Vol. 441: 185–196, 2011
doi: 10.3354/meps09387

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published November 15

Bayesian model for semi-automated zooplankton classification with predictive confidence and rapid category aggregation

Lin Ye[1,2], Chun-Yi Chang[1], Chih-hao Hsieh[1,3,*]

[1]Institute of Oceanography, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan
[2]State Key Laboratory of Freshwater Ecology and Biotechnology, Institute of Hydrobiology, The Chinese Academy of Sciences, Wuhan 430072, PR China
[3]Institute of Ecology and Evolutionary Biology, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan

ABSTRACT: Zooplankton play a critical role in aquatic ecosystems and are commonly used as bioindicators to assess anthropogenic and climate impacts. Nevertheless, traditional microscope-based identification of zooplankton is inefficient. To overcome the low efficiency, computer-based methods have been developed. Yet, the performance of automated classification remains unsatisfactory because of the low accuracy of recognition. Here we propose a novel framework for automated plankton classification based on a naïve Bayesian classifier (NBC). We take advantage of the posterior probability of NBC to facilitate category aggregation and to single out objects of low predictive confidence for manual re-classifying in order to achieve a high level of final accuracy. This method was applied to East China Sea zooplankton samples with 154 289 objects, and the Bayesian automated zooplankton classification model showed a reasonable overall accuracy of 0.69 in unbalanced and 0.68 in balanced training for 25 planktonic and 1 aggregated non-planktonic categories. More importantly, after manually checking 17 to 38% of the objects of low confidence (depending on how one defines 'low confidence'), the final accuracy increased to 0.85–0.95 in the unbalanced training case, and after checking 18 to 42% of the low-confidence objects in the balanced training case, the final accuracy increased to 0.84–0.95. Our semi-automated approach is significantly more accurate than automated classifiers in recognizing rare categories, thereby facilitating ecological applications by improving the estimates of taxa richness and diversity. Our approach can make up for the deficiencies in current automated zooplankton classifiers and facilitates an efficient semi-automated zooplankton classification, which may have a broad application in environmental monitoring and ecological research.

KEY WORDS: Automated classification · Naïve Bayesian classifier · Predictive confidence · Rapid category aggregation · Zooplankton community · ZooScan

INTRODUCTION

Zooplankton play an important role in aquatic ecosystems as (1) indicators of environmental changes (Beaugrand et al. 2002, Hays et al. 2005, Beaugrand & Kirby 2010) and (2) an important link between primary production and higher trophic levels (Suthers & Rissik 2008, Sarmento et al. 2010). Therefore, zooplankton have been routinely collected during many environmental monitoring programs around the world (Hsieh et al. 2005, Tadokoro et al. 2005, Edwards et al. 2009, Heine & Koslow 2009). Traditionally, zooplankton specimens are identified and counted using microscopy. However,

*Corresponding author. Email: chsieh@ntu.edu.tw

the traditional approach is labor-intensive and time-consuming, limiting our ability to analyze zooplankton samples and understand processes controlling aquatic ecosystem dynamics (Suthers & Rissik 2008, Gorsky et al. 2010). Therefore, an important issue in aquatic ecology concerns how to improve the efficiency in plankton analysis and make the data more comparable at different spatial and temporal scales (Wiebe & Benfield 2003, Perry et al. 2004).

Automated or semi-automated computer-aided systems can improve the efficiency in sample analyses and may even deliver more accurate and consistent results than human taxonomists in some cases (Culverhouse et al. 2003, Benfield et al. 2007, MacLeod et al. 2010). Recently, a large amount of effort has been invested in developing automated plankton identification systems (Ortner et al. 1979, Balfoort et al. 1992, Boddy et al. 2000, Hu & Davis 2005, Luo et al. 2005, Benfield et al. 2007, Sosik & Olson 2007, Gasparini 2009, Gislason & Silva 2009). ZooScan integrated with imaging software (ZooProcess) and classification software (Plankton Identifier) has been suggested to be a useful system for automated zooplankton image acquisition, biomass calculation, and taxonomic classification (Grosjean et al. 2004, Gorsky et al. 2010). ZooScan has a standard protocol for zooplankton image acquisition, so that the measured variables for each of the extracted objects can be inter-comparable between different machines. Although ZooScan has been shown to be an efficient and reliable system for enumeration and measurement of particles, its performance in automated classification remains unsatisfactory (Gorsky et al. 2010).

Due to similarity among or incomplete information for digitalized specimens, not all digitalized images can be recognized correctly by machine-based methods (Edwards & Morse 1995, Gaston & O'Neill 2004). Therefore, a critical question is how to pick out the objects most likely being misclassified in automated classification for further processing (e.g. manual reclassification) (Grosjean et al. 2004). Here, we propose a solution by estimating the predictive confidence for automated classification based on Bayesian probability. With predictive confidence, we can reject objects of low confidence from the automated classification and pass them on to human experts for further checking and re-classification. After objects of low confidence are manually checked and reclassified, the final accuracy based on the semi-automatic method may be greatly improved.

Another limitation in current automated zooplankton classifications is the low efficiency in optimizing the number of categories. Based on the literature (Fernandes et al. 2009), the best number of categories in automated zooplankton classification is determined by repeating different iterations with different combinations of categories. After that, the end-user needs to choose the best one among different iterations by considering the balance between taxonomic resolution and number of categories, which is time-consuming and generally results in loss of taxonomic resolution as a trade-off for recognition accuracy (Fernandes et al. 2009). The approach based on Bayesian probability can overcome this deficiency, because the theoretical probability of any aggregated category is simply computed by summing the posterior probabilities of the categories aggregated (Wang et al. 2007). In other words, the classifier can provide the results of automated classification for any possible taxonomic level (any level of aggregation) and the corresponding predictive confidence with only a single calculation at the most detailed level.

In the present study, we propose a framework based on the Bayesian theorem for automated zooplankton classification with an emphasis on predictive confidence and rapid category aggregation. We used a 4 yr zooplankton dataset from the East China Sea to: (1) test the performance of a naïve Bayesian classifier (NBC) in automated zooplankton classification; (2) estimate the predictive confidence from the empirical relationship between Bayesian probabilities and recognition accuracies in different categories; and (3) test the final performance of our proposed semi-automated method in zooplankton classification, assuming that human experts can correctly re-classify all rejected objects of low predictive confidence.

