In: Perspectives on Pattern Recognition
Editor: Monica D. Fournier

ISBN: 978-1-61209-118-1

© 2012 Nova Science Publishers, Inc.

Chapter 9

Class Distribution Estimation in Imprecise Domains based on Supervised Learning

***Víctor González-Castro*[*]*, Rocío Alaiz-Rodríguez*[†] *and Enrique Alegre*[‡]
Dpto. de Ingeniería Eléctrica y de Sistemas y Automática,
Universidad de León, Campus de Vegazana, 24071 León, Spain

Abstract

Quantification - or proportion estimation - plays an important role in many practical classification problems. On the one hand, a machine that automatically classifies an element into a group of predefined classes will make suboptimal decisions if the class distribution in the test (real) domain differs from the one assumed in learning. Estimating the new class distribution is necessary in order to adapt the classifier to the new operational conditions. On the other hand, there are some real domains where the quantification task itself is the main goal. Some fields, such as quality control, direct marketing, tendency study or some textual recognition tasks, require methods that can reliably estimate the proportion of elements within each category without any concerns about how each element has been classified individually. We describe several quantification techniques that rely on supervised learning and provide these estimations based on: (a) the classifier confusion matrix, (b) the posterior probability estimations, and (c) distributional divergence measures. We illustrate these techniques, as well as their robustness against the base classifier performance in a practical seminal quality control setting where the ultimate goal is to quantify the proportion of sperm cells with damaged/intact acrosome.

1. Introduction

Many works in the machine learning domain are focused on extracting the best possible features from a set of objects and optimizing a classifier. Once the classifier is designed, it is applied as-is to a data set in order to predict the class each individual belongs to. This

[*] E-mail address: victor.gonzalez@unileon.es

[†] E-mail address: rocio.alaiz@unileon.es

[‡] E-mail address: enrique.alegre@unileon.es

Complimentary Contributor Copy

process has been widely studied and is called *classification*. Image [19, 4] or speech and audio processing [22] are just some applications that involve classification tasks.

In many supervised learning studies the fact that training and test data follow the same, although unknown, distribution is taken for granted [10]. In particular, prior class probabilities estimated from the training data set are considered to truly reflect the target class distribution. However, time or space class stationarity cannot be assumed in many practical fields. Indeed, if the class distribution of the present sample differs from the one of the training set, the classifier will be suboptimal. For example, if a word sense disambiguation system is trained using instances of words from a certain domain (e.g., Sports news), but it is then used with instances from a different domain (e.g., Political news), where the sense priors are different, the accuracy is affected [6]. It is important to highlight that there will be a fracture in classifier performance if the test class distribution differs from the one assumed in learning. Whenever there is such a change, estimating this new class proportion will be important to adapt the classifier to the new context. This is a common problem that has brought forth some high attention lately [18, 20, 23, 24].

In order to prevent a drop in classifier performance, several approaches have been proposed. A minimax approach [1] tackles this uncertainty problem by optimizing the classifier for a class prior probability that minimizes the maximum possible classifier error that may result from a shift in the class prior probabilities. Others, however, attempt to adapt the classifier to new operating conditions. Some authors rely on an eventual perfect knowledge of the new conditions by the end user [9], but when this is not possible, Saerens et al. [20] propose a re-estimation process of the new conditions as long as the classifier provides estimates of the posterior probabilities of class membership and an unlabeled data set is available. Based on the new estimated conditions (priors), the classifier is adapted in order to minimize the error rate or risk.

There are also other problems where the goal is not to classify each item in a set, but to estimate the proportion of elements of each class, which is known as *quantification* [11]. Although it has been less explored in the literature, quantification has been applied to some real domains, such as quality control [2, 21], news categorization [12, 13], analysis of technical-support call logs [14] or word text disambiguation [5, 6].

Reliably estimating the class proportion of positive/negative samples with no concerns about the individual classification could be addressed in a naïve way by counting the instances classified as positives and negatives by the classifier. This has been referred as *Classify & Count* (CC) by Forman [11, 12] in the context of news categorization. Except for (nearly) perfect classifiers, this approach is not adequate. In his studies, methods based on the classification confusion matrix (*Adjusted Count* (AC) and *Median Sweep* (MS)) are found to outperform the naïve CC method.

On the other hand, Saerens et al. [20] found that in order to improve the classifier accuracy by readjusting the classifier outputs for the new priors, methods based on the classifier posterior probability estimates outperform those that rely on the confusion matrix. In this chapter, we explore this method based on posterior probabilities (PP), but in this case in order to directly estimate the new class distribution.

When there is a shift in the *a priori* probabilities of the classes between the training and the test (real) set, the data distributions, as well as the class posterior probability distribution also change. Measuring the divergence between this test distribution and different generated

Complimentary Contributor Copy

calibration labeled sets would allow to find the class distribution for which this divergence is minimum [15]. The Hellinger Distance [8] is used as distributional divergence metric in this method.

To summarize, in this chapter we consider quantification methods based on:

- The classifier confusion matrix (AC, MS)
- The posterior probability estimations (PP)
- Distributional divergence measures (HD)

The remainder of the chapter is organized as follows: The theoretical approach and algorithms of the quantification methods are exposed in Section 2.. In section 3. an illustration of the quantification methods using real data set of a semen quality control application is given and, finally, the conclusions are pointed out in Section 4..

