Multiclass Support Vector Machines with Example-dependent Costs Applied to Plankton Biomass Estimation

Pablo González, Eva Álvarez, Jose Barranquero, Jorge Díez, Rafael González-Quirós, Enrique Nogueira, Ángel López-Urrutia and Juan José del Coz*

***Abstract*—In many applications, the mistakes made by an automatic classifier are not equal, they have different costs. These problems may be solved using a cost-sensitive learning approach. The main idea is not to minimize the number of errors, but the total cost produced by such mistakes. This paper presents a new multiclass cost-sensitive algorithm, in which each example has attached its corresponding misclassification cost. Our proposal is theoretically well-founded and is designed to optimize cost-sensitive loss functions. This research was motivated by a real-world problem, the biomass estimation of several plankton taxonomic groups. In this particular application, our method improves the performance of traditional multiclass classification approaches that optimize the accuracy.**

***Index Terms*—Cost-sensitive learning, plankton recognition, example-dependent costs, SVM, kernel methods.**

I. Introduction

In supervised learning, the learner is given a set of training examples, each one formed by a feature vector and the desired output. The goal is to infer a model, called classifier when the possible outputs are a finite set of values, able to predict the output of unseen examples. Usually, the quality of the classifier is measured by the number of mistakes made, the fewer the better. Despite being one of the most useful learning paradigms, this approach does not fit properly with several real applications. This is the case, for instance, of those decision support systems for approving bank loan applications or predicting medical diseases. In these applications, the different types of mistakes are not equal, they have a different *cost*. From the point of view of the bank, the cost of mistakenly classifying a costumer depends on the amount of money borrowed; incorrectly diagnosing a healthy person as being sick is preferable to the opposite. Learning classifiers that consider the actual costs of their decisions should lead to improve the quality of these applications.

The techniques aimed to address this sort of problems are known under the name of *cost-sensitive* (CS) learning [1]. The core idea is to induce models that reduce the total cost. Turney defines a taxonomy of the different types of costs [2] that can be considered. This paper focuses in the most important one: the cost of classification errors. From that point of view, two types of CS problems can be distinguished: those that have class-dependent costs [3] (for instance, medical diagnosis problems) and others that have example-dependent costs [4] (e.g. loan application approval). In this paper we present an example-dependent cost method to deal with the plankton biomass estimation problem.

The study of plankton is crucial because i) plankton is the base of the food chain that sustains life in oceans [5], and ii) its ecosystems play a crucial role in many biogeochemical cycles, including the oceans' carbon cycle. Scientists study plankton by means of surveys that employ nets and other samplers to collect specimens. In the past, such surveys were manually processed by trained microscopists, limiting their temporal and spatial resolution. For that reason, the Scientific Committee recognized the importance of developing automatic plankton identification systems creating a working group (http://www.scor-wg130.net). Three basic elements are needed to build these systems: i) a plankton sampling device to automatically obtain high resolution images from the plankton samples,

This research has been supported by Ministerio de Economía y Competitividad (TIN2011-23558), and FICYT (IB09-059-C2).

P. González, J. Barranquero, J. Díez and J.J. del Coz are with the Artificial Intelligence Center, University of Oviedo, Spain.

E. Álvarez, R. González-Quirós, E. Nogueira and Á. López-Urrutia, are with the Oceanographic Centre of Gijón (IEO), Spain.

J.J. del Coz is the corresponding author (juanjo@aic.uniovi.es).

Manuscript received ; revised

ii) computer vision methods to process such images, and iii) a classification algorithm able to identify the species of each organism. The goal is to provide answers to questions like: Which is the abundance (number of individuals) of each taxonomic group? What is the total amount of biomass of each group? Interestingly, current systems [6] are better designed for answering the first question, because they are based on classifiers that maximizes the accuracy, without considering the biomass of the individuals.