METHODS

Bayesian probabilistic model

The key issue is to develop a method to estimate posterior probability and the associated predictive confidence for each automated classification. Here, we adopted a Bayesian probabilistic model. In the Bayesian theorem (Duda et al. 2000, Bolstad 2007), the probability of the hypothesis is defined as the likelihood multiplied by the prior probability. The likelihood is determined by the data, while the prior probability is a given value based on existing knowledge. In the present study, the classifier is based on NBC (Duda et al. 2000). For each object, the posterior probability of each potential category is defined as follows:

$$P(C_j \mid F) = \frac{p(F \mid C_j)P(C_j)}{\sum_{j=1}^{n} p(F \mid C_j)P(C_j)} \quad (1)$$

where $P(C_j \mid F)$ is the posterior occurrence probability of zooplankton category C_j given the input feature vectors $\{F_1,\ldots, F_n\}$. For each object, a vector of posterior probabilities for predetermined zooplankton categories is obtained, and the predicted category is determined as the category with the maximum value of posterior probability (Fig. 1). The expression $p(F \mid C_j)$ is the likelihood probability density function (PDF) for the category C_j, which is estimated using kernel density estimation (Chiu 1996, Jones et al. 1996). $P(C_j)$ is the prior probability of C_j in all potential categories $\{C_1, \ldots, C_n\}$, which is simply taken as the proportion of the categories in a given community: (the number of C_j)/(total number of objects). For each object, the machine-predicted category and the associated posterior probability is thus determined. The Bayesian probabilistic model was implemented based on the Naive-Bayes functions in the Statistics Toolbox of MatLab.

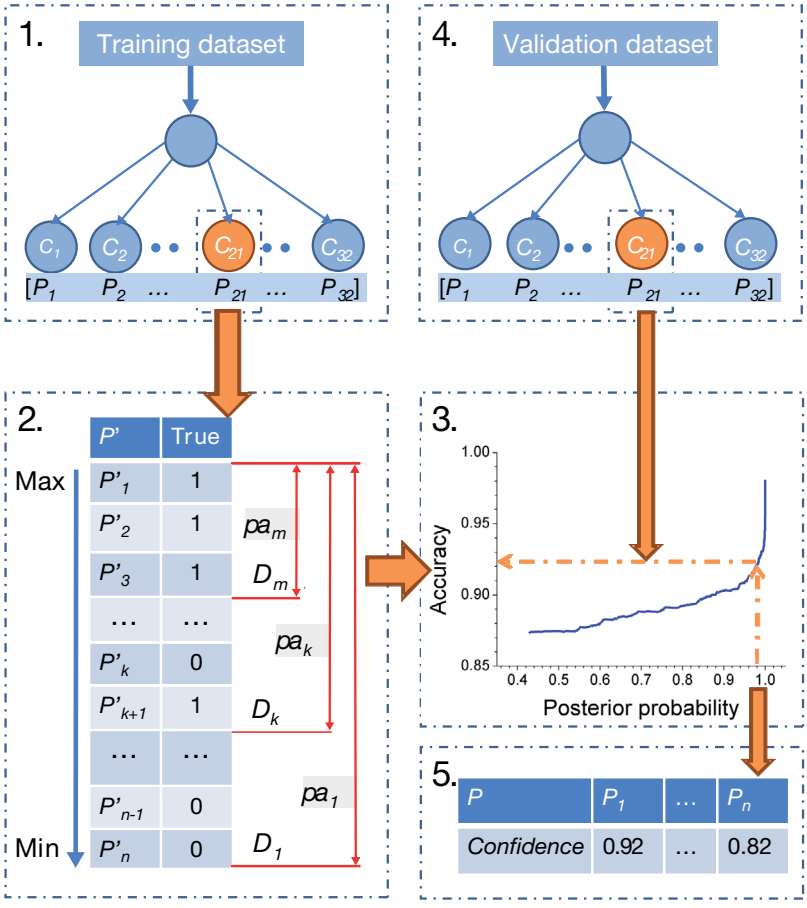

Fig. 1. Determining predictive confidence. Step 1: For each image, a vector of posterior probabilities $\{P_1, \ldots, P_{32}\}$ for the predetermined 32 categories are generated with the trained classifier (using the training dataset), and the predicted group is determined as the category that has the maximum value of posterior probability. For example, if a total of n images in the training dataset were classified into Category 21 (C_{21}), we can get an $n \times 2$ matrix (posterior probability, true or false predicted) for C_{21}. Step 2: After ranking the posterior probabilities (P') of these n images, we obtain the cumulative predictive accuracy (termed predictive confidence) for every unique posterior probability, $\{D_1,\ldots, D_m\}$, for C_{21}. Step 3: With $\{D_1,\ldots, D_m\}$ and the corresponding predictive confidence $\{pa_1, \ldots, pa_m\}$, we can generate the empirical relationship between posterior probabilities and predictive confidences (generated by a large amount of real data from the training dataset) for C_{21}. Step 4: For each novel image from the validation dataset, the naïve Bayesian classifier can classify the image into a specific category (in this example, C_{21}) with a posterior probability. Step 5: We project this value of posterior probability onto the empirical curve generated from Step 3. As such, we can determine the predictive confidence for a specific classification using the empirical distribution generated in Step 3

At this stage, the predicted category for an object has been determined by the value of posterior probability from NBC. Next, we need to calculate the predictive confidence of the object for that category. To do this, we estimated the predictive confidence of a classification from the empirical relationship between posterior probabilities and recognition accuracy based on the training dataset. As shown in Fig. 1, we collected all objects in the training dataset that are predicted to be the same category (say, Category X) and obtain the empirical distribution of posterior probabilities for Category X. Among these objects, certainly some are correctly predicted but others are not. We then rank (from big to small) those posterior probabilities for the objects classified as Category X and plot the cumulative recognition accuracy versus the ranked posterior probability. By doing so, we can obtain the predictive confidence for any classification based on the empirical relationship between posterior probability and cumulative recognition accuracy (accuracy for the collection of objects being classified) of each category in the training dataset as follows:

$$F(x, j) = \begin{bmatrix} 0, x < D_{1,j} \\ pa_{k,j}, D_{k,j} \le x < D_{k+1,j}; k = 1, 2, \ldots, m-1 \\ pa_{m,j}, x \ge D_{m,j} \end{bmatrix} \quad (2)$$

where $F(x, j)$ is the empirical predictive confidence for a specific posterior probability x in the predicted category j. $\{D_{1,j}, \ldots, D_{m,j}\}$ represents m distinct values in the original vector of posterior probability $\{P_{1,j}, \ldots, P_{n,j}\}$ of category j in the training dataset with an ascending order. And $pa_{k,j}$ is the empirical recognition accuracy of category j when considering all objects with the posterior probability equal to or above $D_{k,j}$ in the training dataset (Fig. 1).

Now, we have constructed the classifier and the empirical relationship between recognition accuracy and posterior probabilities for each category. Based on the posterior probability of NBC, any novel image from the validation set can be classified into a certain category with the predictive confidence (in terms of the potential accuracy that one can anticipate) based on the empirical equation (Eq. 2) of that category.

Framework for semi-automated classification

Our framework for a Bayesian probabilistic model for automated zooplankton classification and category aggregation is presented in Fig. 2. In our framework, firstly, the classifier was constructed at the most detailed taxonomic level with the training set. For each training object, the classifier provided the posterior probability for each pre-defined category at the most detailed taxonomic level. The end-user can then determine the accepted taxonomic level in classification (or aggregation), and the posterior probability values of that level can be simply calculated by aggregating related categories. With the posterior probability, the predictive confidence can be estimated using the empirical equation in Eq. (2). Thus, when a novel image is examined, it can be classified into a specific category with the posterior probability and predictive confidence.