2. Class Distribution Estimation Methods

Consider a classification problem with a training data set $S_t = \{(x^k, d^k), k = 1, \dots, K\}$ where x^k is the feature vector of the $k-th$ element of the set, d^k is its class label, which takes its value in $\Omega = \{d_1, d_2, \dots d_M\}$.

The *a priori* probability of belonging to class d_i in S_t is denoted by $p_t(d = i) = p_t(d_i)$ [1]. Consider that all the elements $x^k \in S_t$ have been independently recorded according to the class probability density function $p(x|d_i)$.

Let us also consider a classifier trained using S_t that makes decisions in two steps: it first computes a soft output $\widehat{y}_i^k$ and then, based on it, makes a hard decision $\widehat{d}^k \in \Omega$. It is well known that if the classifier is trained minimizing an appropriate cost function [4], the soft outputs $\widehat{y}_i^k$ will provide an estimation of the *a posteriori* probability of the observation x_k belonging to class d_i provided by the classifier $(\widehat{p}_t(d_i|x))$.

Let us now suppose that the trained classification model is applied on another data set with unknown *a priori* probabilities $p(d_i)$, which are probably different (due to many factors, as we have previously pointed out) from the ones of the training set. The naïve approach to estimate the actual class distribution is usually based on counting the labels assigned by the classifier (CC method). The estimations made with this method will not be reliable: (i) the classifier performance will drop if there is a difference between $p(d_i)$ and $p_t(d_i)$, and (ii) there is no guarantee that the errors made will compensate between them. In the following subsections some quantification techniques to estimate the true *a priori* probability will be described. These techniques are based on the classifier confusion matrix, on the posterior probability estimations provided by the classifier and on the divergence between distributions measured by means of the Hellinger Distance.

2.1. Estimation Based on the Confusion Matrices

The performance of a classifier can be summarized by its confusion matrix, which is an observation of the number of elements classified as belonging to the class i while they actually belong to the class j:

[1] The subscript t will be used for estimates on the basis of the training set in the remaining of the chapter.

Complimentary Contributor Copy

Table 1. Confusion Matrix for a binary classification problem

		Predicted	
		$\widehat{d_1}$	$\widehat{d_0}$
True class	d_1	TP	FP
	d_0	FP	TN

The observed count of positives P' from the classifier will include both true positives and false positives. Similarly, the number of real positive examples is $P = TP + FN$ while the number of actual negatives is $N = FP + TN$. This is used to compute the following rates:

- True Positive rate: $tpr = \widehat{p}(\widehat{d_1}|d_1) = TP/P$
- False Positive rate: $fpr = \widehat{p}(\widehat{d_1}|d_0) = FP/N$
- False Negative rate: $fnr = \widehat{p}(\widehat{d_0}|d_1) = FN/P$
- True Negative rate: $tnr = \widehat{p}(\widehat{d_0}|d_0) = TN/N$

Now, let us say that the probability that a classifier gives a positive prediction (in a binary case) is:

$$\begin{aligned}
\widehat{p}(\widehat{d_1}) &= \widehat{p}(\widehat{d_1}|d_1)\cdot\widehat{p}(d_1) + \widehat{p}(\widehat{d_1}|d_0)\cdot\widehat{p}(d_0) = \\
&= \widehat{p}(\widehat{d_1}|d_1)\cdot\widehat{p}(d_1) + \widehat{p}(\widehat{d_1}|d_0)\cdot(1-\widehat{p}(d_1)) = \\
&= tpr\cdot\widehat{p}(d_1) + fpr\cdot(1-\widehat{p}(d_1))
\end{aligned}$$

Solving the equation we get:

$$\begin{aligned}
\widehat{p}(\widehat{d_1}) &= tpr\cdot\widehat{p}(d_1) + fpr - fpr\cdot\widehat{p}(d_1) = \\
&= fpr + \widehat{p}(d_1)\cdot(tpr - fpr)
\end{aligned}$$

what leads to the estimation of the *a priori* probability of the positive class as:

$$\widehat{p}(d_1) = \frac{\widehat{p}(\widehat{d_1}) - fpr}{tpr - fpr} \qquad (1)$$

It is assumed that there is no fundamental variation in the fpr and tpr characteristics between the training and testing distributions because the within-class densities ($p(x|d_i)$) do not change from the training to the new data sets [20]. Therefore the confusion matrix can be estimated from the training data set, as the class labels are not available on the test set. This estimation can be made by using techniques such as stratified k-fold cross-validation. This value of k is recommended to be as high as possible. When performing 10-fold cross-validation, for example, each classifier is trained on 90% of the data, which could be a

Complimentary Contributor Copy

substantial difference in the early part of the learning curve if the positive cases are scarce in the data set. For this reason, 40 or 50-fold-cross-validation should be used instead [13].

If we are dealing with a problem which has n classes, the following system of n linear equations with respect to $p(\widehat{d_j})$ should be solved in order to estimate the new class prior probabilities of the n classes:

$$\widehat{p}(\widehat{d_i}) = \sum_{j=1}^{n} \widehat{p}_t(\widehat{d_i}|d_j)\widehat{p}(d_j), \quad j = 0, 1, \ldots, n \tag{2}$$

where $\widehat{p}(d_j)$ is the estimation of the *a priori* probability of the class j and $\widehat{p}(\widehat{d_i})$ is the observed class probability by looking at the classifier labels $\widehat{d}$.

The solution of the expression (1) (or the equation system (2)), however, can be non-consistent with the basic probability laws (i.e., values outside the interval $[0,1]$) as it has been highlighted in [13]. In a binary problem, it is suggested to clip the negative values to zero and fix the probability of the other class to one.

