Brief Papers

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 brief 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, example-dependent costs, kernel methods, plankton recognition, SVM.**

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 customer depends on the amount of money borrowed. Incorrectly diagnosing a healthy person as being sick is preferable to the opposite that may result in the loss of a life. Learning classifiers that consider the actual costs of their decisions should lead to improve the quality of these applications.

The techniques aimed to address these 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 [2] defines a taxonomy of the different types of costs that can be considered. This brief focuses on 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 brief, we present an example-dependent cost method to deal with the plankton biomass estimation problem.

The study of plankton is crucial because 1) plankton are the base of the food chain that sustains life in oceans [5] and 2) their ecosystem plays 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: 1) a plankton sampling device to automatically obtain high-resolution images from the plankton samples, 2) computer vision methods to process such images, and 3) a classification algorithm able to identify the species of each organism. The goal is to provide answers to questions such as: 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 maximize 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. Considering 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] were originally designed to solve binary classification tasks. Following such formulation, new methods have been proposed to build

Manuscript received September 26, 2012; accepted June 18, 2013. This work was supported in part by the Ministerio de Economía y Competitividad under Grant TIN2011-23558, and FICYT under Grant IB09-059-C2.

P. González, J. Barranquero, J. Díez, and J. J. del Coz are with the Artificial Intelligence Center, University of Oviedo, Gijón 33204, Spain (e-mail: pgonzalez@aic.uniovi.es; barranquero@aic.uniovi.es; jdiez@aic.uniovi.es; juanjo@aic.uniovi.es).

E. Álvarez, R. González-Quirós, E. Nogueira, and Á. López-Urrutia are with the Oceanographic Centre of Gijón, Gijón 33212, Spain (e-mail: eva.alvarez@gi.ieo.es; rgq@gi.ieo.es; enrique.nogueira@gi.ieo.es; alop@gi.ieo.es).

Digital Object Identifier 10.1109/TNNLS.2013.2271535

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-versus-all [8], one-versus-one (OVO) [9], or those using decision trees [10]. The second alternative considers all data in a single optimization problem [11]. This brief applies several SVMs to the plankton biomass estimation problem. In fact, the main contribution of this brief is the development of a new multiclass CS SVM. The proposed method is the extension of the 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 nonCS 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.

The paper is organized as follows. Section II introduces a formal setting for CS learning. The proposed method is discussed in Section III. Section IV describes the plankton data set used. Finally, some experimental results and the conclusion are presented in the last section.

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 according to the distribution $Pr(\mathcal{X}, \mathcal{Y}, \mathbb{R}^+)$ as follows:

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

in which $\delta_{\mathrm{CS}}(h(\boldsymbol{x}), y, c)$ 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

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

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 to measure the cost over a set of examples. The most common one is the loss function that returns the average cost per example,

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

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

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

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 brief 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 brief: OVO [9], because it obtains better performance [12] and the Crammer & Singer method [11], for being more efficient than others.

The OVO approach defines $k(k-1)/2$ binary problems, where the l-versus-m problem implies subsets $\mathcal{S}_l$ and $\mathcal{S}_m$ that 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 the control of the amount of regularization and ξ_i are the slack variables used to avoid overfitting and to cope with nonseparable 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 the Crammer & Singer method, a model $\boldsymbol{w}_l$ is induced for each class l following the one-versus-rest approach. The key difference is that all of them, $\boldsymbol{W} = \{\boldsymbol{w}_l : l \in \{1, \ldots, k\}\}$, are learned together as follows:

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

where $e_{y_i}^r$ 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, because the set of constraints of each example $\boldsymbol{x}_i$ shares one single slack variable ξ_i.

[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 class predicted by the algorithm will be determined following the winners-takes-all rule as follows:

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

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, because we can modify this method for optimizing a CS loss function, like Δ_{PMC} (4).

B. Cost-Sensitive Algorithms

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

The CS version of the OVO approach is based on the CS binary classifier presented in [13], in which the authors additionally provide 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 $\boldsymbol{x}_i$ as follows:

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

Note that the number of constraints is the same as in (5). 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 by 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. In addition, it shall be proven that 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 data set $\mathcal{S}$, the value of $\sum_{i=1}^{n} c_i\xi_i^*$ defines an upper bound of the total cost associated with misclassified examples.

Proof: 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. 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 tradeoff between the complexity of the model and the misclassification cost. To obtain an upper bound for Δ_{AC}, just let $C = C'/n$.

These two CS methods were implemented[2] by 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 dual variables of an example are optimized together.

