Estimating Class Proportions in Boar Semen Analysis using the Hellinger Distance

R. Alaiz-Rodríguez*, V. González-Castro, L. Fernández-Robles and E. Alegre

Dpto. de Ingeniería Eléctrica y de Sistemas *rocio.alaiz@unileon.es
University of León, Campus de Vegazana s/n, 24071 León, Spain

Abstract. Advances in image analysis make possible the automatic semen analysis in the veterinary practice. The proportion of sperm cells with damaged/intact acrosome, a major aspect in this assessment, depends strongly on several factors, including animal diversity and manipulation/conservation conditions. For this reason, the class proportions have to be quantified for every future (test) semen sample. In this work, we evaluate quantification approaches based on the confusion matrix, the posterior probability estimates and a novel proposal based on the Hellinger distance. Our information theoretic-based approach to estimate the class proportions measures the similarity between several artificially generated calibration distributions and the test one at different stages: the data distributions and the classifier output distributions. Experimental results show that quantification can be conducted with a Mean Absolute Error below 0.02, what seems promising in this field.

1 Introduction

Artificial insemination techniques provide great advantages in the veterinary field. On the one hand, they allow farmers to work with a reduced number of animals, saving both time and money. On the other hand, it makes possible to get better individuals each generation. Companies that sell semen samples to farmers need to guarantee that they will be optimal for fertilization. There is a direct relationship between sperm fertility and the state of the acrosome: a sample containing a high percentage of spermatozoa with a damaged acrosome when is recollected will no be useful for fertilizing purposes. This assessment is traditionally carried out manually, using stains, which makes this process tedious, time-consuming, costly and what is more important, non objective.

There are some works that automatically characterize images of spermatozoa according to their membrane integrity by means of texture descriptors (e.g. [2]). Features are evaluated in terms of the achieved classification accuracy. In this field, however, the aim is to estimate the proportion of damaged cells with no concerns about the individual classification of each one.

Unlike a typical supervised learning problem, the class prior probabilities estimated from the labeled training data cannot be considered representative of

[1] This work has been partially supported by the research project DPI2009-08424 from the Spanish Ministry of Education and Science.

future samples since they are subject to vary due to factors like the animal/farm variability, or the manipulation and conservation conditions. It is well known that a mismatch between the test (real) class prior probabilities and those for which the classifier has been optimized, leads to suboptimal solutions. Different works have tackled this problem (e.g. [11, 10]) from several perspectives, but always with the goal of improving the individual classification performance. Linguistics [5] or medicine [12] are just some fields where it has been applied.

To the best of our knowledge, only a few works cope with the problem of estimating the *a priori* probabilities – the actual class distribution – of unlabeled data sets (also known as quantification). They mostly aim to the analysis of a company's technical support logs where there are changes in trends[8]. Some previous work has also been conducted on veterinary applications [1] with promising results evaluated, though, on very small data sets.

The techniques proposed in the literature to estimate the class proportions are either based on the classifier confusion matrix [8] or on the posterior probability estimations provided by the classifier [1].Forman has also explored a method based on the estimation of the class conditional probability densities [8], but it turned out to be outperformed by simple methods that rely on the confusion matrix.

When there is a shift in class prior probabilities between training and test sets, the data distributions as well as the *a posteriori* probability distribution also change. Our proposal to estimate the class distributions is based on indirectly quantifying the similarity between distributions (comparing the test set distribution with the distribution of several generated labeled sets). A distributional divergence metric (Hellinger Distance [7]) is applied at different stages of the classification process: (1) between data distributions and (2) between the *a posteriori* probability distributions.

The goal of this paper is: (a) to explore an information theoretic approach to automatically quantify the class distribution in a given semen sample and (b) to evaluate some quantification methods and check whether or not reliable estimations can be achieved for this specific application.

In this work, we consider an image data set of boar sperm samples and use a back-propagation neural network as a classifier.The rest of this paper is organized as follows: Sections 2 and 3 present the class distribution quantification methods assessed in this work. Experimental results are shown in Section 4 and finally Section 5 summarizes the main conclusions.

2 Class Distribution Estimation

Consider a binary classification problem with a calibration labeled data set $T = \{(\mathbf{x}^k, d^k), k = 1, \ldots, K\}$ where $\mathbf{x}^k$ is a feature vector, d^k is the class label with $d \in \{0, 1\}$[1] and $\widehat{Q}_i = P\{d = i\}$ is the class prior probability estimated from T. A classifier can be built using the available data in two steps: Firstly, it computes a soft output $\widehat{y}^k$ and based on it, makes the final hard decision $\widehat{d}^k \in \{0, 1\}$

[1] Class-1 will also be denoted as the positive class and class-0 as the negative one.