Based on this Adjusted Count (AC) method, Forman also proposes the Median Sweep (MS) method [12]. Briefly, it can be described as follows: first, several confusion matrices are computed for different classification thresholds; Then, the AC method is applied for each matrix and finally, the class distribution estimation is computed as the median of the estimations derived from each confusion matrix.

2.2. Estimation Based on the Posterior Probability

Given a model whose outputs $\widehat{y}_i$ provide estimates of posterior probabilities, Saerens et al. [20] propose an iterative procedure based on the EM algorithm in order to adjust the classifier outputs for the new deployment conditions without re-training the classifier. This is carried out by indirectly computing the new class prior probabilities, what is the goal in this chapter.

The model's outputs $\widehat{y}_i^k$ yields an approximation of the *a posteriori* probabilities of each class, while the class frequencies in the training set are an estimation of the *a priori* probabilities, so we initialize the prior and *a posteriori* probabilities with them:

$$\widehat{p}^{(0)}(d_i|x_k) = \widehat{y}_i^k \tag{3}$$

$$\widehat{p}^{(0)}(d_i) = \frac{N_t^i}{N_t} \tag{4}$$

Consider $\widehat{p}^{(r)}(d_i)$ the estimation of the new *a priori* probabilities and $\widehat{p}^{(r)}(d_i|x_k)$ the new *a posteriori* probabilities at the $r-th$ iteration of the algorithm. These estimations are given by equations (5) and (6) respectively.

$$\widehat{p}^{(r)}(d_i) = \frac{1}{N}\sum_{l=1}^{N} \widehat{p}^{(r-1)}(d_i|x_k) \tag{5}$$

Complimentary Contributor Copy

$$\widehat{p}^{(r)}(d_i|x_k) = \frac{\dfrac{\widehat{p}^{(r)}(d_i)}{\widehat{p}^{(0)}(d_i)}\widehat{p}^{(0)}(d_i|x_k)}{\sum_{j=1}^{M}\dfrac{\widehat{p}^{(r)}(d_j)}{\widehat{p}^{(0)}(d_j)}\widehat{p}^{(0)}(d_j|x_k)} \; . \qquad (6)$$

This procedure is repeated during R iterations, or until the difference between the $r-th$ iteration and the $(r-1)-th$ iteration is lower than a certain threshold (e.g., 10^{-4}).

This method is called the Posterior Probability method (PP).

2.3. Estimation Based on the Hellinger Distance

As it has been mentioned before, we focus on problems where the within-class conditional densities are fixed, but the class prior probabilities may shift after the classification model is generated. In this case, the joint probabilities $p(x,d_0)$ and $p(x,d_1)$ also vary and so the unconditional density $p(x)$, as well as the posterior probabilities $p(d_0|x)$ and $p(d_1|x)$.

The effect of shifting class distributions on the data distribution $p(x)$ for a binary classification problem is illustrated in figures 1 and 2. In this example each class is defined by a univariate Gaussian distribution.

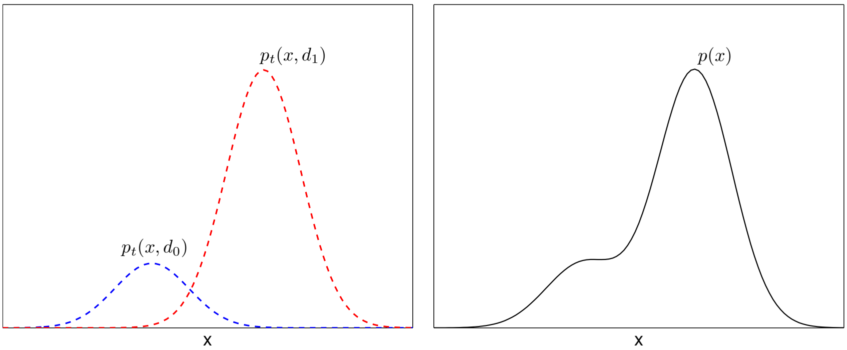

Figure 1. Training data. Joint probabilities $p_t(x,d_0)$ and $p_t(x,d_1)$ (left) and unconditional density $p_t(x)$ (right) for prior class probabilities $(p_t(d_0),p_t(d_1))$ equal to $(0.2,0.8)$.

The joint probabilities $p_t(x,d_0)$ and $p_t(x,d_1)$ for the training dataset with class prior probability $(p_t(d_0),p_t(d_1))$ and the data density $p(x)$ are shown in Fig. 1, while Fig. 2 plots the data distribution for the test set when the prior probabilities have changed $(p(d_0) \neq p_t(d_0), p(d_1) \neq p_t(d_1))$. It is clear that this shift in class proportions, implies also a significant change in the data distribution $p(x)$.

Generating several validation data sets with different prior probabilities, and calculating the differences between these sets and the test set data distributions would allow us to detect these changes and therefore, estimate the new class proportions (the proportion of the validation set which would minimize that difference).

When it comes to measure the difference between two probability distributions, the Kullback-Leibler divergence D_{KL} [17] becomes the most widely used option.