Our proposal is to use CS learning to accurately estimate the total biomass of plankton species. Taking into account that sample devices give us an approximation of organisms' biomass, we can use this information and reformulate our learning task following a CS approach with example-dependent costs. The misclassification cost of each individual will be its biomass and our problem will consist of minimizing the amount of biomass misclassified. The expected result will be that our approach should provide better predictions for the plankton biomass estimation problem, while previous methods [6] should perform better for the abundance problem.

Support Vector Machines (SVM) [7], [8] was originally designed to solve binary classification tasks. Following such formulation, new methods have been proposed to build multiclass SVMs. Mainly, there are two types of approaches. The first one decomposes the multiclass task into a set of binary problems; this approach includes algorithms such as one-vs-all (OVA) [8], one-vs-one (OVO) [9], or those using decision trees [10]. The second alternative considers all data in a single optimization problem [11]. This paper applies several SVMs to the plankton biomass estimation problem. In fact, the main contribution of this work is the development of a new multiclass CS SVM. The proposed method is the extension of Crammer & Singer formulation [11] to a CS setting with example-dependent costs. This new algorithm is efficient enough for the application at hand. The second contribution, from a learning perspective, is the comparison in a real problem between non-CS and CS SVM variants, and also between decomposition and single optimization strategies. The conclusion of our study is that, in this case, it is better to apply a CS algorithm using a single optimization approach.

II. Cost-sensitive learning

Being $\mathcal{X}$ the input space and $\mathcal{Y} = \{1, \ldots, k\}$ a finite set of classes, a CS multiclass task is defined by a training set $\mathcal{S} = \{(\boldsymbol{x}_1, y_1, c_1), \ldots, (\boldsymbol{x}_n, y_n, c_n)\}$, obtained from an unknown probability distribution $Pr(\mathcal{X}, \mathcal{Y}, \mathbb{R}^+)$. In terms of CS learning, the value $c_i > 0$ associated with each example $\boldsymbol{x}_i$ represents the penalty of misclassifying it. In our problem c_i stands for the biomass of organism $\boldsymbol{x}_i$.

The aim of the learning task defined by $\mathcal{S}$ is to find a hypothesis h from the input space to the output space; in symbols $h : \mathcal{X} \longrightarrow \mathcal{Y}$, optimizing the *expected prediction performance (or risk)* on samples $\mathcal{S}'$ independently and identically distributed (i.i.d.) according to the distribution $Pr(\mathcal{X}, \mathcal{Y}, \mathbb{R}^+)$:

$$R^{\delta_{CS}}(h) = \int \delta_{CS}(h(\boldsymbol{x}), y, c)\; d(Pr(\boldsymbol{x}, y, c)), \quad (1)$$

in which $\delta_{CS}(h(\boldsymbol{x}), y, c)$ is a CS loss function that measures the penalty due to the prediction $h(\boldsymbol{x})$ when the real class of object $\boldsymbol{x}$ is y and the misclassification cost is c. The straightforward definition for δ_{CS} in our setting is

$$\delta_{CS}(h(\boldsymbol{x}), y, c) = c[\![h(\boldsymbol{x}) \neq y]\!], \quad (2)$$

where $[\![\pi]\!]$ is 1 when the predicate π is true and 0 otherwise. This definition implies that δ_{CS} is the extension of zero-one loss function, $\delta_{0/1}(h(\boldsymbol{x}), y) = [\![h(\boldsymbol{x}) \neq y]\!]$, to the CS case. Notice that correct decisions of h involving examples with a higher cost are favored. Some kind of average is usually performed in order to measure the cost over a set of examples. The most common one is the loss function that returns the average cost per example,

$$\Delta_{AC}(h, \mathcal{S}') = \frac{1}{n} \sum_{\boldsymbol{x}_i \in \mathcal{S}'} \delta_{CS}(h(\boldsymbol{x}_i), y_i, c_i), \quad (3)$$

being n the number of examples in the testing set $\mathcal{S}'$. In this paper, we prefer a more informative loss function for our target application:

$$\Delta_{PMC}(h, \mathcal{S}') = \frac{1}{\sum_{\boldsymbol{x}_i \in \mathcal{S}'} c_i} \sum_{\boldsymbol{x}_i \in \mathcal{S}'} \delta_{CS}(h(\boldsymbol{x}_i), y_i, c_i), \quad (4)$$

that is, the proportion of misclassified costs. For instance, in our application, the idea is to measure the proportion of biomass that is misclassified. Obviously, both metrics are closely connected: the only difference between them is that, given a concrete testing set, they use a different constant in the denominator. The learning method presented in this paper is able to optimize both loss functions.