Importantly, with the estimated predictive confidence, we can single out the objects of confidence lower than the accepted level (determined by the end-user) and pass them on to human experts for manual re-classification. In addition, with the posterior probability, we can carry out category aggregation (for example, into a lower taxonomical resolution or based on the user's research purposes) to achieve a better accuracy.

Data availability and application

The performance of the proposed framework in automated zooplankton classification was tested

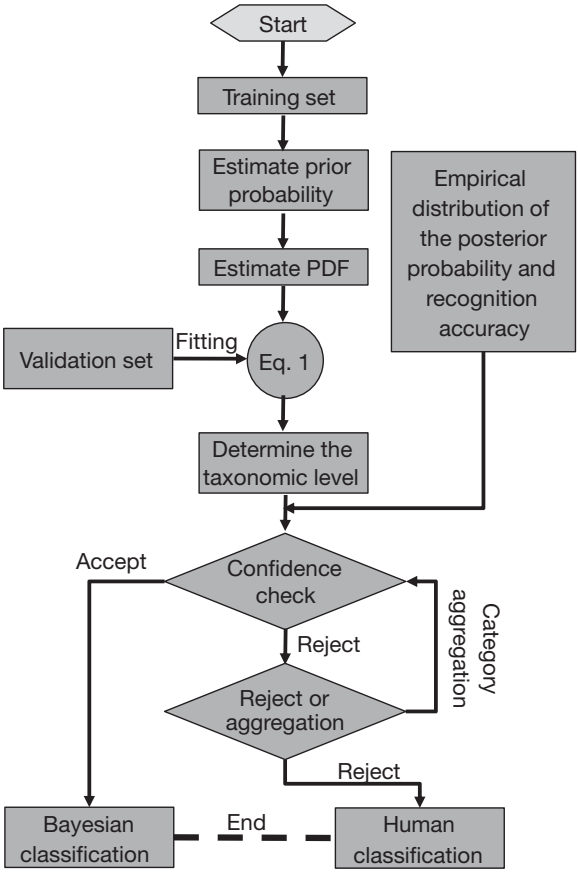

Fig. 2. Framework of the Bayesian probabilistic model for automated zooplankton classification and category aggregation. Eq. (1): see 'Materials and methods'. PDF: probability density function

using the East China Sea zooplankton dataset, a 4 yr dataset with 154 289 objects. Zooplankton samples were collected in the East China Sea from 2006 to 2009 using a 330 µm mesh net with a mouth of 160 cm diameter (sampling sites and methods detailed in Supplement 1 at www.int-res.com/articles/suppl/m441p185_supp.pdf). Pretreatment and scanning of the samples followed the procedures in the ZooScan user manual. We scanned >1500 images per sample (site). The operation of scanning per sample took <30 min. After scanning, objects were extracted and measured using Zooprocess (version 5.07) automatically (see Gorsky et al. 2010 for descriptions of measured variables). Then, the objects were sorted into different taxonomic categories manually by human experts for developing the automated classification model and for validation, following previous studies (e.g. Grosjean et al. 2004, Gorsky et al. 2010, García-Comas et al. 2011). A total of 154 289 objects were sorted into 25 planktonic categories and 3 non-

living categories with the sorting procedures following Grosjean et al. (2004) and Fernandes et al. (2009). The 5 most abundant planktonic categories were Copepoda_like (copepods that could not be further identified by human experts based on ZooScan images), Copepod Cyclopoida, Copepod Temoridae, Gelatinous forms, and Invertebrate egg; they contributed 47.4, 7.9, 5.1, 4.9, and 4.8% of the total abundance of zooplankton, respectively. The non-planktonic particles summed to 57 954 objects (37.6%).

A preliminary analysis found that a large proportion of the non-planktonic category 'Detritus' would contaminate other categories in the automated classification. To investigate the issue of contamination from different kinds of 'Detritus', the category 'Detritus' was divided into 5 sub-groups according to their size, because size is reported as an important factor in machine-based classification (Bell & Hopcroft 2008). Finally, we arrived at a total of 25 planktonic and 7 non-planktonic categories (see Fig. 3).

We first constructed an unbalanced training set by randomly selecting 50% of the objects in each category from the whole dataset, and used the other 50% of the objects as the validation set. We further constructed a balanced training set by randomly selecting 300 objects in each category within the unbalanced training set while using the same validation set as the unbalanced training case. We used 300 objects because a previous study reported that 200 to 300 training vignettes per category are sufficient (Gorsky et al. 2010). If the number of objects in any rare category is below 300, we extra-scanned other zooplankton samples in the East China Sea and collected specific objects to supplement the balanced training set.

The classifier was trained with feature data extracted by the ZooScan system (see Gorsky et al. 2010 for the measured variables). The empirical relationship between the posterior probability and cumulative recognition accuracy was estimated using the training set. Because zooplankton composition varied in space, the site-specific prior probabilities were used in the present study.

In model validation, the likelihood $p(F|C_j)$ of each input vector was calculated according to the pre-calibrated PDF. Then with site-specific prior probabilities, the posterior probability $P(C_j|F)$ of each predefined category was calculated following Eq. (1). As described above (section 'Bayesian probabilistic model'), a novel object can be classified into a certain category associated with posterior probability and predictive confidence. Based on the predictive confidence for each prediction, one can decide whether the item should be accepted or rejected for human re-classification or passed to category aggregation.

Performance assessment

The performance of the model was evaluated using the true positive rate (recall), false positive rate (contamination), and precision based on the confusion matrix (Kohavi & Provost 1998). The true positive rate is the proportion of individuals in the dataset being correctly classified. The false positive rate is the proportion of individuals incorrectly classified as a certain category, and precision is the proportion of individuals belonging to a category that is correctly recognized.

Note that, in evaluating the performance of our classifier, if an object belonging to the category of 'Copepoda_like' (a group of all kinds of copepods that could not be further identified by human experts) was predicted to belong to any specific category of copepods, we would consider such a classification to be correct, because any category of copepods is within the scope of 'Copepoda_like'. The reasoning is that since even human experts could not correctly identify it, we would not expect our machine to identify it.

Justification for using NBC

We compared the performance of NBC with other classifiers provided in the ZooScan integrated system (Gasparini 2009, Gorsky et al. 2010), including 5-NN, S-SVC linear, S-SVC RBF, Random Forest, C4.5, and Multilayer Perceptron. All algorithms were trained with the same training set as NBC. Then the performances of these classifiers were also tested with the same validation dataset as for the NBC. The comparison indicates that NBC provides a reasonable overall and taxon-specific accuracy, although overall accuracy is not the best (Supplement 2 at www.int-res.com/articles/suppl/m441p185_supp.pdf). Nevertheless, we choose NBC because it provides posterior probability, which facilitates quick category aggregation and estimation of predictive confidence. We note that a relatively high-accuracy algorithm named 'Discriminant vector forest' was reported by Grosjean et al. (2004), which can provide a severity rating as some sort of measure of predictive confidence. However, the algorithm was not detailed in the report and cannot be disclosed due to the copyright issue (P. Grosjean pers. comm.).