Complimentary Contributor Copy

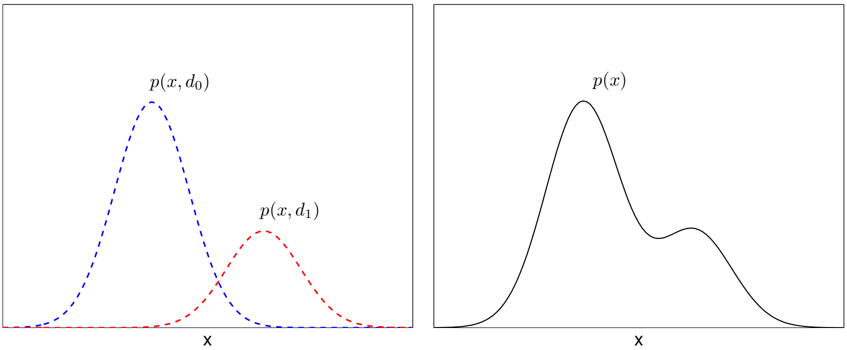

Figure 2. Test (future) data. Joint probabilities $p(x,d_0)$ and $p(x,d_1)$ (left) and unconditional density $p(x)$ (right) for prior class probabilities $(p(d_0), p(d_1))$ in the test set equal to $(0.7, 0.3)$.

The KL divergence between probability distributions p and q on a finite set $\mathscr{X}$ is given by

$$D_{KL}(p||q)) = \sum_{x \in \mathscr{X}} p(x) \log \frac{p(x)}{q(x)} \tag{7}$$

This measure is always non-negative, taking values from 0 to ∞, and $D_{KL}(p||q)) = 0$ if $p = q$. Strictly speaking, however, the KL divergence is not a distance, since (a) in general it is asymmetric $(D_{KL}(p||q)) \neq D_{KL}(q||p))$ and (b) it does not satisfy the triangle inequality. The fact that it is not defined when $q(x) = 0$ limits also its use in certain applications where these situations arise.

The Kullback-Leibler divergence as well as the χ^2 measure and the Hellinger Distance are particular cases of the family of f-Divergences [8] used for measuring the divergence between distributions. χ^2 measure and, as it has previously been pointed out, KL divergence are both asymmetric, and not strictly distance metrics, which makes the Hellinger Distance very appealing for our purpose. Recently, it has been receiving attention in the machine learning community in order to detect failures in classifier performance due to shifts in data distributions. In particular, Cieslak and Chawla [7] have shown that the HD measure is very effective in detecting breakpoints in classifier performance due to shifts in class prior probabilities. Here, we address the problem of class distribution estimation following a HD-based approach.

The HD between two probability density functions $q(x)$ and $p(x)$ can be expressed as

$$H(q,p) = \sqrt{\int \left(\sqrt{q(x)} - \sqrt{p(x)}\right)^2 dx} \tag{8}$$

where HD is non-negative, bounded (it takes values from 0 to $\sqrt{2}$) and it is symmetric (i.e., $H(q,p) = H(p,q)$). Additionally, it is defined for whatever value of $p(x)$ and $q(x)$ and does not make any assumptions about the distributions themselves.

Similarity between the training data distribution and future distributions in the discrete case can also be measured with HD by converting them into binned distributions with a

Complimentary Contributor Copy

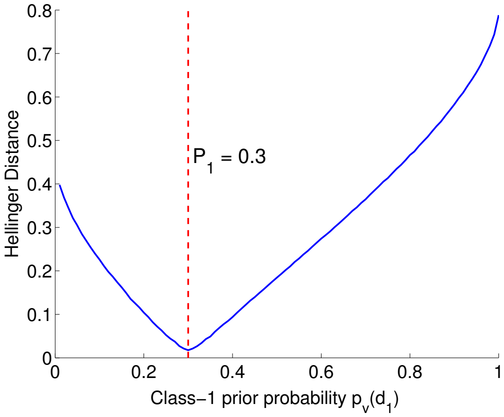

Figure 3. Hellinger distance between the test data distribution and different validation data distributions. The dashed line is the class-1 prior probability of the test data set.

probability associated with each of the b bins. The HD between the training data T and the unlabeled test data U with n_f features is then calculated as

$$H(T,U) = \frac{1}{n_f}\sum_{f=1}^{n_f} H_f(T,U) \tag{9}$$

where the distance between T and U according to feature f is computed as

$$H_f(T,U) = \sqrt{\sum_{i=1}^{b}\left(\sqrt{\frac{|T_{f,i}|}{|T|}} - \sqrt{\frac{|U_{f,i}|}{|U|}}\right)^2} \tag{10}$$

Note that b is the number of bins, $|T|$ the total number of training examples and $|T_{f,i}|$ the number of training examples whose f feature belongs to bin i. Similarly, $|U|$ and $|U_{f,i}|$ correspond to the same characteristics for the test set.

Let us go back to the problem previously presented with its test data distribution depicted in Fig.2. It corresponds to a test set with class prior probabilities $p(d_0) = 0.7$ and $p(d_1) = 0.3$. Fig. 3 plots the Hellinger distance between the distribution of this test data set and different data distributions obtained from the available training data set by subsampling it. These validation data sets differ in the class distributions (from $p_v(d_1) = 0$ to $p_v(d_1) = 1$). Note that the minimum HD is achieved for that validation data set which has the same *a priori* probabilities as the test set $(p_v(d_1) = 0.3)$.

We address the problem of estimating the class proportions of a new unlabeled data set by finding the calibration data distribution which has the minimum distance to the former. These artificially generated distributions can be extracted from the available training data set either by stratified sub-sampling or over-sampling the examples accordingly.