III. Learning methods

A. Multiclass classification algorithms

As we stated before, there are two groups of approaches to build multiclass SVM: decomposition and single optimization strategies. One method of each kind has been applied in this study: OVO [9], because it obtains better performance [12], and Crammer & Singer method [11], for being more efficient than others.

OVO approach defines $k(k-1)/2$ binary problems, where the l-vs-m problem implies subsets $\mathcal{S}_l$ and $\mathcal{S}_m$ which contain examples of classes l and m respectively. Using soft-margin binary SVMs, OVO solves the following kind of optimization problems[1]:

$$\begin{aligned} \min_{\boldsymbol{w}_{lm},\boldsymbol{\xi}^{lm}} \quad & \frac{1}{2}\langle \boldsymbol{w}_{lm}, \boldsymbol{w}_{lm}\rangle + C\sum_{y_i\in\{l,m\}} \xi_i^{lm}, \\ \text{s.t.} \quad & \langle \boldsymbol{w}_{lm}, \boldsymbol{x}_i\rangle \geq +1-\xi_i^{lm}, \quad \text{if } \boldsymbol{x}_i\in\mathcal{S}_l, \\ & \langle \boldsymbol{w}_{lm}, \boldsymbol{x}_i\rangle \leq -1+\xi_i^{lm}, \quad \text{if } \boldsymbol{x}_i\in\mathcal{S}_m, \\ & \xi_i^{lm}\geq 0, \quad \forall \boldsymbol{x}_i\in\mathcal{S}_l\cup\mathcal{S}_m, \end{aligned} \tag{5}$$

where factor C allows to control the amount of regularization and ξ_i are the slack variables used to avoid overfitting and to cope with non-separable cases. For an example $\boldsymbol{x}_i$, the output of each model $\boldsymbol{w}_{lm}$ is counted as one vote for the predicted class l or m. The final decision is the highest-voted class.

In Crammer & Singer method, a model $\boldsymbol{w}_l$ is induced for each class l following the one-vs-rest approach. The key difference is that all of them, $\boldsymbol{W}=\{\boldsymbol{w}_l : l\in\{1,\ldots,k\}\}$, are learned together:

$$\begin{aligned} \min_{\boldsymbol{W},\boldsymbol{\xi}} \quad & \frac{1}{2}\sum_{l=1}^{k}\langle \boldsymbol{w}_l, \boldsymbol{w}_l\rangle + C\sum_{i=1}^{n}\xi_i, \\ \text{s.t.} \quad & (\langle \boldsymbol{w}_{y_i}, \boldsymbol{x}_i\rangle - \langle \boldsymbol{w}_r, \boldsymbol{x}_i\rangle) \geq e_i^r - \xi_i, \\ & \forall i=1,\ldots,n \quad \forall r\in\{1,\ldots,k\}, \end{aligned} \tag{6}$$

where e_i^r is 1 when $r\neq y_i$ and 0 otherwise. Notice that the number of constrains might be large, especially for those problems with many classes. Still, efficiency is achieved since most of the constraints are inactive, due to the fact that the set of constraints of each example $\boldsymbol{x}_i$ shares one single slack variable ξ_i. The class predicted by the algorithm will be determined following the winners-takes-all rule:

$$h(\boldsymbol{x}_i) = \underset{l\in\{1,\ldots,k\}}{\operatorname{argmax}}\langle \boldsymbol{w}_l, \boldsymbol{x}_i\rangle. \tag{7}$$

[1]For ease of reading, bias term will be always omitted. It could be included by adding a feature of constant value to each $\boldsymbol{x}_i$

The main advantage of this approach over the previous one is that a specific loss function can be optimized. In this formulation, obtained by softening the constraints using the continuous hinge loss function, the zero-one loss function is optimized for the whole multiclass classifier. This is particularly interesting for our purposes, since we can modify this method for optimizing a CS loss function, like Δ_{PMC} (4).

