost importance to humans. Anthropogenic pressures such as pollution and eutrophication are only some examples of the threats that freshwater ecosystems face. Assessing the status of freshwaters is thus essential and has been included into the environmental legislation of many countries. Freshwater ecosystems are inhabited by a diverse spectrum of organisms with varying sensitivities to different anthropogenic pressures. Because of their functional importance and varying sensitivity to environmental pressures benthic macroinvertebrate taxa are commonly used in biomonitoring to detect human-induced changes (Rosenberg and Resh, 1993). For example, many taxa belonging to the orders (Ephemeroptera, Plecoptera and Trichoptera) are known to be sensitive to organic pollution. Benthic macroinvertebrate community composition not only shifts quickly with severe water quality deterioration but also provides a perspective of prevailing pressures on water quality over intermediate to long-term time scales (Rosenberg and Resh, 1993). Although macroinvertebrates are commonly used in aquatic biomonitoring, their manual identification by experts is laborious, time-consuming and expensive.

Gaston and O'Neill (2004) envisaged automated species identification as one possible solution for the mismatch problem of resources and workload that taxonomists are facing. However, automated identification has met not only pure ideological opposition (MacLeod et al., 2010) but also proven to be a challenging problem due to the nature's diversity. Nevertheless, there are a growing number of studies describing successful automated identification of biological object identification such as, e.g., bees (Arbuckle et al., 2001; Gaston and O'Neill, 2004), bird species (Härmä, 2003), butterfly species (Kang et al., 2012) and live moths (Mayo and Watson, 2007).

* Corresponding author. Tel.: +358 503185860; fax: +358 32191001.
E-mail addresses: Henry.Joutsijoki@uta.fi (H. Joutsijoki), Kristian.Meissner@ymparisto.fi (K. Meissner), Moncef.Gabbouj@tut.fi (M. Gabbouj), Serkan.Kiranyaz@tut.fi (S. Kiranyaz), Jenni.Raitoharju@tut.fi (J. Raitoharju), Johanna.Arje@jyu.fi (J. Ärje), Salme.Karkkainen@jyu.fi (S. Kärkkäinen), ville.e.t.tirronen@jyu.fi (V. Tirronen), Tuomas.Turpeinen@jyu.fi (T. Turpeinen), Martti.Juhola@uta.fi (M. Juhola).
[1] This work was supported by the Maj and Tor Nessling Foundation.

1574-9541/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ecoinf.2014.01.004

Automated taxa identification of benthic macroinvertebrates has received some although more limited attention during the last years. In Lytle et al. (2010), Sarpola et al. (2008) the BugID system was introduced for automated identification and processing of benthic macroinvertebrate samples. Moreover, in Joutsijoki (2012a,b, 2013a), Joutsijoki and Juhola (2011a,b, 2012c, 2013b, submitted for publication), Kiranyaz et al. (2010a,b, 2011), Tirronen et al. (2009), Ärje et al. (2010, 2013) the feasibility of automated taxa identification of benthic macroinvertebrates was studied with various methods such as *k*-nearest neighbor (KNN), random forest (RF), random Bayes forest named later random Bayes array (RBA), Radial Basis Function networks (RBFN), Multi-Layer Perceptrons (MLP), linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and different multi-class extensions of Support Vector Machines. Neural networks were also studied in the aforementioned studies but not as thoroughly as in the present study. Despite differences in performance the aforementioned studies have demonstrated the capability of machine learning methods to surpass the accuracy of average manual benthic macroinvertebrate taxa identification and indicate the potential to make the taxa identification more efficient than is manually possible. Automated benthic macroinvertebrate taxa identification does not mean remove the need for human taxonomic expertise completely, as in depth taxonomic knowledge is and will be needed for training of machines. However, automated taxa identification is likely to shift the human effort from routine work to focussing on specialized cases.

The purpose of this research is to explore the suitability and performance of artificial neural networks (ANN) for automated taxa identification of macroinvertebrates. Several learning algorithms have been introduced for Multi-Layer Perceptron networks (MLP) and several variations for Radial Basis Function networks (RBFN) and probabilistic neural networks (PNN) have been made. Nonetheless, there is a lack of an extensive comparison of different ANN methods in the case of automated taxa identification of macroinvertebrates. Coskun and Yildirim (2003) showed in their research concerning image classification that the choice of training algorithm in MLP networks is an important question. Our specific objective is to identify which ANN method is the most effective to classify our benthic macroinvertebrate image dataset. To do so, we perform thorough experimental tests with PNN, RBFN and 13 additional MLP training algorithms.