Data sparseness is a problem that usually have to be faced in real practical applications. In real life training data sets are very likely not to be fully representative in all regions of the n_f (number of features) dimensional space. When this happens, the curve in Figure 3 (which has been obtained from a large enough data set) would be noisy, like the ones represented

Complimentary Contributor Copy

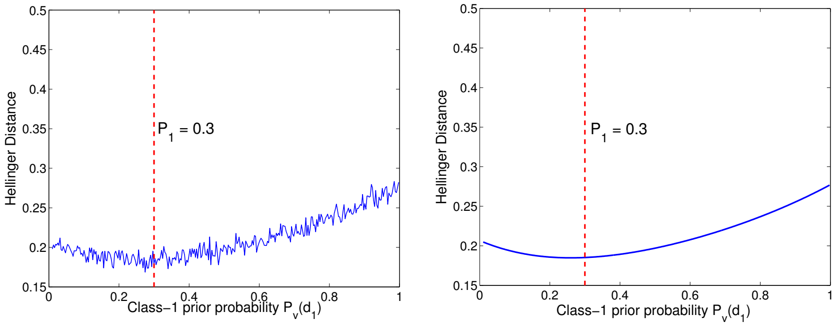

Figure 4. Sperm cell data set. Hellinger distance (HD) between the distributions of a test set with $p(d_1) = 0.3$ and different calibration sets (left). The resulting curve is fitted on the right to a polonomial of degree 5 (right).

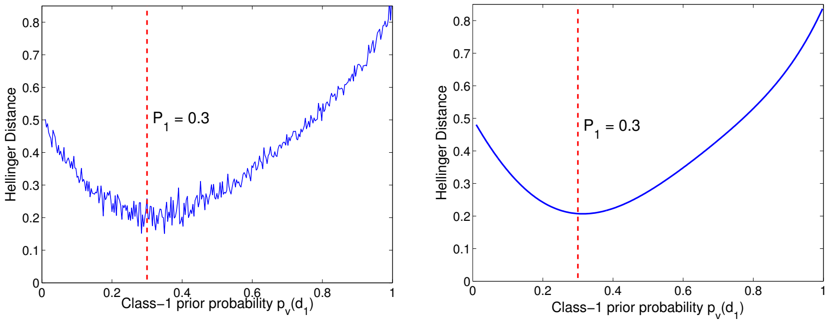

Figure 5. Sperm cell data set. Hellinger distance (HD) between the distributions of the outputs of a test set and different calibration sets given by a classifier (left). Once again the curve is fitted on the right to a polonomial of degree 5 (right).

in figures 4 and 5 on the left. This problem might be partially solved by filtering the curve, downsampling it (e.g., by computing the median among that value and the nearest points), or fitting the resulting curve with a polinomial of degree n (see figures 4 and 5 on the right).

The difference between a calibration data set (with known labels) and the test set can be measured either by computing the HD between both data distributions or by computing the HD between the outputs assigned by a classifier that provides posterior probability estimates (e.g., a neural network). In this case, the problem is simplified because distributional divergences are measured with data defined in a one-dimensional space (for a two class problem) or in a $L-1$ space for a general multi-class problem with L classes. Figures 4 and 5 shows a comparison of the HD between distributions and between the distributions of the same data sets, respectively.

Finally, the method to compute the Hellinger Distance between two discrete distributions T and V is shown in Algorithm 1. The proposed prior probability estimation method based on it is shown in Algorithm 2.

Complimentary Contributor Copy

Algorithm 1 Compute_HD (V, T, b)

```
Require: $N_{features}(V) = N_{features}(T)$, number of bins $b > 0$
$HD = 0$, $n_f$ = number of features in $V$ and $T$
for $i = 1$ to $n_f$ do
    $h = 0$
    Combine $V$ and $T$, and discretize into $b$ bins
    $nb_T \leftarrow$ number of elments of $T$ in each bin
    $n_T \leftarrow$ number of elments of $T$
    $nb_V \leftarrow$ number of elments of $V$ in each bin
    $n_V \leftarrow$ number of elments of $V$
    for $j = 1$ to $b$ do
        $h = h + \left( \sqrt{\frac{nb_T[j]}{n_T}} - \sqrt{\frac{nb_V[j]}{n_V}} \right)^2$
    end for
    $HD = HD + \sqrt{h}$
end for
return $\frac{HD}{n_f}$
```

Algorithm 2 Estimate_HD (X_{Valid}, L_{Valid}, X_{Test}, b)

```
Require: Training set labels $L_{Valid} = \{d/d \in \{0,1\}\}$, $N_{feats}(X_{Valid}) = N_{feats}(X_{Test}) \geq 1$,
    $N_{elems}(X_{Valid}) = N_{elems}(L_{Valid})$ and num. bins $b > 0$
for $i = 0.01$ to 1 in small steps do
    Extract a subset $V$ from $X_{Valid}$ with a proportion $i$ of positives
    $HD\_curve[i] = Compute\_HD(V, X_{Test}, b)$
end for
return $\widehat{p} = arg\ min(HD\_curve)$
```

3. Experimental Study

An experimental illustration of the methods explained in Section 2. will be shown in this section. We have assessed the performance of the CC, AC, MS, and PP methods, as well as the mentioned approaches that rely on the Hellinger distance.

