Adjusting the Outputs of a Classifier to New a Priori Probabilities May Significantly Improve Classification Accuracy: Evidence from a Multi-class Problem in Remote Sensing

Patrice Latinne PLATINNE@ULB.AC.BE
IRIDIA Laboratory, Université Libre de Bruxelles, cp 194/06, 50 avenue Franklin Roosevelt, B-1050 Brussels, BELGIUM

Marco Saerens SAERENS@ULB.AC.BE
IRIDIA Laboratory, Université Libre de Bruxelles, cp 194/06, 50 avenue Franklin Roosevelt, B-1050 Brussels, BELGIUM
SmalS-MvM, Research Section, 102 rue du Prince Royal, B-1050 Brussels, BELGIUM

Christine Decaestecker CDECAES@ULB.AC.BE
Laboratory of Histopathology – Faculty of Medicine, Université Libre de Bruxelles, cp 620, 808 route de Lennik, B-1070 Brussels, BELGIUM

Abstract

In the present study, we introduce a simple iterative procedure that allows to correct the outputs of a classifier with respect to the new a priori probabilities of a new data set to be scored, even when these new a priori probabilities are unknown in advance. We also show that a significant increase in classification accuracy can be observed when using this procedure properly. More specifically, by applying the correcting procedure on the outputs of a simple logistic regression model, we observed an increase of 5.8% of classification rate on a difficult real-world multi-class problem – the automatic labeling of geographical maps based on remote sensing information. Moreover, the resulting classifier – the logistic regression model whose outputs have been adjusted according to our procedure – outperformed by more that 4% all of our previous models in terms of classification accuracy, including Bagfs (a multiple classifier system based on C4.5 decision trees), the best obtained model up-to-now.

1. Introduction

In many real-world classification problems, the training set is built by selecting a number of samples from each class, without respecting the true a priori probabilities of the classes, just because these true a priori probabilities are unknown.

As an example, let us consider the following scenario which concerns the problem investigated in the present paper: the automatic labeling of geographical maps based on remote sensing information. Each pixel of the map has to be labeled according to its nature (forest, agricultural zone, urban zone, etc). In this case, the a priori probabilities of the classes of the real-world data are unknown in advance and may vary considerably from one image to another, since they directly depend on the geographical area that has been observed (urban area, country area, etc).

The knowledge of the 'true' a priori probabilities in the real-world data is often highly desirable for the following important reasons:

- Optimal Bayesian decision making is based on the *a posteriori probabilities* of the classes given the observation (we have to select the class label that has the maximum estimated a posteriori probability). Now, following Bayes' rule, these a posteriori probabilities depend in a nonlinear way on the *a priori probabilities*. Therefore, a change of the a priori probabilities (as is the case for the real-world geographical data versus the training set) may have an important impact on the a posteriori probabilities of membership, which themselves affect the classification rate. In other words, even if we use an optimal Bayesian model, if the a priori

probabilities of the classes change, the model will not be optimal any more in these new conditions. However, knowing the new a priori probabilities of the classes would allow us to correct (by Bayes' rule) the outputs of the model in order to recover the optimal decision.

- Many classification methods, including neural network classifiers or logistic regression models, provide estimates of the a posteriori probabilities of the classes. From the preceding point, this means that applying such a classifier as-is on new data having different a priori probabilities from the training set can result in a loss of classification accuracy, in comparison with an equivalent classifier that relies on the 'true' a priori probabilities of the new data set.

- Estimation of class proportions in a real data set may also be an essential goal by itself, as for example in epidemiology where an important problem is the estimation of the disease prevalence in a population.

In this paper, we present a simple iterative procedure that estimates the new a priori probabilities of a new data set and adjusts the outputs of a classifier – supposed to approximate the a posteriori probabilities – accordingly, without having to refit the model, even when these new a priori probabilities are unknown in advance. This procedure is a simple instance of the EM algorithm (Dempster et al., 1977; McLachlan & Krishnan, 1997) that aims to maximize the likelihood of the new observed data (for more details, see Saerens et al., 2001).

Our main goal is therefore to show that this readjustment procedure can be very useful (i.e. it increases classification accuracy) when a classifier has been trained on a training set that does not reflect the true a priori probabilities of the classes in real-world conditions.