RESULTS

Performance of NBC

In validation, the NBC achieves a reasonable performance (Fig. 3) (Supplements 2 & 3 at www.int-res.com/articles/suppl/m441p185_supp.pdf). The overall recall accuracy of NBC for the 25 planktonic categories was 0.71 in unbalanced and 0.70 in balanced training. In terms of category-specific recall accuracy (Fig. 3a), both unbalanced and balanced training of NBC was unstable for rare taxa (Categories 1 to 8, each containing <0.5% of all objects in our validation set), with accuracy ranging from 0.18 to 0.87. For abundant taxa, recall accuracy is relatively stable, with the value ranging from 0.57 to 0.87, except for Category 17 (Copepod Eucalanidae), most of which were recalled as other copepods (see confusion matrices in Supplement 3).

Judging from the false positive rates, the contamination of NBC is very low in both unbalanced and balanced training, with the value below 0.02 for most categories (Fig. 3b). However, the low contamination still caused a low precision for many taxa (Categories 1 to 18, each containing <2% of all objects in our validation set), and only the 7 most abundant taxa (Categories 19 to 25) had a relative stable value for precision, which ranged from 0.51 to 0.96.

For the 7 non-planktonic categories, the overall recall accuracy and precision was 0.56 and 0.74, respectively, in unbalanced training, and 0.52 and 0.70, respectively, in balanced training. As shown in Fig. 3, both recall accuracy and precision decreased with the increase of their size. Because there is no ecological significance in distinguishing different non-planktonic objects, we aggregated them into 1 non-planktonic group to achieve better accuracy by summing the Bayesian probability of these categories. After aggregation, the final overall recall accuracy and precision of the aggregated non-planktonic category was improved to 0.67 and 0.88, respectively, in unbalanced training and 0.66 and 0.88, respectively, in balanced training.

In summary, the overall accuracy of 25 planktonic categories and 1 aggregated non-planktonic category with unbalanced and balanced training was 0.69 and 0.68, respectively. Such performance is similar to the algorithms provided in Plankton Identifier of the ZooScan integrated system (Supplement 2). Note that unbalanced and balanced training achieved similar performance for the 25 planktonic and 7 non-planktonic categories (Supplements 2, 3 & 4 at www.int-res.com/articles/suppl/m441p185_supp.pdf).

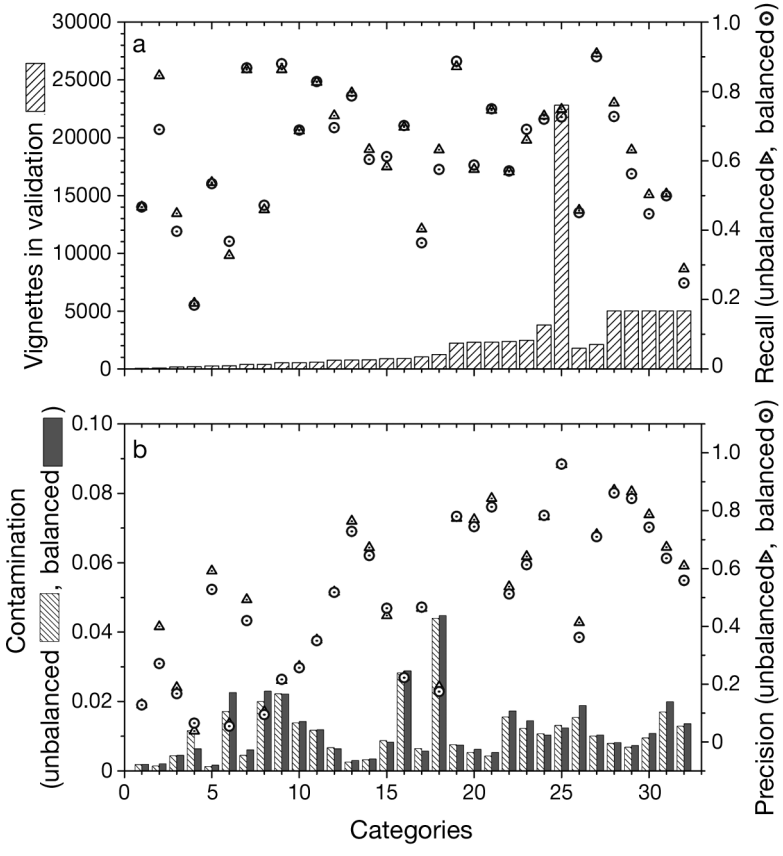

Fig. 3. (a) Number of objects and recall (true positive accuracy) and (b) contamination (false positive rate) and precision in validation at the most detailed taxonomic level with the balanced and unbalanced training set using the naïve Bayesian classifier. Categories: 1: Ostracod Cypridinidae, 2: Cladocera *Podon*, 3: Copepod Sapphirinidae, 4: Larva_zoea, 5: Pteropod, 6: Larva_nauplii, 7: Euphausiids, 8: Amphipoda, 9: Larva_veliger, 10: Other Ostracoda, 11: Copepod Oithonidae, 12: Luciferidae, 13: Other chaetognath, 14: Chaetognath *Flaccisagitta*-like, 15: Appendicularia, 16: Larva_furcilia, 17: Copepod Eucalanidae, 18: Larva_calyptopis, 19: Cladocera *Evadne*, 20: Copepod Euchaetidae, 21: Invertebrate egg, 22: Gelatinous forms, 23: Copepod Temoridae, 24: Copepod Cyclopoida, 25: Copepoda_like, 26: Shadow, 27: Fiber, 28: Detritus_1, 29: Detritus_2, 30: Detritus_3, 31: Detritus_4, 32: Detritus_5

Semi-automatic classification based on predictive confidence

The empirical relationship between posterior probability and cumulative recognition accuracy of each category

showed that recognition accuracy increases with posterior probability both in unbalanced and balanced training (Supplement 4), which indicates that the objects with a higher posterior probability are more likely to be correctly classified. However, the acceptable value of the posterior probability for a pre-defined accuracy varied dramatically among categories (Supplement 4). For example, in unbalanced training, a posterior probability of 0.83 can achieve a recognition accuracy of 0.60 for Category 17 (Copepod Eucalanidae), while a posterior probability of 1.00 cannot attain a recognition accuracy of 0.60 for some categories (such as Categories 1, 3, 6, 9, and 10). To overcome this variation, we determined the accepted objects based on predictive confidence rather than using their posterior probability. We pre-defined a predictive confidence, investigated the empirical distribution of posterior probabilities versus cumulative recognition accuracy for each category (Supplement 4), and decided the limit of posterior probability required to achieve the pre-determined predictive confidence for each category.