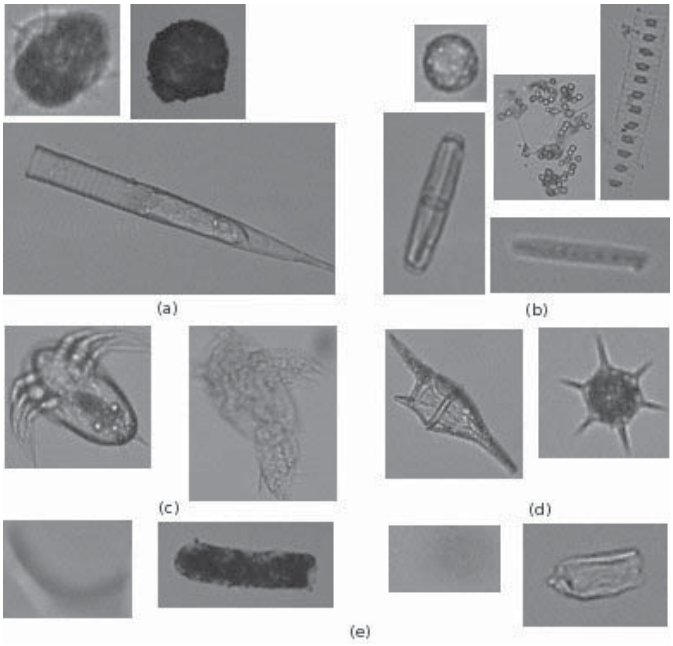

Fig. 1. Sample plankton images from five classes. (a) Ciliata. (b) Diatoms. (c) Crustacea. (d) Flagelata. (e) Others.

IV. Plankton Biomass Estimation Data Set

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 (Fig. 1) in a continuous flow. Then, our data set of 5145 examples were classified by a taxonomist into five classes: 1) Ciliata, 2) Diatoms, 3) Crustacea, 4) Flagelata, and 5) 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

[2]Download from http://www.aic.uniovi.es/~juanjo/csbsvm.zip

TABLE I
ERROR RESULTS ($\Delta_{0/1}$ AND Δ_{PMC}) FOR THE NONCS (OVO AND C&S) AND CS (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

classifiers significantly degrades 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 center, and so on. The rest of the attributes were obtained by applying several image analysis techniques to represent the texture and the shape of the individuals.

To describe the shape, firstly, we used elliptic Fourier descriptors [17] to obtain a closed 2-D 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 positions or orientations, 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, eight 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 multiresolution and multiscale function that allows hierarchical decomposition of a signal. Fourth-order Daubechies was chosen as the mother wavelet function and we analyzed four scales, with three detail sub-bands each. Energy firm, $E_n^m = 1/N \times N \sum_{i,j=1}^{N} (s_n^m(i, j))^2$, 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 $\boldsymbol{x}_i$ as follows:

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

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 data set. 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, nonCS 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 3 2CV (twofold cross validation repeated three times).

Table I shows the results obtained in a 2 5CV. 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 nonCS versions. Their differences are statistically significant in a Wilcoxon signed-ranks test with $p < 0.01$. CS versions, however, present a lower Δ_{PMC} error, as 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 nonCS and CS are smaller, but still statistically significant: with $p < 0.01$ in the case of OVO versus cs-OVO and with $p < 0.04$ for C&S versus cs-C&S. Analyzing the scores for Δ_{PMC}, the 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 before, CS versions outperformed their counterpart nonCS algorithms for Δ_{PMC}, in a higher degree in the case of cs-OVO, but the only statistically significant difference was between cs-C&S and C&S ($p < 0.10$). Notice that the differences among all C&S versions are now smaller, because of 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 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 that 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 the crustacea class has been correctly labeled, while 96.07% of the biomass

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	

that cs-C&S assigns to the crustacea class, actually belongs to this class. The greater difficulties lie in the ciliata class in which both precision and recall are around 80%, and in the precision of diatoms.

VI. CONCLUSION

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

REFERENCES

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

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

[3] T. Fawcett and F. Provost, "Adaptive fraud detection," *Data Mining Knowl. Discovery*, vol. 1, no. 3, pp. 291–316, 1997.

[4] A. Lenarcik and Z. Piasta, "Rough classifiers sensitive to costs varying from object to object," in *Proc. Int. Conf. Rough Sets Current Trends Comput.*, 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, no. 2, pp. 172–187, 2007.

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

[8] V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[9] U. Kreßel, "Pairwise classification and support vector machines," in *Advances in Kernel Methods*. Cambridge, MA, USA: 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," *Inf. Sci.*, vol. 223, pp. 42–55, Feb. 2013.

[11] K. Crammer and Y. Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *J. Mach. Learn. Res.*, vol. 2, pp. 265–292, Jan. 2001.

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

[13] U. Brefeld, P. Geibel, and F. Wysotzki, "Support vector machines with example dependent costs," in *Proc. ECML*, Sep. 2003, pp. 23–34.

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

[15] J. Lopez and J. Dorronsoro, "Simple proof of convergence of the SMO algorithm for different SVM variants," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1142–1147, Jul. 2012.

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

[17] F. Kuhl and C. Giardina, "Elliptic fourier features of a closed contour," *Comput. Graph. Image Process.*, vol. 18, no. 3, pp. 236–258, Mar. 1982.

[18] M. Hu, "Visual pattern recognition by moment invariants," *IEEE Trans. Inf. Theory*, vol. 8, no. 2, pp. 179–187, Feb. 1962.

[19] M. R. Teague, "Image analysis via the general theory of moments," *J. Opt. Soc. Amer.*, vol. 70, no. 8, pp. 920–930, Aug. 1980.

[20] G. Matheron, *Random Sets and Integral Geometry*. New York, NY, USA: Wiley, 1974.

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

[22] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. 3, no. 6, pp. 610–621, Nov. 1973.

[23] S. Menden-Deuer and E. Lessard, "Carbon to volume relationships for dinoflagellates, diatoms and other protist plankton," *Limnol. Oceanograph.*, vol. 45, no. 3, pp. 569–579, 2000.