Notice, however, that this output readjustment procedure can only be applied if the classifier supplies estimates of the a posteriori probabilities. This is for instance the case if we use the least-squares error or the Kullback-Leibler divergence as a criterion for training, and if the minimum of the criterion is reached (see for instance Richard & Lippmann, 1991 or Saerens, 2000, for a recent discussion). In the case of approaches which generate categorical outputs, such as decision trees or rule-based systems, probability estimates can be generated with the Laplace estimate (Bradford et al., 1998).

The paper is organized as follows. The outputs adjustment procedure is described in section 2; the experiments are outlined in section 3; the conclusion is presented in section 4.

2. Correcting the outputs of the classifier with respect to new a priori probabilities

2.1 Training the classification model

Let us suppose a classification problem in n classes with the class labels taking their value in $\Omega = (\omega_1, \ldots, \omega_n)$. In order to train a classification model, we rely on a *training set*, i.e. a collection of observation vectors, $\mathbf{x}_k$, measured on individuals and allocated to one of the n classes $\in \Omega$.

For building this training set, we suppose that, for each class ω_i, observations on N_t^i individuals belonging to the class (with $\sum_{i=1}^{n} N_t^i = N_t$, the total number of training examples) have been independently recorded according to the within-class probability density $p(\mathbf{x}|\omega_i)$. The *a priori probability* of belonging to class ω_i in the *training set* will be denoted as $p_t(\omega_i)$ (in the sequel, subscript t will be used for estimates carried out on the basis of the *training set*) and is therefore estimated by the class frequency $\widehat{p}_t(\omega_i) = N_t^i/N_t$.

Let us now assume that the classification model has been trained, i.e. its parameters have been estimated on the basis of the training set (as indicated by subscript t). The classification model could be an artificial neural network, a logistic regression, or any other model that provides as outputs estimates of the a posteriori probabilities of the classes given the observations. We therefore assume that the model has n outputs, $g_i(\mathbf{x})$ $(i = 1, \ldots, n)$, providing estimated *a posteriori probabilities* of membership $\widehat{p}_t(\omega_i|\mathbf{x}) = g_i(\mathbf{x})$ (i.e., probability of belonging to class ω_i, given that observation vector $\mathbf{x}$ has been observed) in the conditions of the training set.

2.2 Adjusting the outputs to new a priori probabilities

2.2.1 An estimate of the new a priori probabilities is known

Let us now suppose that the trained classification model has to be applied to another data set (new cases, e.g. real-world data to be scored) for which the class frequencies, estimating the a priori probabilities $p(\omega_i)$ (no subscript t), are known. The case where the new a priori probabilities are unknown is covered in the next sub-section.

In the sequel, we will make the natural assumption that the generation of the observations within the classes, and thus each within-class density, does not change from the training set to the new data set ($\widehat{p}_t(\mathbf{x}|\omega_i) = \widehat{p}(\mathbf{x}|\omega_i)$): *only the number of measurements observed from each class has changed*. As stated in the Introduction (section 1), if used without modification, the classification model provides estimated a posteriori probabilities ($g_i(\mathbf{x}) = \widehat{p}_t(\omega_i|\mathbf{x})$) which, for these new cases, are biased by the prior probabilities of the training set ($\widehat{p}_t(\omega_i)$), and thus have to be corrected accordingly.

On the new data set to be scored, Bayes' theorem provides:

$$\widehat{p}_t(\mathbf{x}|\omega_i) = \frac{\widehat{p}_t(\omega_i|\mathbf{x})\widehat{p}_t(\mathbf{x})}{\widehat{p}_t(\omega_i)} \tag{1}$$

where the a posteriori probabilities $\widehat{p}_t(\omega_i|\mathbf{x})$ are obtained by applying the *trained model* as-is (subscript t) on some observation $\mathbf{x}$ of the new data set (i.e. by scoring the data).