Table 1. Ratio of accepted objects to all objects in the validation dataset (Ratio), overall accuracy of accepted objects (Acc. 1), overall accuracy of rejected objects (Acc. 2), and the final overall accuracy, assuming that all rejected objects were correctly re-classified (Acc. 3) in unbalanced and balanced training at different accepted confidence levels (statistics based on 25 planktonic and 1 non-planktonic categories)

Confidence level	Unbalanced training				Balanced training			
	Ratio	Acc. 1	Acc. 2	Acc. 3	Ratio	Acc. 1	Acc. 2	Acc. 3
0.10	0.99	0.70	0.03	0.71	0.98	0.69	0.03	0.70
0.20	0.96	0.72	0.04	0.73	0.95	0.72	0.05	0.73
0.30	0.88	0.77	0.10	0.80	0.87	0.77	0.10	0.80
0.40	0.85	0.80	0.11	0.83	0.84	0.79	0.10	0.83
0.50	0.83	0.82	0.12	0.85	0.82	0.81	0.11	0.84
0.60	0.79	0.84	0.15	0.87	0.77	0.84	0.16	0.88
0.70	0.73	0.87	0.22	0.90	0.69	0.87	0.25	0.91
0.80	0.69	0.89	0.27	0.92	0.67	0.89	0.27	0.92
0.90	0.62	0.92	0.33	0.95	0.58	0.92	0.36	0.95
0.95	0.48	0.95	0.46	0.98	0.44	0.95	0.47	0.98
0.96	0.46	0.95	0.47	0.98	0.41	0.96	0.49	0.98
0.97	0.40	0.96	0.51	0.99	0.18	0.97	0.62	0.99
0.98	0.15	0.97	0.64	1.00	0.11	0.98	0.65	1.00

Our approach successfully estimated the predictive confidence in automated zooplankton classification; that is, the validation showed that the objects with the lowest confidence have the highest probability of being misclassified both in balanced and unbalanced training (Table 1). For example, with a confidence value of 0.10, the accepted objects only achieved an overall accuracy (25 planktonic and 1 aggregated non-planktonic categories) of 0.70 and 0.69 in unbalanced and balanced training, respectively. When the accepted confidence levels increased from 0.10 to 0.98, the overall accuracy in accepted objects was improved from 0.70 to 0.97 in unbalanced training and from 0.69 to 0.98 in balanced training. Moreover, if all rejected objects (the objects of predictive confidence below user-accepted levels) were correctly re-classified by human experts, the corresponding final precision could be improved to 1.00. Suppose the end-user-accepted confidence was 0.80, then among the objects 31% would be rejected in the unbalanced training case (33% of the objects rejected in balanced training). And if all rejected objects were correctly re-classified, the final overall accuracy could reach 0.92 in unbalanced and balanced training, with a significant improvement (2-tailed paired *t*-test, $p < 0.001$) for both recall accuracy and precision for all categories, especially for rare taxa (Fig. 4) (Supplement 3). Generally, the performances of unbalanced and balanced training sets are similar, suggesting that 300 images for each category in training should suffice for our purposes in the semi-automatic approach. Note that manually re-checking 30% of images (~500 images) took less than 1 h.

DISCUSSION

Performance of the semi-automated Bayesian model

In the present study we have proposed a semi-automated Bayesian model for plankton classification with the guidance of predictive confidence. Currently, although many efforts have been invested in optimizing different machine learning algorithms in automated zooplankton classification, the recognition accuracy of machine-based methods is still not good enough to replace manual processing of zooplankton samples (Bell & Hopcroft 2008, Fernandes et al. 2009, Irigoien et al. 2009, Gorsky et al. 2010). To improve the efficiency of manual processing, Grosjean et al. (2004) pioneered a semi-automated method by calculating the severity rating to determine and re-classify 'suspected' objects. Even though it was not detailed in that report, we suspect that the severity parameter was estimated by synthesizing

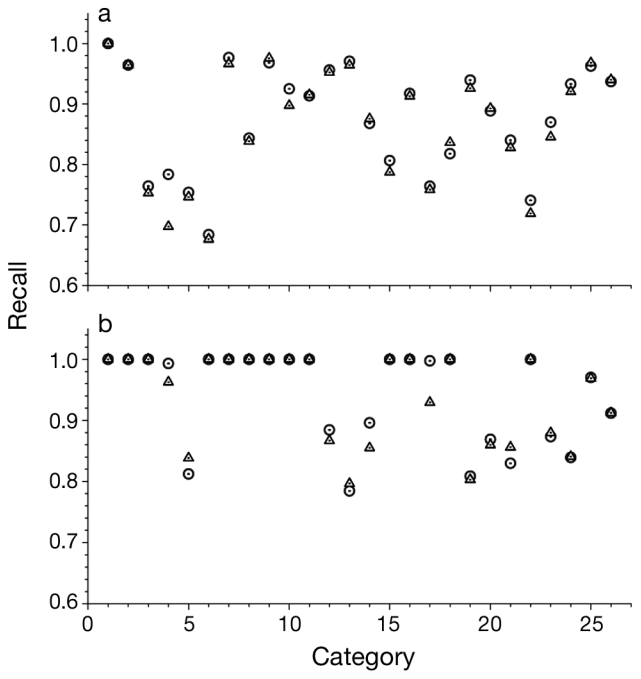

Fig. 4. Final category-specific (a) recall accuracy and (b) precision in balanced (⊙) and unbalanced (△) training cases, assuming that the objects with a low confidence (<0.80) are correctly re-classified. See Fig. 3 legend for the names of Categories 1 to 25. Category 26 is the aggregated non-plankton

the results of linear discriminate analysis, learning vector quantization, and random forest classifiers. According to their results, final accuracy could be improved from 0.75 to a maximum of 0.94 by re-classifying all suspected objects (Grosjean et al. 2004). So far, this approach is the only reported method for estimating predictive confidence in machine-based zooplankton classification. Unfortunately, no technical details were provided in the published paper (Grosjean et al. 2004), and the algorithm is not available to the public due to copyright issues (P. Grosjean pers. comm.).

Here, we take advantage of the posterior probability provided by the NBC to estimate predictive confidence. The application on the East China Sea zooplankton dataset suggested that our method for estimating predictive confidence may be an improvement over the severity rating in several ways, by comparing the published data (Grosjean et al. 2004). Firstly, our approach is more efficient. Predictive confidence in our model is directly 'born' with the classifier, while Grosjean's method needs to synthesize the results from several classifiers to estimate confidence. Secondly, our approach is more accurate in estimating misclassified objects. In Grosjean's case, final accuracy was improved by 15 percentage points (from 0.75 to 0.90) after 31% of the objects were re-classified, while with our approach using unbalanced training, the final accuracy was improved by 23 percentage points (from 0.69 to 0.92) after 31% of the suspected objects were re-classified. And in balanced training, the final accuracy was improved by 24 percentage points (from 0.68 to 0.92) after 33% of the suspected objects were re-classified. Moreover, in our model, the maximum accuracy of semi-automated classification can reach 1.00, whereas in Grosjean's case, the best accuracy can only be improved to 0.94, because 6% of the objects that were simultaneously misclassified by the 3 classifiers used (Grosjean et al. 2004) cannot be singled out. Thirdly, the predictive confidence devised here is more than just a parameter indicating how likely an object can be accurately classified by the machine in a relative sense; rather, it links directly to recognition accuracy (Fig. 1).