The data used in these experiments have been obtained from a real boar semen quality control application based on computer vision[2]. This dataset has 1861 instances: 951 damaged (class 1) and 910 intact (class 0) spermatozoon acrosomes. The acrosome is a membrane that is over the spermatozoon's head and allows the penetration into the egg. For this reason, if a sample has a high proportion of damaged acrosomes, it will be useless for fertilization purposes. Some texture descriptors derived from the Discrete Wavelet Transform (DWT) have been used. Each image is characterized by a vector of 20 features which are known as Wavelet Co-occurrence Features [3]. For further details about how images have been aquired and processed we refer the interested reader to [16].

[2] We would like to thank CENTROTEC for providing us the semen samples and for their colaboration in the acquisition of the images.

Complimentary Contributor Copy

A back-propagation Neural Network with one hidden layer and a logistic sigmoid transfer function for the hidden and the output layer has been used in the classification stage. Learning was carried out with a momentum and adaptive learning rate algorithm. The training is carried out minimizing a loss function in order to take the classifier outputs as estimates of posterior probabilities [4]. The loss function that has been used in these experiments is the mean square error. Data were normalized with zero mean and standard deviation equal to one.

As a previous step, 10-fold cross validation with several training cycles and neurons in the hidden layer has been carried out in order to determine the performance of the network with each different configuration. A network with 2 neurons in the hidden layer trained during 400 cycles proved to be the best one, in terms of both the miss classification rate (which was 4.27%) and simplicity, so we have used this configuration in the following experiments.

The mismatch between the real class distribution and the estimation provided by the different approaches assessed in this work is measured by means of the Mean Relative Error (MRE). It focuses on the possitive class (the class of interest) and measures the importance of the error (i.e., it is not the same to have an absolute error of 1% when the *a priori* probability is 5% than when it is 50%). The MRE (measured in %) is defined as follows:

$$MRE(p(d_1),\widehat{p}(d_1)) = \frac{|p(d_1) - \widehat{p}(d_1)|}{p(d_1)}100 \tag{11}$$

where $p(d_1)$ and $\widehat{p}(d_1)$ are the true and the estimated *a priori* probabilities of class 1, respectively.

3.1. Comparison of Methods

The performance of the methods has been evaluated in a test set for a proportion of class-1 examples that varies between 0.05 and 0.50 with a fixed set size of 280 elements. The class proportions in the training set are always balanced, and it has the 70% of the minority class of the whole dataset. Both training and test set are always disjoint and randomly extracted amongst the whole dataset.

For the estimation of each different test set proportion, 100 runs were conducted in order to avoid random effects. These runs result from the extraction of 20 random training sets and, for each of them, 5 sets with the desired distribution extracted from the remaining examples are tested. The final results are the average of these 100 runs.

First of all, we have compared the methods based on the Hellinger Distance using the data distribution $p(x)$ and the classifier's outputs distribution $p(y)$, both with and without fitting the curve, in order to compare their performance. In order to avoid adjusting the bins in the HD based methods, each one is applied for a number of bins varying from 10 to 110 in steps of 10, and the final estimated *a priori* probability is the median of these 11 estimations.

Table 2 shows the relative errors of the estimations, as well as their ranks when estimating the prior probabilities of several sets with different proportions of class-1 elements. *HDx* and *Adj. HDx* are the HD of the data distribution itself and the HD obtained by fitting

Complimentary Contributor Copy

the curve to a polynomial of degree 5, respectively (as it is shown in Figure 4). Similarly, *HDy* and *Adj. HDy* are conducted with the outputs of a classifier (see figure 5).

Table 2. MRE (%) of the methods based on the Hellinger Distance.

Test set class-1 prop.	HDx	Adj. HDx	HDy	Adj. HDy
0.05	68.03 (4)	45.15 (3)	21.35 (2)	15.97 (1)
0.15	13.85 (3)	23.59 (4)	7.46 (2)	5.83 (1)
0.25	7.35 (3)	14.23 (4)	3.58 (2)	3.53 (1)
0.35	6.73 (3)	9.47 (4)	2.50 (2)	2.16 (1)
0.45	7.22 (4)	4.23 (3)	1.97 (1)	3.44 (2)
Avg. Rank	3.4	3.6	1.8	1.2

According to these results, the Hellinger Distance is a more reliable method when it is computed using the outputs instead of the data itself, and, what is more important, it performs even better when the HD curve is fitted to a polynomial of degree 5 instead of taking the minimum of the curve itself.

Next, we compare this method with CC, AC, MS, and PP. The confusion matrices required for AC and MS were estimated from the training set by 50-fold cross validation, as suggested in [13]. Table 3 shows the relative errors and the ranks of these methods.

Table 3. MRE (%) of CC, AC, MS, PP, and HDy with adjusted curve.

Test set class-1 prop.	CC	AC	MS	PP	Adj. HDy
0.05	81.93 (5)	20.85 (3)	24.28 (4)	11.39 (1)	15.97 (2)
0.15	20.38(5)	6.25 (3)	7.66 (4)	4.61 (1)	5.83 (2)
0.25	8.81(5)	3.26 (2)	3.63 (4)	2.74 (1)	3.53 (3)
0.35	4.10(5)	2.30 (3)	2.36 (4)	2.04 (1)	2.16 (2)
0.45	1.65(4)	1.56 (2)	1.65 (3)	1.49 (1)	3.44 (5)
Avg. Rank	4.8	2.6	3.8	1	2.8

Results clearly show the benefits from using an estimation method instead of relying on the naïve Classify and Count (CC), as expected.