The *corrected* a posteriori probabilities, $\widehat{p}(\omega_i|\mathbf{x})$ (relying on the a priori probabilities of the new data set), obey the same equation, but with $\widehat{p}(\omega_i)$ as the new a priori probabilities and $\widehat{p}(\mathbf{x})$ as the new probability density function (no subscript t):

$$\widehat{p}(\mathbf{x}|\omega_i) = \frac{\widehat{p}(\omega_i|\mathbf{x})\widehat{p}(\mathbf{x})}{\widehat{p}(\omega_i)} \tag{2}$$

Since the within-class densities $\widehat{p}(\mathbf{x}|\omega_i)$ do not change from training to real-world data ($\widehat{p}_t(\mathbf{x}|\omega_i) = \widehat{p}(\mathbf{x}|\omega_i)$), by equating equation (1) to (2) and defining $f(\mathbf{x}) = \widehat{p}_t(\mathbf{x})/\widehat{p}(\mathbf{x})$, we find

$$\widehat{p}(\omega_i|\mathbf{x}) = f(\mathbf{x})\frac{\widehat{p}(\omega_i)}{\widehat{p}_t(\omega_i)}\widehat{p}_t(\omega_i|\mathbf{x}) \tag{3}$$

And since $\sum_{i=1}^{n}\widehat{p}(\omega_i|\mathbf{x}) = 1$, we easily obtain $f(\mathbf{x}) = \left[\sum_{j=1}^{n}\frac{\widehat{p}(\omega_j)}{\widehat{p}_t(\omega_j)}\widehat{p}_t(\omega_j|\mathbf{x})\right]^{-1}$, and consequently

$$\boxed{\widehat{p}(\omega_i|\mathbf{x}) = \frac{\frac{\widehat{p}(\omega_i)}{\widehat{p}_t(\omega_i)}\widehat{p}_t(\omega_i|\mathbf{x})}{\sum_{j=1}^{n}\frac{\widehat{p}(\omega_j)}{\widehat{p}_t(\omega_j)}\widehat{p}_t(\omega_j|\mathbf{x})}} \tag{4}$$

This well-known formula can be used in order to compute the new a posteriori probabilities, $\widehat{p}(\omega_i|\mathbf{x})$, in terms of the outputs provided by the trained model, $g_i(\mathbf{x}) = \widehat{p}_t(\omega_i|\mathbf{x})$, and the new a priori probabilities $\widehat{p}(\omega_i)$.

We observe that the corrected a posteriori probabilities $\widehat{p}(\omega_i|\mathbf{x})$ are simply the outputs provided by the classifier, $g_i(\mathbf{x})$, weighted by the ratio of the new priors to the old priors, $\widehat{p}(\omega_i)/\widehat{p}_t(\omega_i)$. The denominator of (4) ensures that the corrected a posteriori probabilities sum to one.

2.2.2 No estimate of the new a priori probabilities is known

However, in many real-world situations, we ignore what the real-world a priori probabilities $p(\omega_i)$ are since we do not know the class labels for these new data. In this context, we now briefly present a new procedure for a priori and a posteriori probabilities adjustment, based on the EM algorithm (Dempster et al., 1977; McLachlan & Krishnan, 1997). This iterative algorithm increases the likelihood of the new data at each iteration until a local maximum is reached. We unfortunately do not have enough space here to develop the complete proof, which can be found in a technical report (Saerens et al., 2001).

As before, let us suppose that we record a set of N new independent realizations of variable $\mathbf{x}$, $\mathbf{X}_1^N = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, sampled from $p(\mathbf{x})$, in a *new data set* to be scored by the model. The likelihood of these new observations is defined as:

$$\begin{aligned} L(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) &= \prod_{k=1}^{N} p(\mathbf{x}_k) \\ &= \prod_{k=1}^{N}\left[\sum_{i=1}^{n} p(\mathbf{x}_k|\omega_i)p(\omega_i)\right] \end{aligned} \tag{5}$$

where the within-class densities, i.e. the probabilities of observing $\mathbf{x}_k$ given the class ω_i, are fixed ($p(\mathbf{x}_k|\omega_i) = p_t(\mathbf{x}_k|\omega_i)$) since we assume that only the a priori probabilities (the proportions of observations of each class) change from the training set to the new data set. We have to determine the estimates, $\widehat{p}(\omega_i)$, that maximize the likelihood (5) with respect to $p(\omega_i)$. While a closed-form solution to this problem cannot be found, we can nevertheless obtain an iterative procedure for estimating new $p(\omega_i)$ by applying the EM algorithm.