Consider now an unlabeled test data set $U = \{(\mathbf{x}^l), l = 1, \ldots, N\}$ with unknown class distribution P_i and the decisions $\widehat{y}^l$ and $\widehat{d}^l$ provided by the classifier for each instance in the data set U. The naive approach to estimate its class distribution is to count the labels assigned by the classifier, what is referred as Classify & Count (CC) in [8]. It is considered here as a baseline for comparison purposes.

A brief description of proportion estimation methods based on the confusion matrix and the posterior probability estimation are provided in Sects.2.1 and 2.2, respectively. The technique based on the distance between probability distributions is presented in Sect. 3.

2.1 Methods based on the Confusion Matrix

Classifier performance can be summarized by its confusion matrix. Based on it and the ratio of labels assigned by the classifier to each class, different techniques have been proposed in [10, 8] to estimate the unknown proportions of an unlabeled data set. Basically, the estimations are obtained by solving a following system of two (in a binary case) linear equations with respect to $\widehat{P}_i$

The solution of the equation system, however, can be non consistent with the basic probability laws (i.e, values outside the interval $[0, 1]$) as it has been highlighted in [8]. In a binary problem, it is suggested to clip the negative values to zero and fix the probability of the other class to one. This solution may be not satisfactory for real practical binary applications, though and moreover, there is no straightforward solution for general multi-class problems. As in [8] we will refer to this method as Adjusted Count (AC). A related method, Median Sweep (MS), computes several confusion matrices for different classification thresholds and finally, the class proportion is given as the median of the estimations derived from each confusion matrix

2.2 Estimation of priors based on posterior probabilities

In [8] the application of techniques based on the posterior probability estimation is discarded arguing that, under a change in class prior probabilities, the classifier is not optimal anymore. Therefore, any estimation derived from the posterior probabilities given by the classification model would not be reliable.

This problem can be overcome with the following strategy. Given a model whose outputs y_i provide estimates of posterior probabilities, Saerens et al. [10] proposes 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, which is the goal in our work.

We will denote this iterative approach as the Posterior Probability (PP) method, and we refer the interested reader to [1] for more details.

3 Hellinger Distance to estimate proportions

As it has been mentioned before, in this work we focus on problems where the class conditional densities are fixed, but the class prior probabilities may shift after the classifier calibration. When this happens, the joint probabilities $p(x, d = 0)$ and $p(x, d = 1)$ also vary and so the unconditional density $p(x)$ and the posterior probabilities $p(d = 0|x)$ and $p(d = 1|x)$.

Figs 1 and 2 show the effects of shifting class distributions on the data distribution $p(x)$ for a binary classification problem where each class is defined by a univariate gaussian distribution. Fig.1 depicts the joint probabilities $p(x, d = 0)$

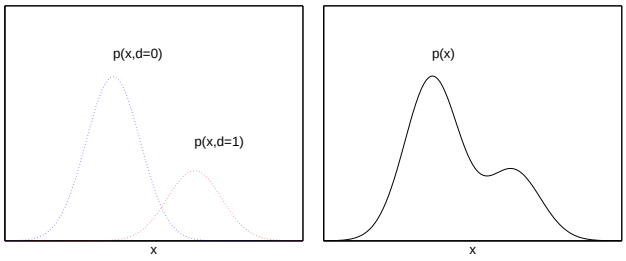

Fig. 1. Training data. Joint probabilities $p(x, d = 0)$ and $p(x, d = 1)$ (left) and unconditional density $p(x)$ (right) for prior class probabilities (Q_0, Q_1) equal to $(0.3, 0.7)$.