B. Cost-sensitive algorithms

The learning methods described before can be modified to work within the CS learning paradigm. As we shall prove, the obtained CS algorithms are as efficient as their non-CS counterparts.

The CS version of OVO approach is based on the CS binary classifier presented in [13], in which the authors additionally provides some nice theoretical results, establishing a risk bound for such binary CS learner. The optimization problem is almost identical to that of (5), with the same number of constraints but including the cost c_i of misclassifying each example $\boldsymbol{x}_i$:

$$\begin{aligned} \min_{\boldsymbol{w}_{lm},\boldsymbol{\xi}^{lm}} \quad & \frac{1}{2}\langle \boldsymbol{w}_{lm}, \boldsymbol{w}_{lm}\rangle + C\sum_{y_i\in\{l,m\}} c_i\xi_i^{lm}, \\ \text{s.t.} \quad & \langle \boldsymbol{w}_{lm}, \boldsymbol{x}_i\rangle \geq +1-\xi_i^{lm}, \quad \text{if } \boldsymbol{x}_i\in\mathcal{S}_l, \\ & \langle \boldsymbol{w}_{lm}, \boldsymbol{x}_i\rangle \leq -1+\xi_i^{lm}, \quad \text{if } \boldsymbol{x}_i\in\mathcal{S}_m, \\ & \xi_i^{lm}\geq 0, \quad \forall \boldsymbol{x}_i\in\mathcal{S}_l\cup\mathcal{S}_m. \end{aligned} \tag{8}$$

Interestingly, fuzzy SVM [14] leads to this optimization problem too. The difference is that c_i stands for the fuzzy membership associated with $\boldsymbol{x}_i$.

The disadvantage of CS OVO is that the global model learned, formed by a set of binary classifiers, has not been induced optimizing any loss function. Next, our extension of the method by Crammer & Singer is presented, allowing for the optimization of CS loss functions, like (3) and (4). To the best of our knowledge, this method has never been presented before. The formulation is based on adding the cost c_i of each example $\boldsymbol{x}_i$ to the objective function,

$$\begin{aligned} \min_{\boldsymbol{W},\boldsymbol{\xi}} \quad & \frac{1}{2}\sum_{l=1}^{k}\langle \boldsymbol{w}_l, \boldsymbol{w}_l\rangle + C\sum_{i=1}^{n} c_i\xi_i, \\ \text{s.t.} \quad & (\langle \boldsymbol{w}_{y_i}, \boldsymbol{x}_i\rangle - \langle \boldsymbol{w}_r, \boldsymbol{x}_i\rangle) \geq e_{y_i}^r - \xi_i, \\ & \forall i=1,\ldots,n \quad \forall r\in\{1,\ldots,k\}. \end{aligned} \tag{9}$$

The most important consequence is that the cost produced by the misclassified examples can be controlled during the learning process. What is more, as

it shall be proved, the second term of the objective function constitutes an upper bound of Δ_{PMC} (4).

Theorem 1: At the solution $\boldsymbol{W}^*$, $\boldsymbol{\xi}^*$ of the optimization problem in (9) on the training dataset $\mathcal{S}$, the value of $\sum_{i=1}^{n} c_i \xi_i^*$ defines an upper bound of the total cost associated with misclassified examples.