To further investigate the severity rating approach, we calculated the consistency index using the 6 classifiers in the ZooScan integrated system to estimate confidence (known as a bagging approach) using the East China Sea zooplankton data. We found that this kind of severity rating (Table 2) is less efficient and accurate compared with the semi-automatic NBC approach. In the unbalanced training case, the final accuracy only improved by 17 percentage points (from 0.79 to 0.96) after 38% of the suspected objects were re-classified. And in balanced training cases, the final accuracy improved by 25 percentage points (from 0.73 to 0.98) after

Table 2. Severity rating of the 6 classifiers in the ZooScan integrated system based on balanced and unbalanced training. Ratio of accepted objects to all objects in the validation dataset (Ratio), overall accuracy of accepted objects (Acc. 1), overall accuracy of rejected objects (Acc. 2), and the final overall accuracy, assuming that all rejected objects were correctly re-classified (Acc. 3). The severity rating indicates the number of classifiers having the same prediction. –: no objects rejected by severity rating

Severity rating	Unbalanced training				Balanced training			
	Ratio	Acc. 1	Acc. 2	Acc. 3	Ratio	Acc. 1	Acc. 2	Acc. 3
1	1.00	0.79	–	0.79	1.00	0.73	–	0.73
2	1.00	0.79	0.00	0.79	1.00	0.73	0.00	0.73
3	0.98	0.80	0.24	0.80	0.93	0.76	0.28	0.78
4	0.89	0.84	0.39	0.86	0.76	0.83	0.39	0.87
5	0.77	0.88	0.48	0.91	0.58	0.90	0.50	0.94
6	0.62	0.93	0.57	0.96	0.36	0.94	0.61	0.98

64% of the suspected objects were re-classified. Moreover, this approach is time-consuming because it requires the execution of several algorithms and the synthesizing of results.

Efficiency of Bayesian approach in category aggregation

Our proposed method is flexible and efficient in aggregating categories. The traditional approach of category aggregation requires the re-running of the training process with new combinations of categories if end-users change the taxonomic resolution (Fernandes et al. 2009, Gorsky et al. 2010). By contrast, the Bayesian classifier can provide classification results for any potential level of aggregation by aggregating the posterior probabilities of related categories, with only 1 training at the most detailed taxonomic level (Wang et al. 2007). Single training in our Bayesian approach is certainly more efficient than multiple training as provided in the previous method. This feature is especially useful for plankton research because plankton ecologists often need different taxonomic resolution for different research purposes. For instance, in biodiversity research, ecologists generally require detailed species information to calculate different kinds of biodiversity indices (Edwards & Morse 1995, Irigoien et al. 2004), whereas in functional-group studies, ecologists need to combine different species with the same ecological traits into a single functional unit to study their relationships with environmental changes (Quéré et al. 2005, Romanuk et al. 2010).

Ecological applications

Our approach may have broad application in environmental monitoring and ecological research. As has been shown in Gorsky et al. (2010), the ZooScan integrated system can provide rapid digital archiving, enumeration, size measurement, and data sharing of plankton specimens in a non-destructive way. Indeed, if we accept all machine predictions (without further human checking), the calculated total abundance and biomass based on outputs of the different classifiers are generally acceptable in both unbalanced and balanced training (Supplement 5 at www.int-res.com/articles/suppl/m441p185_supp.pdf). However, when we considered category-specific abundance estimation, the results for rare taxa were generally not accurate (Supplement 3). In such cases, the semi-automatic approach can significantly improve the estimation (Supplement 3).

By contrast, the calculated taxa richness (number of planktonic groups defined in this study) and Shannon diversity (calculated from group composition) is not consistent with the observed values (Supplement 6 at www.int-res.com/articles/suppl/m441p185_supp. pdf). As a consequence, the low accuracy for rare taxa hampers some ecological applications of the system, such as biodiversity-related research. Our semi-automatic approach can overcome this difficulty by improving accuracy for both rare and abundant taxa with only some fraction of suspected objects re-classified by human experts. For example, after re-checking 31 and 33% of suspected objects (predictive confidence <0.80) in unbalanced and balanced training cases respectively, the accuracy of final taxa richness and Shannon diversity calculated from the output of our method was improved significantly (2-tailed paired *t*-test, $p < 0.001$). In the unbalanced training case (Fig. 5), before re-classifying the suspected objects, the reliability (R^2 of observed versus estimated from machine) of richness and Shannon diversity estimate is 0.52 and 0.73, respectively. Notably, after 31% of the suspected objects were re-classified, the reliability (R^2) of the richness and Shannon diversity estimate was improved dramatically to 0.91 and 0.97, respectively. In the balanced training case (Fig. 6), after 33% of the suspected objects were re-classified, the reliability of the richness and Shannon diversity estimate was improved from 0.55 to 0.91 and from 0.72 to 0.96, respectively.

One might argue that the taxonomic resolution obtained from the ZooScan integrated system is too low for diversity research. However, according to our ZooScan data in the East China Sea, we found that taxa richness and Shannon diversity increases from high latitudes to low latitudes (Supplement 7 at www.int-res.com/articles/suppl/m441p185_supp. pdf), which is consistent with the general global pattern of latitudinal distribution of zooplankton diversity (Hillebrand 2004, Irigoien et al. 2004, Beaugrand et al. 2010). In fact, the usefulness of lower taxonomic-resolution data in diversity and environmental research has been suggested in research across a large spatiotemporal scale (e.g. Törnblom et al. 2011, Llope et al. 2011). Here, we suggest that the capability and efficiency of the ZooScan system in collecting consistent taxonomic data is still useful in diversity research, especially in research encompassing a large spatiotemporal scale when synthesizing zooplankton data from different organizations is needed. Moreover, in the future, when high taxonomic-resolution data are available from improved