As before, let us define $g_i(\mathbf{x}_k)$ as the model's output value corresponding to class ω_i for observation $\mathbf{x}_k$ from the new data set to be scored. The model's outputs provide an approximation of the a posteriori probabilities of the classes given the observation in the conditions of the training set (subscript t), while the a priori

probabilities are estimated by the class frequencies:

$$\widehat{p}_t(\omega_i|\mathbf{x}_k) = g_i(\mathbf{x}_k) \quad (6)$$

$$\widehat{p}_t(\omega_i) = \frac{N_t^i}{N_t} \quad (7)$$

Let us define as $\widehat{p}^{(s)}(\omega_i)$ and $\widehat{p}^{(s)}(\omega_i|\mathbf{x}_k)$ the estimates of the new a priori and a posteriori probabilities at step s of the iterative procedure. If the $\widehat{p}^{(s)}(\omega_i)$ probabilities are initialized by the frequencies of the classes in the training set (equation 7), the EM algorithm provides the following iterative steps (see Saerens et al., 2001), for each new observation $\mathbf{x}_k$ and each class ω_i:

$$\boxed{\begin{array}{c} \widehat{p}^{(0)}(\omega_i) = \widehat{p}_t(\omega_i) \\ \\ \widehat{p}^{(s)}(\omega_i|\mathbf{x}_k) = \dfrac{\dfrac{\widehat{p}^{(s)}(\omega_i)}{\widehat{p}_t(\omega_i)}\widehat{p}_t(\omega_i|\mathbf{x}_k)}{\displaystyle\sum_{j=1}^{n}\dfrac{\widehat{p}^{(s)}(\omega_j)}{\widehat{p}_t(\omega_j)}\widehat{p}_t(\omega_j|\mathbf{x}_k)} \\ \\ \widehat{p}^{(s+1)}(\omega_i) = \dfrac{1}{N}\displaystyle\sum_{k=1}^{N}\widehat{p}^{(s)}(\omega_i|\mathbf{x}_k) \end{array}} \quad (8)$$

where $\widehat{p}_t(\omega_i|\mathbf{x}_k)$ and $\widehat{p}_t(\omega_i)$ are given by (6) and (7). At each iteration step s, both the a posteriori ($\widehat{p}^{(s)}(\omega_i|\mathbf{x}_k)$) and the a priori probabilities ($\widehat{p}^{(s)}(\omega_i)$) are re-estimated sequentially for each observation $\mathbf{x}_k$ and each class ω_i. The iterative procedure proceeds until the convergence of the estimated probabilities, $\widehat{p}^{(s)}(\omega_i)$. The reader will notice the similarity between equations (4) and (8).

Notice that, although we did not encounter this problem in our simulations, we must keep in mind that, potentially, local maxima problems may occur with the EM algorithm.

In Saerens et al., 2001, we also showed that a likelihood ratio test (see for instance Papoulis, 1991) can be used in order to decide if the a priori probabilities have significantly changed from the training set to the new data set. The readjustment procedure should be applied only when we find a significant change of a priori probabilities.

Of course, in order to obtain good a priori estimates, it is necessary that

1. The a posteriori probabilities provided by the model (the readjustment procedure can only be applied if the classifier provides as output an estimate of the a posteriori probabilities) are reasonably well-approximated, which means that it provides predicted probabilities of belonging to the classes that are sufficiently close to the observed probabilities.
2. The new data set to be scored is large enough in order to be able to estimate accurately the new a priori class probabilities.
3. The training set selection (the sampling) has been performed on the basis of the discrete dependent variable (the classes), and not of the observed input variable $\mathbf{x}$ (the explanatory variable), so that the within-class probability densities, $p(\mathbf{x}|\omega_i)$, do not change. In other words, only the a priori probabilities change from the training set to the real-world data sets.

About these conditions, simulation results (Saerens et al., 2001) showed that the EM procedure improves classification rate, even if the classifier's output provides imperfect a posteriori estimates. Additionally, the quality of the estimates does not appear to depend on the size of the new data set (point 2). About condition 3, if sampling also occurs on the basis of $\mathbf{x}$, the usual sample survey solution to this problem is to use weighted maximum likelihood estimators with weights inversely proportional to the selection probabilities, which are supposed to be known.

We now present experimental results supporting the use of our output readjustment procedure.

3. A Remote Sensing Application

3.1 Material

In this section, we present a real-world application that illustrates the practical usefulness of the iterative adjustment of the classifier's outputs. We tackled a difficult and important problem in the remote sensing field. The goal is to interpret automatically the land cover of a whole remote sensing image. The image is constituted of (1201 x 1201) pixels extracted from the LANDSAT Thematic Mapper (7 bands) data.