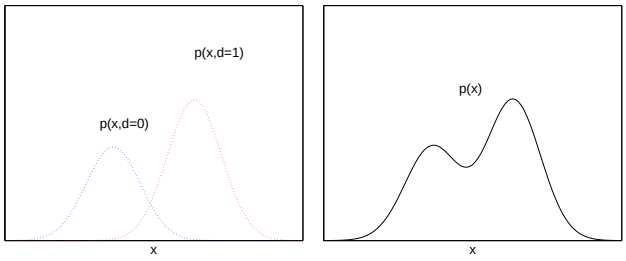

Fig. 2. Test (future) data. Joint probabilities $p(x, d = 0)$ and $p(x, d = 1)$ (left) and unconditional density $p(x)$ (right) for prior class probabilities (P_0, P_1) in the test set equal to $(0.6, 0.4)$.

and $p(x, d = 1)$ for the training dataset with class prior probability (Q_0, Q_1) and the data density $p(x)$. Fig.2 plots the data distribution for the test set when a change in the prior probabilities has taken place. We note that a shift in class proportions from training and test, implies also a significant change in the data distribution. Generating calibration data sets with different prior probabilities

and measuring differences between the calibration and the test data distribution would allow to detect these changes and therefore, estimate the new class proportions.

The Kullback-Leibler divergence as well as the χ^2 measure and the Hellinger distance are particular cases of the family of f-Divergences [7] that measure distributional divergence. Unlike the KL divergence or χ^2 measure that are both asymmetric, and not strictly distance metrics, the Hellinger Distance (HD) has interesting properties that make it 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 [6]. In particular, Cieslak and Chawla have shown that the HD measure is very effective in detecting breakpoints in classifier performance due to shifting class prior probabilities. In this work we address the problem of class distribution estimation following a HD-based approach.

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

$$H(q,p) = \sqrt{\int \left(\sqrt{q(\mathbf{x})} - \sqrt{p(\mathbf{x})}\right)^2 dx} \tag{1}$$

where HD is non negative and bounded (it ranges from 0 to $\sqrt{2}$) and 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 can be measured with HD by converting them into binned distributions with a 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

$$H(T,U) = \frac{1}{n_f} \sum_{f=1}^{n_f} H_f(T,U) \tag{2}$$

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

$$H_f(T,U) = \sqrt{\sum_{i=1}^{b} \left(\sqrt{\frac{|T_{f,i}|}{|T|}} - \sqrt{\frac{|U_{f,i}|}{|U|}} \right)^2} \tag{3}$$

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 statistics for the test set.

Let us go back to the problem presented previously with its test data distribution depicted in Fig.2. It corresponds to a test set with class prior probabilities $P_1 = 0.4$ and $P_0 = 0.6$. Fig.3 plots the Hellinger distance between this test data distribution and different data distributions obtained from the available training data set. These calibration data sets differ in the class distributions (from $Q_1 = 0$ to $Q_1 = 1$). Note that the minimum HD is achieved for that calibration data set with the same a priori probabilities as the test set ($Q_1 = 0.4$).

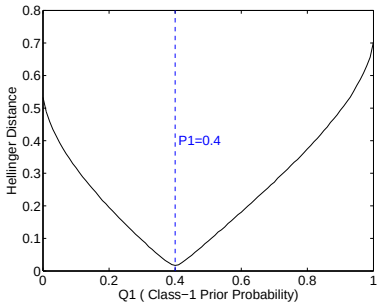

Fig. 3. Hellinger distance between the test data distribution and different calibration data distributions. The dashed line is the class-1 prior probability of the test data set.

In this work we address the problem of estimating the class proportions of a new unlabeled data set by finding the calibration data distribution for which the distance with the test data distribution is minimum. These artificially generated distributions can be extracted from the available training data set either by stratified sub-sampling, over-sampling or weighting the examples accordingly.

In real practical applications we usually face the problem of data sparseness. It is not uncommon to have a training data set that is not fully representative in all regions of the nf (number of features) dimensional space. In these cases, the curve in Fig. 3 (obtained from a large enough data set) turns noisy like the ones represented in Fig.4. This has been partially solved by downsampling the resultant curve and estimate the HD for a given training data set distribution by computing the median among that value and the nearest four points.

Note that we can measure the difference between a calibration set (with known labels) and the test set either by computing the HD between both data distributions (see Fig.4, left) or by computing the HD between the outputs assigned by a classifier that provides a posteriori probability estimates (see Fig.4, right).

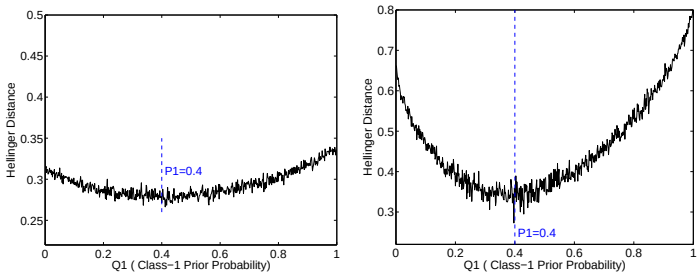

Fig. 4. Sperm cell data set. Hellinger distance (HD) between a test set and different calibration sets. HD between data distributions $p(x)$ (left). HD between posterior probability distribution (right).