We organize this paper as follows: In Section 2 we describe the benthic macroinvertebrate image dataset. Section 3 describes the theoretical background of the methods used. Section 4 depicts the design of experiments, how classification was performed and provides the technical details concerning it. Results are expressed in Section 5 through averaged confusion matrices with standard deviations and accuracies with standard deviations. Section 6 provides the discussion and conclusions drawn from our results.

2. Data

The dataset is composed of 1350 benthic macroinvertebrate images from eight different taxa. Table 1 shows taxa specific frequencies in the dataset. All other taxa were identified at the species rank except, *Isoperla* sp., which was identified only to genus. Fig. 1 shows several example images from every taxonomical group included to the dataset so the intra-class variability on each taxonomical group can be seen. Actual image sizes vary but in Fig. 1 images have been scaled such that the height of each image is 0.75 in.

Table 1
Frequencies and percentages of benthic macroinvertebrate classes in the dataset.

Taxonomical group		Frequencies	%
Baetis rhodani	BAE	116	8.6
Diura nanseni	DIU	129	9.6
Heptagenia sulphurea	HEP	172	12.7
Hydropsyche pellucidulla	PEL	102	7.6
Hydropsyche siltalai	SIL	271	20.1
Isoperla sp.	ISO	311	23.0
Rhyacophila nubila	RHY	83	6.1
Taeniopteryx nebulosa	TAE	166	12.3

The images were obtained using a HP Scanjet 4850 desktop scanner with Vuescan 8.4.57 software. The samples were put into a transparent container filled with alcohol placed on the glass panel of the scanner. To avoid lightning artifacts a cardboard cover was used and the lightning conditions of the room were stabilized. Movement artifacts were minimized by setting the scanner on the most stable surface in the room (i.e. the floor). The samples were scanned at 1600 dpi resolution in RGB space wih the same settings for all scans.

The segmentation of the samples was performed in grayscale space produced by averaging the RGB channels. The uneven illumination was corrected by using the rolling ball algorithm (Sternberg, 1983) with a manually determined rolling ball radius (the radius was chosen to be a bit larger than the radius of the objects at largest). The binarization was performed by thresholding and the sample labeling by assuming them to be non-touching. The label image was used to separate each sample into an individual RGB image (sample image) and a binary mask for further analysis.

From each individual image a total of 25 ImageJ (2013) default features were extracted. Here, we restricted all our analysis to data from a subset of 15 geometrical and intensity based features only (see also Ärje et al., 2010; Joutsijoki (2012a, 2013a); Joutsijoki and Juhola (2011a,b, 2013b, submitted for publication; Kiranyaz et al., 2010a,b, 2011). Geometrical features included in analysis were Area, Perimeter, Width, Height, Feret's Diameter, Major, Minor and Circularity. Intensity based features were Mean, Standard Deviation, Mode, Median, Integrated Density, Kurtosis and Skewness.

Geometrical features define the shape of the object and are determined from the mask. Geometrical features used were defined as follows based on ImageJ (2013):

- Area is the number of the sample pixels contained in the mask.
- Perimeter is calculated from the mask such that edge pixels get value 1 and corners $\sqrt{2}$.
- Width and Height are the width and height of the smallest rectangle that encloses the sample in mask.
- Feret's Diameter is the longest distance between two points inside the sample.
- Major and minor axes are the length of the primary and secondary axis of the best fitting smallest ellipse (ellipse that has the same 0th, 1st and 2nd image moments as the mask).
- Circularity is defined as $4\pi \times \frac{\text{Area}}{\text{Perimeter}^2}$.

Intensity based features are calculated from the RGB image using the mask to define the pixels belonging to the object. The features are defined for each individual channel and grayscale image. In this work only the grayscale features were utilized. According to ImageJ (2013) intensity based features used in this research were defined as follows:

- Mean is the average gray value of the object pixels.
- Standard deviation defines the variation from the mean intensity value.
- Mode is the most frequently occurring gray value of the object pixels.
- Median is the centermost value (or average of two most centermost values) when arranging all the gray values of the object pixels into one ordered vector.
- Integrated density is the sum of the gray values of the object pixels.
- Kurtosis defines the “peakedness” of the histogram. It is the fourth standardized moment of the intensity values of the object pixels.
- Skewness defines the symmetry of the histogram. It is the third standardized moment of the intensity values of the object pixels.