Figure 1 shows the experts' visual interpretation of the image which is used as reference data in order to test the classification accuracy (the reference image). Each pixel of the image (1,442,401 pixels in total) is labeled with a legend of 11 classes (listed in Table 1).

We computed 50 features from textural statistical filters applied to circular regions surrounding each pixel of the image, in order to associate the inner spectral information of each pixel with the information given by homogeneous patterns and spatial arrangements that the pixel's intensity or color alone do not sufficiently describe (as detailed in Debeir et al., 2001). Each pixel

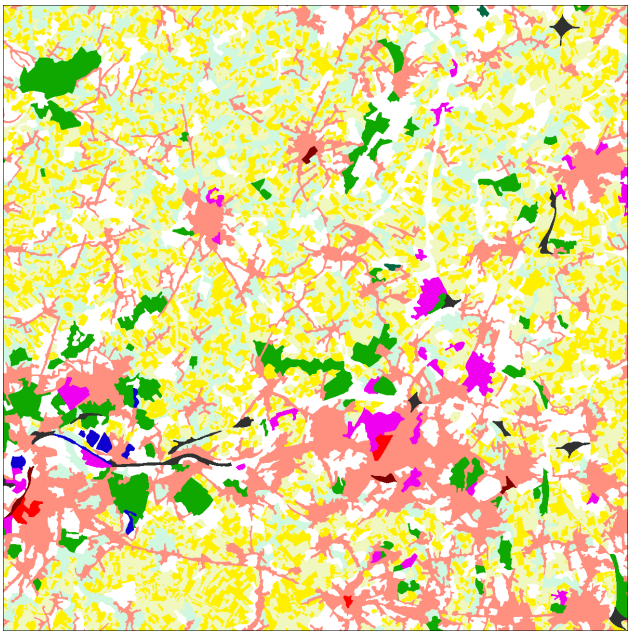

Figure 1. Visual interpretation (reference image).

Table 1. Priors estimation on the test set (in %).

Class label	**True priors**	**Log-EM**
Arable land : cultivated soil	29.7	28.9
Discontinuous urban fabric	23.1	21.1
Arable land without vegetation	22.0	21.0
Pastures	15.0	16.0
Broad-leaved forest	6.6	7.1
Industrial, commercial units	2.0	2.9
Road networks and assoc. land	0.7	1.4
Water bodies	0.3	0.4
Continuous urban fabric	0.2	0.4
Rail networks and assoc. land	0.2	0.4
Coniferous forest	0.1	0.1