Proof: In order to prove the theorem we must verify that, $\sum_{i=1}^{n} c_i \xi_i^* \geq \sum_{i=1}^{n} c_i [\![h(\boldsymbol{x}_i) \neq y_i]\!]$, in which $h(\boldsymbol{x}_i)$ is the prediction made for $\boldsymbol{x}_i$ by the set of models $\boldsymbol{W}^*$ when (7) is used as the decision rule. And this is trivial because the slack variables ξ_i in (9) are defined according to the hinge loss function. That is, $\xi_i^* \geq 1$ whenever the example $\boldsymbol{x}_i$ is misclassified, and $0 \leq \xi_i^* < 1$ if the true class of $\boldsymbol{x}_i$ is predicted. Thus, $\xi_i^* \geq [\![h(\boldsymbol{x}_i) \neq y_i]\!]$ is always true, and so is the expression above. ■

Therefore, if we define $C = C' / \sum_{i=1}^{n} c_i$ then the second term of (9) is an upper bound of our target loss function Δ_{PMC} (4). This allows our learner (through C') to trade-off between the complexity of the model and the misclassification cost. In order to obtain an upper bound for Δ_{AC}, just let $C = C'/n$.

These two CS methods were implemented[2] extending the code of [12]. Our proposal is a kind of Sequential Minimal Optimization algorithm [15], but instead of optimizing a pair of dual variables on each step, as happens in binary SVM, the set of dual variables of an example is optimized together.

IV. Plankton Biomass Estimation Dataset

The plankton samples from the Cantabrian Sea were processed using the FlowCam [16]. This is a device capable of analyzing and capturing an image of each organism (Figure 1) in a continuous flow. Then, our dataset of 5145 examples were classified by a taxonomist into 5 classes: Ciliata, Diatoms, Crustacea, Flagelata and a category named "Other", comprising rare taxa and unidentifiable objects.

Each example is described by 170 attributes, formed by different groups of characteristics. The performance of studied classifiers significantly degrade if we remove any of these groups. There are 26 morphological features calculated by the FlowCam, some of those are the particle perimeter, its area, the mean distance to perimeter from the centre, etc. The rest of the attributes were obtained applying several image analysis techniques to represent the texture and the shape of the individuals.

[2]Downloadable from http://www.aic.uniovi.es/˜juanjo/csbsvm.zip

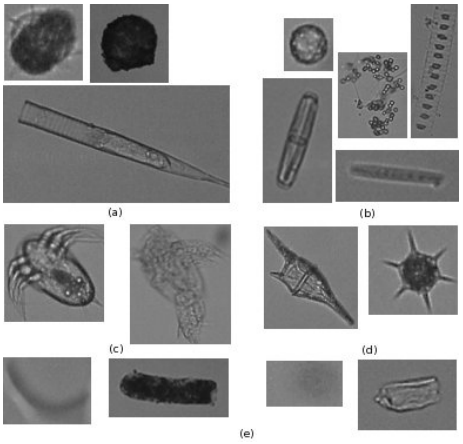

Fig. 1. Sample plankton images from five classes (a) Ciliata; (b) Diatoms; (c) Crustacea; (d) Flagelata; (e) Others

In order to describe the shape, firstly, we used Elliptic Fourier Descriptors (EFD) [17] to obtain a closed 2D contour. After some experiments, we chose 15 harmonics. Secondly, we added Hu moments [18] because they are translation, rotation and scaling invariant. This means that two organisms with the same shape but different size, and placed in different position or orientation, will have equivalent Hu moments. We also calculated 49 Zernike moments [19] using the centroid of the organism. They have interesting properties in terms of noise sensitivity, information redundancy and reconstruction capability. Finally, 8 granulometric features [20] were included. Previous works in plankton recognition [21] found that these features were crucial.

On the other hand, we employed Haralick features [22] and wavelets to represent the texture. Haralick attributes are metrics computed from gray-level co-occurrence matrices, in which element $[i, j]$ is the number of times pixels of values i and j are adjacent. Wavelets are a type of multi-resolution and multi-scale functions that allow hierarchical decomposition of a signal. Four-order Daubechies was chosen as the mother wavelet function and we analyzed 4 scales, with 3 detail sub-bands each. Energy firm, $E_n^m = \frac{1}{N \times N} \sum_{i,j=1}^{N} (s_n^m(i,j))^2$, was computed for each band, where s_n^m is the detail sub-band m, with scale n, and size $N \times N$.