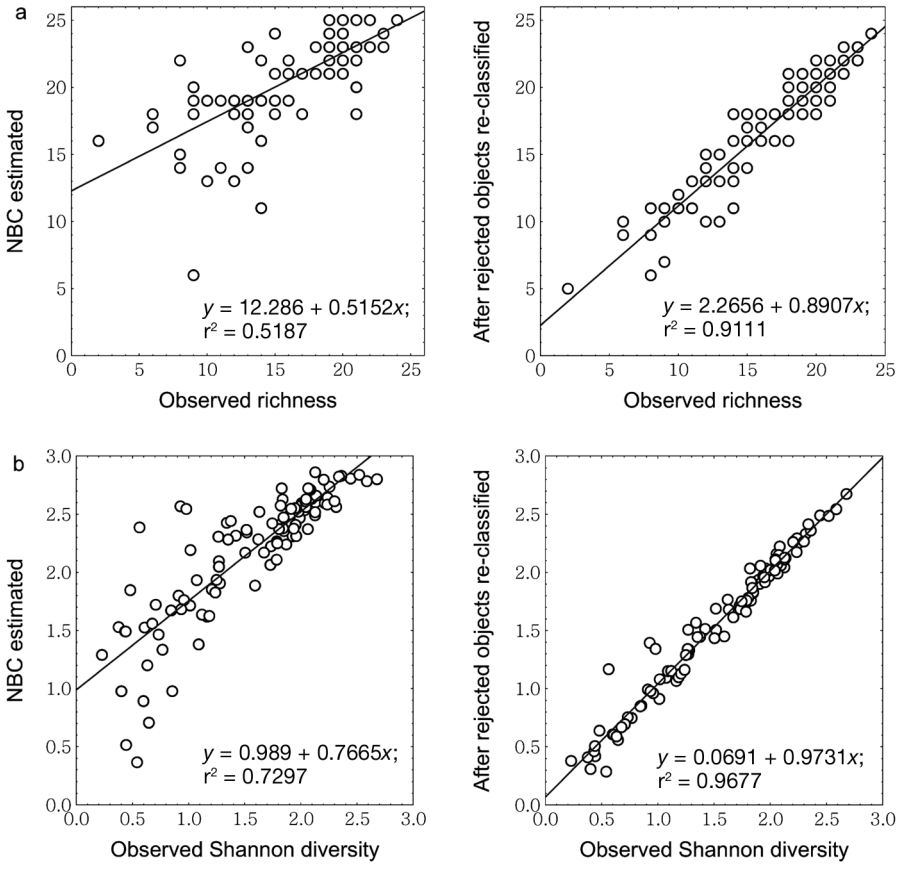

Fig. 5. Estimates of (a) taxa richness and (b) Shannon diversity were improved significantly after the objects of low confidence (<0.80) were reclassified in the unbalanced training case. NBC: naïve Bayesian classifier

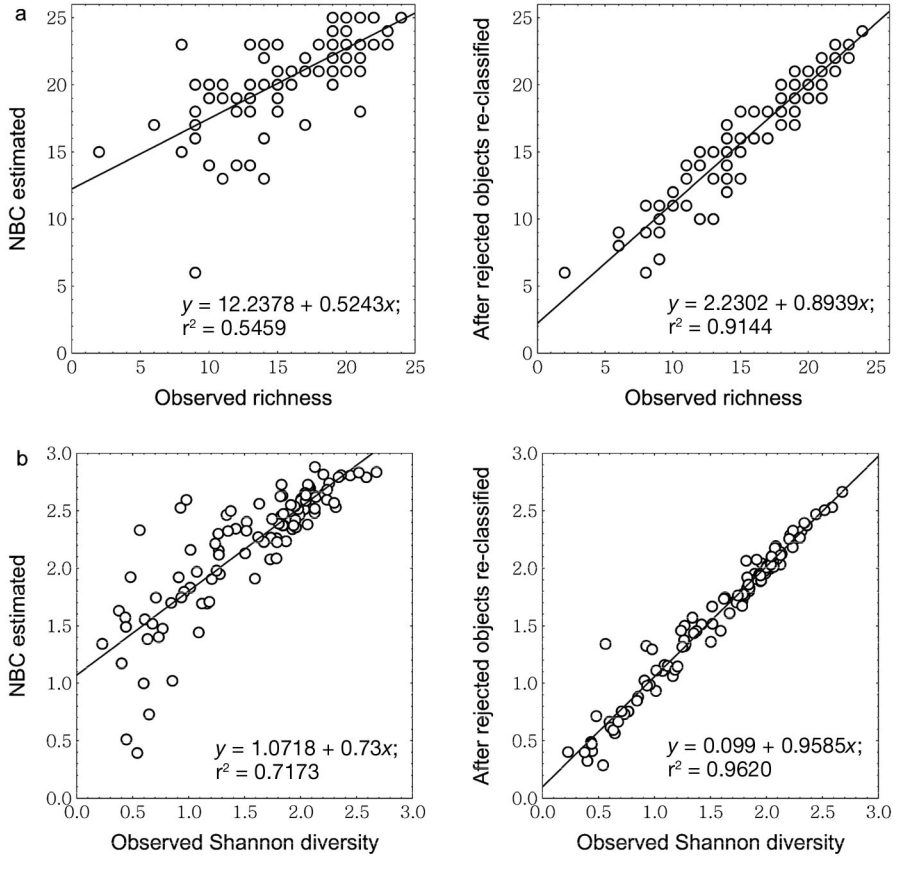

Fig. 6. Estimates of (a) taxa richness and (b) Shannon diversity were improved significantly after the objects of low confidence (<0.80) were reclassified in the balanced training case. NBC: naïve Bayesian classifier

automated systems (e.g. with improved image resolution or feature extraction), our method can be readily applied to obtain more accurate and highly resolved classification.

CONCLUSIONS

We have proposed a novel approach for automated zooplankton classification with an emphasis on predictive confidence and rapid category aggregation based on posterior probability within a Bayesian framework. With the estimated predictive confidence, we can single out objects of low confidence and pass them on for human identification or for further category aggregation to improve accuracy. The application of the proposed framework on the East China Sea zooplankton samples showed that the Bayesian semi-automated zooplankton classification model has good performance with either unbalanced or balanced training. The performances of unbalanced (50% of the data) and balanced (300 images for each category only) training sets are similar, suggesting that 300 images for each category in training should suffice for our purposes in the semi-automatic approach. Our method of defining predictive confidence is more efficient and accurate for estimating misclassified objects than the previously reported method. Our semi-automated approach achieves significant improvement in recognition accuracy, which improves the ecological applications of automated plankton classification, such as an improved estimation of taxa richness and Shannon diversity. In addition, our framework is adaptive. Our approach can be applied to the plankton data collected from FlowCAM (Alvarez et al. 2011), ZooImage (Bachiller & Fernandes 2011), as well as other imaging systems, which may provide higher-resolution taxonomic data (as these techniques can capture more detailed taxonomic characteristics). Furthermore, other classification algorithms could be modified to fit in our Bayesian framework to provide predictive confidence and flexibility in category aggregation. Currently, automatic classification systems have been developed for many fields, such as phytoplankton, insects, plants, etc. (MacLeod et al. 2010). Our approach may help improve classification accuracy in these systems.

Acknowledgements. We are grateful to the Computer and Information Networking Center, National Taiwan University for the support of high-performance computing facilities. We also thank N. Grimm, E. Marquis, and P. Ho for their valuable comments, which greatly improved the quality of this article. This study was supported by a grant for Frontier and Innovative Research of National Taiwan University, National Science Council of Taiwan, and Major Science and Technology Program for Water Pollution Control and Treatment (2009ZX07528-003).