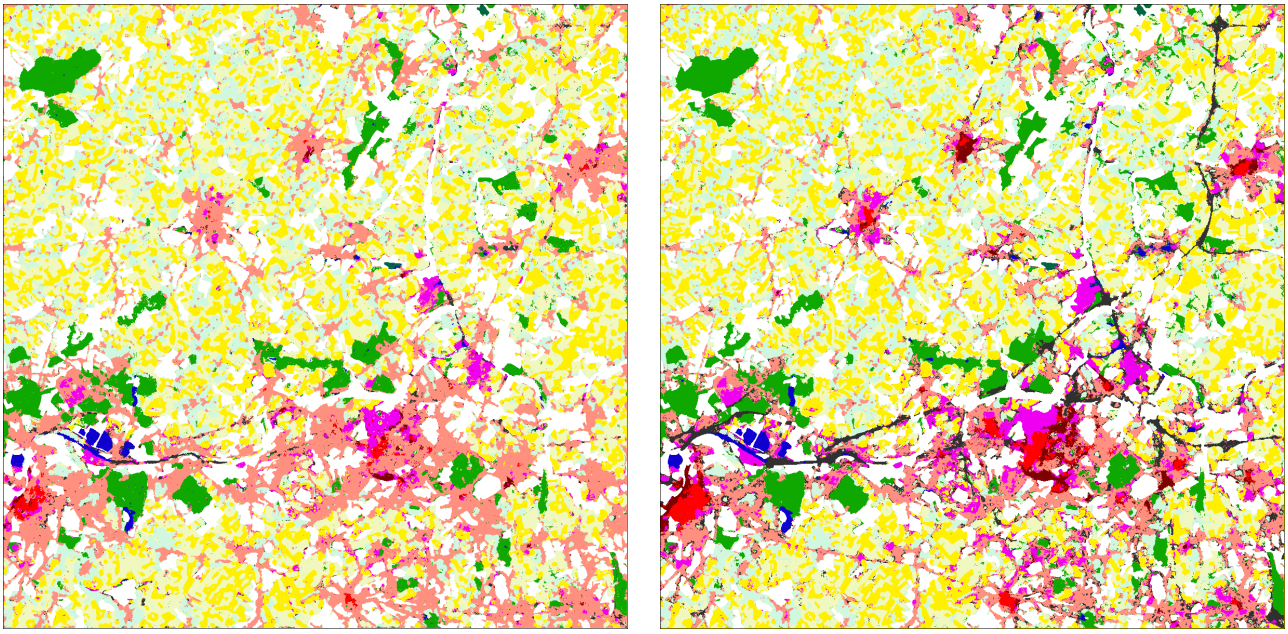

Figure 2. Remote sensing application: map resulting from classification with Log-EM (left) and Bagfs (right).

of the image is thus associated with an observation vector of 50 features ($\mathbf{x}_k$).

3.2 Experimental design

The main goal is, of course, to classify all the pixels of the image in one of the 11 class labels. Once this is done, the image is ready for post-processing (post-filtering, etc).

The learning set was built by randomly selecting 200 different pixels in each class from the whole image (i.e., a total of 2200 examples; the remaining 1,440,201 pixels were left for the test set). This random selection simulated the real expert's labeling task. Indeed, in a real-world situation, an expert is asked to select by visual inspection a given number of pixels of each class (200 pixels, in our case), and to provide a label for each of them. Once this is done, the complete labeling

Table 2. Per class and global error rates (in %) obtained on the test set (the reference image).

Class label (class frequency)	**Log regr.**	**With true priors**	**Log-EM**	**Bagfs**
Arable land : cultivated soil (29.7%)	33.1	26.6	26.8	35.7
Discontinuous urban fabric (23.1%)	49.2	25.3	27.5	48.2
Arable land : without vegetation (22.0%)	22.5	22.2	22.5	17.3
Pastures (15.0%)	33.9	37.2	35.9	31.8
Broad-leaved forest (6.6%)	16.7	19.9	19.5	17.1
Industrial, commercial units (2.0%)	42.1	54.3	50.4	33.2
Road networks and associated land (0.7%)	32.2	85.0	75.9	17.6
Water bodies (0.3%)	8.2	10.8	10.7	3.4
Continuous urban fabric (0.2%)	19.1	41.8	37.7	2.1
Rail networks and associated land (0.2%)	25.2	57.4	49.3	5.1
Coniferous forest (0.1%)	2.3	7.6	7.1	0.5
Global error rate	**33.6**	**27.5**	**27.8**	**32.3**

of the map (the remaining 1,440,201 pixels) is left to an automatic system (the classifier) trained on the expert's labeled pixels. In the present case, however, we have a reference map that has been completely labeled in order to assess the classifier's performances. Notice in Table 1 ('True priors') that the distribution of the classes is not well-balanced in the image, in contrast with the learning set.

The EM readjustment procedure is perfectly suited in this case where true a priori probabilities are unknown and, furthermore, may vary considerably from one image to another.

In order to evaluate the actual impact of the proposed algorithm, we compared the class prediction on the test set (the reference image of Figure 1, from which the pixels of the training set have been excluded) obtained by means of (1) a standard logistic regression, 'Log regr.', without outputs readjustment, (2) a standard logistic regression with outputs readjusted by means of the EM algorithm, 'Log-EM' (i.e. by applying the iterative procedure (equations (8)), to the model outputs), and (3) a standard logistic regression with outputs readjusted (with equation (4)) by using the true priors of the test set (which are unknown in a real-world situation), labeled 'With true priors'. This latter result can be consider as an optimal reference in the present experimental context.

We also compared these results with those obtained by a multiple classifier system, 'Bagfs', based on the association of Breiman's bagging (Breiman, 1996) and Ho's random subspaces (Ho, 1998) applied to Quinlan's (1993) C4.5 decision tree Release 8 (see more details in Latinne et al., 2000). This latter method was found to achieve the best global and per class accuracies for this problem among other standard classification systems (see Debeir et al., 2001), up to now. Notice that the outputs of the Bagfs model cannot be adjusted by the EM-based procedure because, in its current version, it does not provide estimates of the a posteriori probabilities.