In these experiments, the average ranks show that PP clearly outperforms the others. As suggested by Saerens, this iterative procedure performs better when the posterior probability estimates yielded by the classfier are well approximated [20]. The performance of AC and HDy is quite similar, and better than MS. Thus, for a test proportion of 0.15, the best method has been PP, with a MRE of 4.61%. Adj. HDy and AC yields errors of 5.83% and 6.25%, respectively, while the error of MS is higher than 7.50%. Finally CC is the worst, yielding a relative error of 20.38%.

In the next section, we will study the robustness of the methods with respect to the performance of the classifier.

Complimentary Contributor Copy

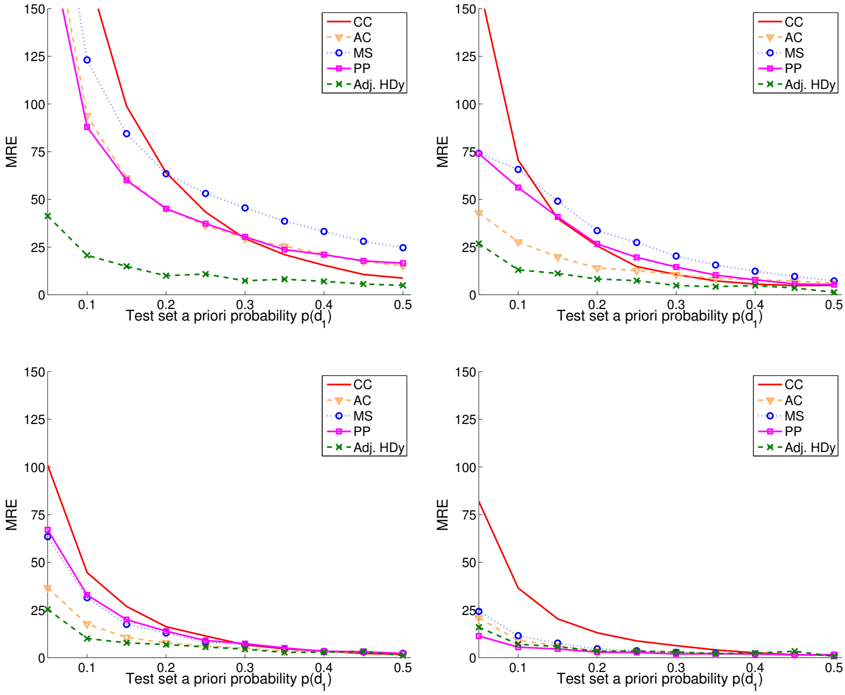

Figure 6. Variation of the MRE with each method when using networks trained during 100, 125, 160, and 400 cycles (from top to bottom and from left to right).

3.2. Robustness

According to Table 3, PP has been the best quantification method when using a neural network trained during 400 cycles. In this section, we explore how this quantification is affected by the classifier performance underlying all the estimation methods. To test this, we used networks trained with a lower number of cycles – in order to obtain a poorer performance when yielding the *a posteriori* probability estimates –. The dataset and number of runs are the same as explained in the previous section. We have carried out experiments with networks trained during 100, 125, 160, and 400 cycles. The aim of these new quantifications is to find out how important is the tuning of the classifier in the estimation of *a priori* probabilities.

Figure 6 shows the evolution of the MRE (in %) for each different number of training cycles.

Although the PP method gives the best estimations for an optimal classifier, the pictures show that its performance is very sensitive to changes in the classifier training conditions. Thus, for a network trained with 100 cycles, the MRE may be higher than 150%, while the maximum relative error when the network is trained during 400 cycles is around 11%.

It is clear that *Adj. HDy* outperforms the other quantification methods when training the network during 100, 125, and 160 cycles, specially with low proportions of class-1 elements in the test set, when the performance of the network is quite poor.

Moreover, it can be observed that the error provided by the *Adj. HDy* method hardly deteriorates when the base classifier performance does. In fact, the maximum MRE is around

Complimentary Contributor Copy

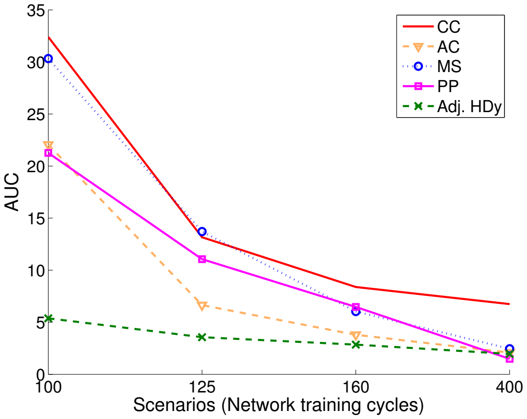

Figure 7. Evolution of the area under curves in figure 6 of each method along the evaluated scenarios.

25% when the networks are trained with 125 and 160 cycles, and around 40% when the number of training cycles is 100. In order to illustrate this, we have computed the areas under the curves in Figure 6 as a measure of the overall performance of the estimation methods. The comparison is shown in figure 7. As the curves have been plotted by joining the MRE of 10 different test distributions, the AUC have been computed by using the numerical integration based on the trapezoid's rule.

It is clear that the *Adj. HDy* method is more robust than the others with respect to the changes in the classifier performance. Therefore, we can conclude that the performance of this method based on the Hellinger distance does not strongly depend on the performance of the network.