LITERATURE CITED

Alvarez E, Lopez-Urrutia A, Nogueira E, Fraga S (2011) How to effectively sample the plankton size spectrum? A case study using FlowCAM. J Plankton Res 33: 1119–1133

Bachiller E, Fernandes JA (2011) Zooplankton image analysis manual: automated identification by means of scanner and digital camera as imaging devices. Rev Invest Mar 18:17–37

Balfoort HW, Snoek J, Smits JRM, Breedveld LW, Hofstraat JW, Ringelberg J (1992) Automatic identification of algae: neural network analysis of flow cytometric data. J Plankton Res 14:575–589

Beaugrand G, Kirby RR (2010) Climate, plankton and cod. Glob Change Biol 16:1268–1280

Beaugrand G, Reid PC, Ibanez F, Lindley A, Edwards M (2002) Reorganization of North Atlantic marine copepod biodiversity and climate. Science 296:1692–1694

Beaugrand G, Edwards M, Legendre L (2010) Marine biodiversity, ecosystem functioning, and carbon cycles. Proc Natl Acad Sci USA 107:10120–10124

Bell JL, Hopcroft RR (2008) Assessment of ZooImage as a tool for the classification of zooplankton. J Plankton Res 30:1351–1367

Benfield MC, Grosjean P, Culverhouse PF, Irigoien X and others (2007) RAPID: Research on Automated Plankton Identification. Oceanography (Wash DC) 20:172–187

Boddy L, Morris CW, Wilkins MF, Al-Haddad L, Tarran GA, Jonker RR, Burkill PH (2000) Idenfification of 72 phytoplankton species by radial basis function neural network analysis of flow cytometric data. Mar Ecol Prog Ser 195: 47–59

Bolstad WM (ed) (2007) Introduction to Bayesian statistics, 2nd edn. John Wiley & Sons, Hoboken, NJ

Chiu ST (1996) A comparative review of bandwidth selection for dernel density estimation. Statist Sinica 6: 129–145

Culverhouse PF, Williams R, Reguera B, Herry V, González-Gil S (2003) Do experts make mistakes? A comparison of human and machine identification of dinoflagellates. Mar Ecol Prog Ser 247:17–25

Duda RO, Hart PE, Stork DG (2000) Bayesian decision theory. In: Duda RO, Hart PE, Stork DG (eds) Pattern classification. Wiley-Interscience, New York, NY, p 20–83

Edwards M, Morse DR (1995) The potential for computer-aided identification in biodiversity research. Trends Ecol Evol 10:153–158

Edwards M, John AWG, Johns DG, McQuatters-Gollop A (2009) CPR annual report 2008. Continuous Plankton Recorder (CPR) Survey, Sir Alister Hardy Foundation for Ocean Science, Plymouth. www.sahfos.ac.uk/annual_reports/Annual2008.pdf

Fernandes JA, Irigoien X, Boyra G, Lozano JA, Inza I (2009) Optimizing the number of classes in automated zooplankton classification. J Plankton Res 31:19–29

García-Comas C, Stemmann L, Ibanez F, Berline L and

others (2011) Zooplankton long-term changes in the NW Mediterranean Sea: decadal periodicity forced by winter hydrographic conditions related to large-scale atmospheric changes? J Mar Syst 87:216–226

Gasparini S (2009) Plankton Identifier [software]. www.obs-vlfr.fr/~gaspari/Plankton_Identifier/

Gaston KJ, O'Neill MAO (2004) Automated species identification: Why not? Philos Trans R Soc Lond B Biol Sci 359: 655–667

Gislason A, Silva T (2009) Comparison between automated analysis of zooplankton using ZooImage and traditional methodology. J Plankton Res 31:1505–1516

Gorsky G, Ohman MD, Pichearl M, Gasparini S and others (2010) Digital zooplankton image analysis using the ZooScan integrated system. J Plankton Res 32:285–303

Grosjean P, Picheral M, Warembourg C, Gorsky G (2004) Enumeration, measurement, and identification of net zooplankton samples using the ZOOSCAN digital imaging system. ICES J Mar Sci 61:518–525

Hays GC, Richardson AJ, Robinson C (2005) Climate change and marine plankton. Trends Ecol Evol 20: 337–344

Heine JN, Koslow T (2009) CalCOFI Reports, Vol 50. California Cooperative Oceanic Fisheries Investigations (CalCOFI), Pacific Grove, CA. http://calcofi.org/publications/calcofireports/vol50/Calcofi_vol.50_finals.pdf

Hillebrand H (2004) On the generality of the latitudinal diversity gradient. Am Nat 163:192–211

Hsieh CH, Chen CS, Chiu TS (2005) Composition and abundance of copepod and ichthyoplankton in the Taiwan Strait (western North Pacific) in relation to seasonal marine conditions. Mar Freshw Res 56:153–161

Hu Q, Davis C (2005) Automatic plankton image recognition with co-occurrence matrices and Support Vector Machine. Mar Ecol Prog Ser 295:21–31

Irigoien X, Huisman J, Harris RP (2004) Global biodiversity patterns of marine phytoplankton and zooplankton. Nature 429:863–867

Irigoien X, Fernandes JA, Grosjean P, Denis K, Albaina A, Santos M (2009) Spring zooplankton distribution in the Bay of Biscay from 1998 to 2006 in relation with anchovy recruitment. J Plankton Res 31:1–17

Jones MC, Marron JS, Sheather SJ (1996) Progress in data-based bandwidth selection for kernel density estimation. Comput Stat 11:337–381

Kohavi R, Provost F (1998) Glossary of terms. Mach Learn 30:271–274

Llope M, Licandro P, Chan KS, Stenseth NC (2011) Spatial variability of the plankton trophic interaction in the North Sea: a new feature after the early 1970s. Glob Change Biol doi:10.1111/j.1365-2486.2011.02492.x

Luo T, Kramer K, Goldgof DB, Hall LO, Samson S, Remsen A, Hopkins T (2005) Active learning to recognize multiple types of plankton. J Mach Learn Res 6:589–613

MacLeod N, Benfield M, Culverhouse P (2010) Time to automate identification. Nature 467:154–155

Ortner PB, Cummings SR, Aftering RP (1979) Silhouette photography of oceanic zooplankton. Nature 277:50–51

Perry RI, Batchelder HP, Mackas DL, Chiba S, Durbin E, Greve W, Verheye HM (2004) Identifying global synchronies in marine zooplankton populations: issues and opportunities. ICES J Mar Sci 61:445–456

Quéré CL, Harrison SP, Prentice IC, Buitenhuis ET and others (2005) Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. Glob Change Biol 11:2016–2040

Romanuk TN, Vogt RJ, Young A, Tuck C, Carscallen MW (2010) Maintenance of positive diversity-stability relations along a gradient of environmental stress. PLoS ONE 5:e10378

Sarmento H, Montoya JM, Vázquez-Domínguez E, Vaqué D, Gasol JM (2010) Warming effects on marine microbial food web processes: How far can we go when it comes to predictions? Philos Trans R Soc B Biol Sci 365: 2137–2149

Sosik HM, Olson RJ (2007) Automated taxonomic classification of phytoplankton sampled with imaging-in-flow cytometry. Limnol Oceanogr Methods 5:204–216

Suthers IM, Rissik D (2008) Plankton: a guide to their ecology and monitoring for water quality. CSIRO Publishing, Collingwood

Tadokoro K, Chiba S, Ono T, Midorikawa T, Saino T (2005) Interannual variation in *Neocalanus* biomass in the Oyashio waters of the western North Pacific. Fish Oceanogr 14:210–222

Törnblom J, Roberge JM, Angelstam P (2011) Rapid assessment of headwater stream macroinvertebrate diversity: an evaluation of surrogates across a land-use gradient. Fundam Appl Limnol 178:287–300

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol 73:5261–5267

Wiebe PH, Benfield MC (2003) From the Hensen net toward four-dimensional biological oceanography. Prog Oceanogr 56:7–136

Editorial responsibility: Marsh Youngbluth, Fort Pierce, Florida, USA

Submitted: April, 19, 2011; Accepted: September 5, 2011
Proofs received from author(s): November 10, 2011