Finally, we calculated the biomass (c_i) of each organism $\boldsymbol{x_i}$. In [23], the carbon biomass/volume relationship was studied and three ways of estimating the biomass were presented. They depend on the volume v_i (approximated from the particle diameter measured by the FlowCam) and the class of $\boldsymbol{x}_i$:

$$\log_{10} c_i = \begin{cases} -0.665 + 0.939 \log_{10} v_i \\ \quad \text{if } \boldsymbol{x}_i \notin \text{Diatoms } \& \ v_i > 3000um^3 \\ -0.933 + 0.881 \log_{10} v_i \\ \quad \text{if } \boldsymbol{x}_i \in \text{Diatoms } \& \ v_i > 3000um^3 \\ -0.583 + 0.86 \log_{10} v_i \\ \quad \text{if } v_i < 3000um^3. \end{cases}$$

V. Experimental Results

The goal of the experiments was to study the performance of OVO (5), C&S (6), cs-OVO (8) and cs-C&S (9) over the plankton biomass estimation dataset. The linear and the gaussian kernel, $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma \parallel \boldsymbol{x}_i - \boldsymbol{x}_j \parallel^2)$, were tested for each algorithm. To select the most appropriate values for parameters *C* and γ a search divided into two phases was made. The first one used *C* values in $[10^{-3}..10^{2}]$ and γ values in $[10^{-3}..10^{1}]$; in the second phase a finer search was carried out using ten values evenly distributed between those preceding and those following the best value obtained in the first phase. In this parameter searching process, non-CS algorithms selected the values that minimize zero-one error ($\Delta_{0/1}(h, \mathcal{S}') = \frac{1}{n} \sum_{\boldsymbol{x}_i \in \mathcal{S}'} [\![h(\boldsymbol{x}_i) \neq y_i]\!]$) , while CS methods optimize Δ_{PMC}. All the estimations for this parameter adjustment were made using a 3x2CV (2-fold cross validation repeated 3 times).

Table I shows the results obtained in a 2x5CV. Overall, decomposition algorithms achieve better results using the linear kernel, both in terms of $\Delta_{0/1}$ (OVO) and Δ_{PMC} (cs-OVO). It should also be noted that CS algorithms make a higher $\Delta_{0/1}$ error than their counterpart non-CS versions. Their differences are statistically significant in a Wilcoxon signed-ranks test with $p < 0.01$. However, CS versions present a lower Δ_{PMC} error, as it was expected. Noticeably, cs-C&S is significantly better than C&S ($p < 0.01$), but if we compare CS methods, cs-OVO is significantly better than cs-C&S ($p < 0.06$).

Nevertheless, previous results can be improved using the gaussian kernel. The best algorithm to optimize $\Delta_{0/1}$ is again OVO, but now C&S obtains almost the same score. Moreover, it seems that the differences between non-CS and CS are smaller, but still statistically significant: with $p < 0.01$ in the case of OVO vs cs-OVO and with $p < 0.04$ for C&S vs cs-C&S. Analyzing the scores for Δ_{PMC}, best results are those corresponding to cs-C&S. Interestingly, the difference between cs-C&S and OVO in Δ_{PMC} error is quite big; cs-C&S reduces the error of OVO in more than 37%. As it happened before, CS versions outperform their counterpart non-CS algorithms for Δ_{PMC}, in a higher degree in the case of cs-OVO, but the only statistically significant difference is between cs-C&S and C&S ($p < 0.10$). Notice that the differences among all C&S versions are now smaller, due to the fact that fairly low errors are always obtained. Finally, comparing CS algorithms, cs-C&S is significantly better than cs-OVO ($p < 0.06$). The main conclusion drawn from these results is that a CS algorithm using single optimization provides the best solution.