Before evaluating the class prediction on the whole image for each of the four algorithms described above (i.e. 'Log regr.', 'With true priors', 'Log-EM' and 'Bagfs'), a preliminary, less computer-intensive, experiment was carried out in order to measure the variance in the classification error rate, computed on smaller test sets (the whole test set includes 1,440,201 observations). From our complete learning set of 2200 cases (200 cases per class), we selected 10 different training sets made up of 90% of cases within each class (as training sets resulting from a conventional stratified 10-fold cross-validation). For each training set, we selected an independent test set of 669 cases in which the class distribution observed in the reference image (see Table 1) was respected as far as possible.

3.3 Experimental Results

The preliminary experiment aiming to estimate variance produced the following results. In terms of mean error rate ± standard deviation, we obtained: 31.4% ± 1.2% for Log regr., 25.2% ± 1.3% for Log-EM and 33.0% ± 1.7% for Bagfs. These first results show the improvement in error rate due to EM-based output adjustment applied to the standard logistic regression (about 6% for a standard deviation below 1.5%). Furthermore, Log-EM also outperformed Bagfs.

These results were confirmed when the classifiers were trained on the whole training set (2200 pixels) and applied to the whole test set (1,440,201 pixels), as detailed in Tables 1 and 2.

Table 1 shows the different classes with their true a priori probabilities in the reference image (from which the pixels of the training set have been excluded, 'True Priors'), in comparison with the prior estimated by the EM algorithm, after adaptation of the outputs of the logistic regression ('Log-EM'). These results indicate that Log-EM was able to provide good estimates even in the case of a strongly unbalanced class distribution (which is generally the case in remote sensing, see Table 1). Notice that the EM procedure converged after 5 passes through the reference image (i.e. after 5 iterations of equations (8)).

Table 2 shows the classification results on the whole image. The error rates are reported for the 4 different classifiers described above. The error rates are indicated for each individual class and, globally, for the whole image : the 'global error rate' (i.e. the sum of the per class error rates weighted by the a priori probabilities).

By examining Table 2, it is worth noticing that the Log-EM model had a global error rate very similar to the one obtained by the logistic regression adjusted with the true priors (which are unknown in a real-world situation), i.e. the optimal correction for the logistic regression model in the present context. Log-EM also outperformed the other models globally in terms of classification accuracy by more than 4%. The analysis of the Log-EM and Bagfs pixel classifications shows that Log-EM correctly classified 144,960 pixels which were erroneously classified by Bagfs as compared to the 92,955 pixels correctly classified by Bagfs and erroneously classified by Log-EM. The statistical significance of the Log-EM performances was confirmed by means of the McNemar nonparametric change test (Siegel & Castellan, 1988; Salzberg, 1997). This test exhibited a significance level of $p < 10^{-6}$ when comparing the results of Log-EM with respect to both Log regr. and Bagfs.

This is a remarkable result since the problem at hand is a difficult one, and all the previously tested classification models were optimized during several man-months, cumulating in a sophisticated model that performed best (in terms of both the classification accuracy assessment and the visual interpretation) on these remote sensing applications (Bagfs; see Latinne et al., 2000). The fact that a simple model, such as a logistic regression after automatic readjustment with respect to the new priors, performed better than these more sophisticated models, was considered as quite surprising.

We also observe that, because of output adjustment, Log-EM performed better than Log regr. on the classes with a high a priori probability (e.g. 'Arable land: cultivated soil' and 'discontinuous urban fabric') and poorer on the other classes. In order words, Log-EM shifts its decision surfaces in order to perform better on the classes with a high a priori probability, so that its global error rate was finally lower than the Log regr. one.

In contrast, the comparison of Bagfs and Log regr. (i.e. the two classifiers not submitted to output adjustment) shows that Bagfs performed especially well for the classes with a low a priori probability (smaller than 1%), while performing as well as Log regr. on the other classes, resulting in a global error rate slightly lower for Bagfs. Finally, Log-EM (by boosting its performance on the classes strongly represented) outperformed Bagfs, even if Bagfs performed better on the majority of the classes (9 of 11).

Finally, the maps illustrated in Figure 2 show that Log-EM produced a map that is much closer to the visual interpretation (reference map of Figure 1) than Bagfs, as confirmed by the experts themselves.

4. Conclusion

We presented a simple procedure allowing to adjust the outputs of a classifier to new a priori class probabilities. This procedure is a simple instance of the EM algorithm and was applied here to the outputs of a logistic regression. The practical usefulness of this readjustment procedure was illustrated in the context of an important and difficult remote sensing application for which the true a priori probabilities of belonging to a class are unknown and strongly unbalanced.