4. Conclusion

In this work, we study the problem of automatically quantifying the proportion of data from different classes in a supervised classification environment. The class distribution estimation helps to adapt a classifier to an environment that shows a shift in the class prior probabilities with respect to the training data set; in order to prevent a drop in classifier performance, but it is also an important task by itself, as we have pointed out.

We present and describe quantification methods based on: (a) the classifier confusion matrix (AC, MS), (b) the posterior probability estimations provided by the classifier (PP) and (c) distributional divergence measures (HD). All these techniques are illustrated on a real empirical study based on computer vision that quantifies the proportion of sperm cells with damaged/intact acrosome in a given boar semen sample. Results show that all techniques get a significant improvement with respect to the naive approach of classify and count (CC) and highlight the robustness of the HD method against the classifier tuning, as well as the superiority of the PP method when the classifier provides good posterior probability estimates. There are many works that can be done on this particular field like developing the selection or fusion of quantification techniques, in order to get an improved performance.

Complimentary Contributor Copy

References

[1] R. Alaiz-Rodríguez, A. Guerrero-Curieses, and J. Cid-Sueiro. Minimax regret classifier for imprecise class distributions. *Journal of Machine Learning Research*, **8**:103–130, January 2007.

[2] Rocío Alaiz-Rodríguez, Enrique Alegre, Víctor González-Castro, and Lidia Sánchez. Quantifying the proportion of damaged sperm cells based on image analysis and neural networks. In *Proceedings of the 8th conference on Simulation, modelling and optimization*, pp. 383–388. WSEAS, 2008.

[3] S. Arivazhagan and L. Ganesan. Texture classification using wavelet transform. *Pattern Recognition Letters*, **24**(9–10):1513–1521, June 2003.

[4] C. M. Bishop. *Neural networks for pattern recognition.* Oxford University Press, 1996.

[5] Yee Seng Chan and Hwee Tou Ng. Word sense disambiguation with distribution estimation. In *Proceedings of the IJCAI05*, pp. 1010–1015, 2005.

[6] Yee Seng Chan and Hwee Tou Ng. Estimating class priors in domain adaptation for word sense disambiguation. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pp. 89–96. Association for Computational Linguistics, 2006.

[7] David A. Cieslak and Nitesh V. Chawla. A framework for monitoring classiffiers performance: When and why failure occurs? *Knowledge and Information Systems*, **18**(1):83–108, January 2009.

[8] I. Csiszar and P. Shields. *Information Theory and Statistics: A Tutorial (Foundations and Trends in Communications and Information Theory).* Now Publishers Inc, December 2004.

[9] Chris Drummond and Robert C. Holte. Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, **65**(1):95–130, 2006.

[10] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wiley and Sons, 2001.

[11] G. Forman. Counting positives accurately despite inaccurate classification. In *European Conference on Machine Learning*, pp. 564–575, 2005.

[12] G. Forman. Quantifying trends accurately despite classifier error and class imbalance. In *Principles and Practice of Knowledge Discovery in Databases*, pp. 157–166, 2006.

[13] George Forman. Quantifying counts and costs via classifcation. *Data Mining and Knowledge Discovery*, 17(2):164–206, October 2008.

Complimentary Contributor Copy

[14] George Forman, Evan Kirshenbaum, and Jaap Suermondt. Pragmatic text mining: Minimizing human effort to quantify many issues in call logs. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 852–861. ACM, 2006.

[15] Víctor González-Castro, Rocío Alaiz-Rodríguez, Laura Fernández-Robles, R. Guzmán-Martínez, and Enrique Alegre. Estimating class proportions in boar semen analysis using the hellinger distance. In *Trends in Applied Intelligent Systems*, volume 6096 of *Lecture Notes in Computer Science*, pp. 284–293. Springer, 2010.

[16] Maribel Gonzlez, Enrique Alegre, Roco Alaiz, and Lidia Snchez. Acrosome integrity classification of boar spermatozoon images using dwt and texture techniques. In *VipIMAGE -Computational Vision and Medical Image Processing*, pp. 165–168. Taylor and Francis Group London, 2007.

[17] S. Kullback and R. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, **22**(1):79–86, 1951.

[18] Patrice Latinne, Marco Saerens, and Christine Decaestecker. 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. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 298–305, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[19] Doaa Mahmoud-Ghoneim, Mariam K Alkaabi, Jacques D de Certaines, and Frank-M. Goettsche. The impact of image dynamic range on texture classification of brain white matter. *BMC Medical Imaging*, **8**:18, 2008.

[20] M. Saerens, P. Latinne, and C. Decaestecker. Adjusting a classifier for new a priori probabilities: A simple procedure. *Neural Computation*, **14**:21–41, January 2002.

[21] Lidia Sánchez, Víctor González, Enrique Alegre, and Rocío Alaiz. Classification and quantification based on image analysis for sperm samples with uncertain damaged/intact cell proportions. In *Proceedings of the 5th international conference on Image Analysis and Recognition*, volume 5112 of *Lecture Notes in Computer Science*, pp. 827–836, July 2008.

[22] Sergios Theodoridis and Konstantinos Koutroumbas. *Pattern Recognition*. Academic Press, third edition, 2006.

[23] Jack Chongjie Xue and Gary M. Weiss. Quantification and semi-supervised classification methods for handling changes in class distribution. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 897–906, New York, NY, USA, 2009. ACM.

[24] Zhihao Zhang and Jie Zhou. Transfer estimation of evolving class priors in data stream classification. *Pattern Recogn.*, **43**(9):3151–3161, 2010.

Complimentary Contributor Copy