TABLE I
Error results ($\Delta_{0/1}$ and Δ_{PMC}) for the non-cost-sensitive (OVO and C&S) and cost-sensitive (cs-OVO and cs-C&S) algorithms

Kernel	Algorithm	$\Delta_{0/1}$	Δ_{PMC}
Linear	OVO	**0.1093**±0.0054	0.0922±0.0174
	cs-OVO	0.1778±0.0287	**0.0861**±0.0181
	C&S	0.1142±0.0057	0.1168±0.0398
	cs-C&S	0.1791±0.0154	0.1005±0.0381
Gauss.	OVO	**0.0640**±0.0062	0.0937±0.0438
	cs-OVO	0.1084±0.0165	0.0804±0.0409
	C&S	0.0653±0.0048	0.0646±0.0280
	cs-C&S	0.0696±0.0049	**0.0585**±0.0181

TABLE II
Biomass confusion matrix for cs-C&S using the gaussian kernel (all quantities are in thousands)

Class	Other	Cili.	Crust.	Flag.	Diat.	Prec.(%)
Other	13,949	136	178	105	67	96.63%
Ciliata	194	1,197	0	51	18	82.02%
Crustea	532	37	14,902	37	3	96.07%
Flagelata	88	61	0	2,646	30	93.65%
Diatoms	422	62	34	209	3,927	84.36%
Rec.(%)	91.9	80.2	98.6	86.8	97.1	

Table II shows the biomass confusion matrix using cs-C&S with the gaussian kernel. Each entry represents the amount of biomass of those examples of the class in the column predicted as the class in the row ($\sum_{\boldsymbol{x}_i \in S_{col}} c_i [\![h(\boldsymbol{x}_i) = row]\!]$). The last column and row, respectively, present the biomass percentage predicted by cs-C&S which truly belongs to that class (named as precision in information retrieval tasks), and the percentage of the real biomass of that class predicted by cs-C&S (recall), e.g. 98.6% of the total biomass corresponding to crustacea class has been correctly labelled, while 96.07% of the biomass that cs-C&S assigns to crustacea class,

actually belongs to this class. The greater difficulties lie in ciliata class in which both precision and recall are around 80%, and in the precision of diatoms.

VI. Conclusions

This study presents an interesting application that allows for the automatic biomass estimation of 5 plankton species. A new multiclass cost-sensitive method has been developed in order to improve such estimation. This algorithm is theoretically well-founded and is designed to optimize cost-sensitive loss functions. In practice, our method ameliorates the biomass prediction in comparison to the traditional multiclass classification approaches that optimize the accuracy. The proposed algorithm can be useful in other multiclass cost-sensitive applications.

References

[1] C. Elkan, "The foundations of cost-sensitive learning," in *Proceedings of IJCAI*. Morgan Kaufmann, 2001, pp. 973–978.

[2] P. D. Turney, "Types of cost in inductive concept learning," in *ICML-Workshop on Cost-Sensitive Learning*, 2000, pp. 15–21.

[3] T. Fawcett and F. Provost, "Adaptive fraud detection," *Data Mining and Knowledge Discovery*, vol. 1, pp. 291–316, 1997.

[4] A. Lenarcik and Z. Piasta, "Rough classifiers sensitive to costs varying from object to object," in *Int. Conf. on Rough Sets and Current Trends in Computing*. Springer, 1998, pp. 222–230.

[5] G. Almazan and C. Boyd, "Plankton production and tilapia yield in ponds," *Aquaculture*, vol. 15, no. 1, pp. 75–77, 1978.

[6] M. Benfield, P. Grosjean, P. Culverhouse, X. Irigoien, M. Sieracki, Á. López-Urrutia, H. Dam, Q. Hu, C. Davis, A. Hansen, C. Pilskaln, E. Riseman, H. Schultz, P. Utgoff, and G. Gorsky, "Rapid: Research on automated plankton identification," *Oceanography*, vol. 20, pp. 172–187, 2007.

[7] V. Vapnik, "An overview of statistical learning theory," *IEEE Trans. on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[8] ——, *Statistical Learning Theory*. NY: John Wiley, 1998.

[9] U. Kreßel, "Pairwise classification and support vector machines," in *Advances in Kernel Methods – Support Vector Learning*. MIT Press, 1999, pp. 255–268.

[10] E. Montañés, J. Barranquero, J. Díez, and J. J. del Coz, "Enhancing directed binary trees for multi-class classification." *Information Sciences*, vol. 223, pp. 42–55, 2013.