This experiment largely confirmed the preliminary results obtained for binary classification on artificial data and data sets from the UCI repository, (Saerens et al., 2001). It showed that the EM-based readjustment procedure, applied here to a multi-class problem, was able to provide good estimates of the true a priori probabilities, and to improve classification accuracy. In the context of our remote sensing application, the EM procedure provided the results which the best agree with the experts' reference image interpretation.

In summary, here are the *main conclusions* of our experimental evaluation: (1) The EM readjustment procedure was able to provide reasonably good estimates of the new a priori probabilities; (2) The logistic regression model with adjusted outputs performed significantly better than all the previous classification models in terms of classification accuracy; (3) The classification performances after adjustment by EM were relatively close to the results obtained by using the true

priors (which are unknown in a real-world situation).

Notice that Provost et al. (2001) recently proposed an alternative method for building a classifier when the a priori probabilities of the classes are unknown, without requiring posteriori probability estimates. However, the problem of determining the a priori probabilities at testing time was not treated. The present method proposes a direct solution to this problem.

Future work will aim to combine the strength of Log-EM on the biggest classes and Bagfs' strength on the smallest in order to obtain better automatically classified images. We are also working on biomedical applications of the EM-based readjustment procedure in the field of disease prevalence estimation.

Acknowledgments

Part of this work was supported by the project RBC-BR 216/4041 from the "Région de Bruxelles-Capitale", and funding from the SmalS-MvM. Patrice Latinne is supported by a grant under an ARC (Action de Recherche Concertée) programme of the Communauté Française de Belgique. Christine Decaestecker is a Research Associate with the 'F.N.R.S.' (Belgian National Scientific Research Fund).

The authors would like to thank Professor Eleonore Wolff and Isabelle Van den Steen (remote-sensing department of the I.G.E.A.T. – ULB, Brussels) for the supply of the remote sensing data and their overall visual interpretation assessment.

References

Bradford, J., Kunz, C., Kohavi, R., Brunk, C., & Brodley, C. (1998). Pruning decision trees with misclassification costs. *Proceedings of the Ninth European Conference on Machine Learning* (pp. 131–136).

Breiman, L. (1996). bagging predictors. *Machine Learning*, *24*, 123–140.

Debeir, O., Van den Steen, I., Latinne, P., Van Ham, P., & Wolff, E. (2001). *Textural and contextual land-cover classification using single and multiple classifier systems* (Technical Report TR/IRIDIA/2001-03). Université Libre de Bruxelles, IRIDIA, cp 194/06.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society*, *39*, 1–38.

Ho, T. (1998). the random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*, 832–844.

Latinne, P., Debeir, O., & Decaestecker, C. (2000). Different ways of weakening decision trees and their impact on classification accuracy. *Proceedings of the First International Workshop of Multiple Classifier System* (pp. 200–210). Cagliari, Italy: Springer (Lecture Notes in Computer Sciences; Vol. 1857).

McLachlan, G., & Krishnan, T. (1997). *The em algorithm and extensions*. Wiley.

Papoulis, A. (1991). *Probability, random variables, and stochastic processes*. McGraw-Hill. 3rd edition.

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, *42*, 203–231.

Quinlan, J. (1993). *C4.5 : Programs for machine learning*. San Mateo, California: Morgan Kaufmann Publishers.

Richard, M., & Lippmann, R. (1991). Neural network classifiers estimate bayesian a posteriori probabilities. *Neural Computation*, *2*, 461–483.

Saerens, M. (2000). Building cost functions minimizing to some summary statistics. *IEEE Transactions on neural networks*, *11*, 1263–1271.

Saerens, M., Latinne, P., & Decaestecker, C. (2001). *Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure* (Technical Report TR/IRIDIA/2001-2). Université Libre de Bruxelles, IRIDIA - cp 194/06. Manuscript available from ftp://iridia.ulb.ac.be at /pub/latinne/MARCH/apriori.pdf.

Salzberg, S. (1997). On comparing classifiers : Pitfalls to avoid and a recommended approach. *Data Mining and knowledge discovery*, *1*, 317–327.

Siegel, S., & Castellan, N. (1988). *Nonparametric statistics for the behavioral sciences*. McGraw-Hill. 2nd edition.