[11] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2001.

[12] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.

[13] U. Brefeld, P. Geibel, and F. Wysotzki, "Support vector machines with example dependent costs," in *ECML*, 2003.

[14] C.-F. Lin and S.-D. Wang, "Fuzzy support vector machines," *IEEE T. on Neural Networks*, vol. 13, no. 2, pp. 464–471, 2002.

[15] J. Lopez and J. Dorronsoro, "Simple proof of convergence of the smo algorithm for different svm variants," *IEEE T. on Neural Networks and Lear. Sys.*, vol. 23, pp. 1142 –1147, 2012.

[16] C. Sieracki, M. Sieracki, and C. Yentsch, "An imaging-in-flow system for automated analysis of marine microplankton," *Marine Ecology Progress Series*, vol. 168, pp. 285–296, 1998.

[17] F. Kuhl and C. Giardina, "Elliptic fourier features of a closed contour," *Computer Graphics and Image Processing*, vol. 18, pp. 236–258, 1982.

[18] M. Hu, "Visual pattern recognition by moment invariants," *IEEE T. on Information Theory*, vol. 8, pp. 179–187, 1962.

[19] M. R. Teague, "Image analysis via the general theory of moments," *Journal of the Optical Society of America (1917-1983)*, vol. 70, pp. 920–930, August 1980.

[20] G. Matheron, *Random sets and integral geometry*. Wiley, 1974.

[21] X. Tang, W. Stewart, H. Huang, S. Gallager, C. Davis, L. Vincent, and M. Marra, "Automatic plankton image recognition," *Artificial Intelligence Review*, vol. 12, pp. 177–199, 1998.

[22] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, no. 6, pp. 610–621, 1973.

[23] S. Menden-Deuer and E. Lessard, "Carbon to volume relationships for dinoflagellates, diatoms and other protist plankton," *Limnology and Oceanography*, vol. 45, pp. 569–579, 2000.

Pablo González received his Ms. degree in Computer Science from the University of Oviedo (Spain), in 2007. In 2009 he joined the Artificial Intelligence Center as a PhD candidate. His research interests include computer vision and pattern recognition.

Eva Álvarez is a PhD student at the Centro Oceanográfico de Gijón that belongs to the Instituto Español de Oceanografía (IEO). Her PhD thesis is focus on microplankton ecology. She has participated in 7 research cruises as the sampling responsible using the FlowCam.

Jose Barranquero is a predoctoral researcher at Artificial Intelligence Center. He has a M.Sc. in Soft Computing and Intelligent Data Analysis from the University of Oviedo. His research interests include quantification and opinion mining.

Jorge Díez received his Ph.D. degree in Computer Science from the University of Oviedo in 2003. In 2002 he joined the Computer Science Department of the University of Oviedo, where he is an Associate Professor. His research interests include preference learning, kernel methods, clustering and classification.

Rafael González-Quirós worked at different institutions, e.g. the Scripps Institution of Oceanography (USA) and joined IEO in 2008. He received his PhD at the University of Oviedo. He studies the effect of the pelagic ecosystem dynamics on the variability of marine fish populations and has published 11 peer reviewed publications.

Enrique Nogueira holds a permanent position at IEO. His research interests are plankton ecology and size-structure of plankton communities. He worked in several European Universities and has participated in 20 oceanographic cruises and 5 international projects. He has authored 21 research papers in peer reviewed journals.

Ángel López-Urrutia After working as a zooplankton ecologist at the Plymouth Marine Laboratory (UK), he joined IEO as a research associate. He received his PhD from the University of Oviedo. He has participated in 4 research cruises and authored over 15 papers in peer reviewed publications, including Nature, PNAS and Ecology.

Juan José del Coz received his Ph.D. in Computer Science from the University of Oviedo. In 1997, he joined the Computer Science Department, where he is currently an Associate Professor. He has authored over 30 papers in peer reviewed journals and conferences, including articles in NIPS, ICML, JMLR and Pattern Recognition.