Finding the differences between the classifier output distribution (the posterior probabilities) simplifies the problem, 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.

4 Experimental Results

In this section, several techniques that estimate the class distribution are evaluated in the context of a boar semen quality control application. We assess the performance of three quantifying approaches that rely on the Hellinger distance (HD), as well as AC, MS[2], PP and CC methods.

4.1 Sperm cell data set

Experiments are carried out with images of boar sperm samples. The image data acquisition was conducted at CENTROTEC and under the guidance of researchers of the Faculty of Veterinary Sciences from the University of León. We use a data set with 1861 instances: 951 damaged and 910 intact spermatozoon heads. An example of these two acrosome states are shown in Fig. 5.

We use texture descriptors derived from the Discrete Wavelet Transform (DWT) to characterize the images. 20 features per image are derived from the co-occurrence matrix using the Wavelet Co-occurrence Features (WCF) [3]. For further details we refer the interested reader to [9].

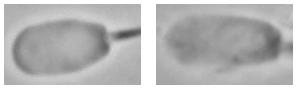

Fig. 5. Grey level images of intact (left) and damaged (right) acrosomes of boar sperm.

4.2 Neural network Classifier

Classification was carried out by means of a back-propagation Neural Network with one hidden layer and a logistic sigmoid transfer function for the hidden and the output layer. Learning was carried out with a momentum and adaptive learning rate algorithm. It is well known that classifier outputs provide estimates of class posterior probabilities when training is carried out minimizing some loss functions such as the mean square error used in this work [4].

Data were normalized with zero mean and standard deviation equal to one. The neural network architecture as well as the number of training cycles were determined by 10-fold cross validation. A two-node hidden layer network with 400 training cycles lead to the optimal configuration evaluated in terms of the overall misclassification rate, which was 4.27%.

[2] According to the methods based on the confusion matrix, we will only show the results for the AC method, since AC and MS presented very similar performance.

4.3 Performance metrics

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 Absolute Error (MAE) and the Mean Relative Error (MRE).

MAE focuses on the class of interest (class-1, positive or damaged cell class) and is defined as the absolute value of the difference between its actual prior probability and the estimated one: $MAE(\mathbf{P}, \widehat{\mathbf{P}}) = |P_1 - \widehat{P_1}|$. MRE measures the importance of the error and is defined as follows: $MRE(\mathbf{P}, \widehat{\mathbf{P}}) = |P_1 - \widehat{P_1}|/P_1$.

4.4 Quantification of the damaged sperm cells

The following experiment was designed to assess the quantification methods CC, AC, PP, and three methods based on the Hellinger distance: HD of the inputs to the classifier (HDx), HD of the input data computing the median of the points in the curve (Median HDx) and HD of the outputs of the classifier (HDy).

Performance has been evaluated for a proportion of class-1 examples lower than 0.5 and a fixed set size (600 instances for training and 300 for test). The ratio of images from class-1 in the test set goes from 0.05 to 0.50 in 0.05 steps. For each scenario, results are the average of 5 training sets randomly extracted from the data set and, for each training set, 4 test sets were randomly extracted among the remaining examples. The confusion matrices required for AC were estimated exclusively from the training set by 50-fold cross validation, as suggested in [8].

	MAE			MRE (%)		
Ptest	HDx	Median HDx	HDy25	HDx	Median HDx	HDy25
0.05	0.039	0.024	**0.009**	77.32	48.56	**18.60**
0.10	0.032	0.026	**0.010**	31.70	26.36	**10.00**
0.15	0.023	0.022	**0.007**	15.33	14.42	**4.92**
0.20	0.021	0.022	**0.008**	10.38	11.16	**4.01**
0.25	0.029	0.029	**0.011**	11.40	11.41	**4.56**
0.30	0.031	0.036	**0.010**	10.22	12.13	**3.43**
0.35	0.025	0.027	**0.014**	7.24	7.74	**4.09**
0.40	0.027	0.032	**0.021**	6.75	8.04	**5.30**
0.45	0.044	0.033	**0.017**	9.80	7.27	**3.82**
0.50	0.071	0.054	**0.023**	14.23	10.87	**4.66**

Table 1. MAE and MRE errors of the HDx, Median HDx and HDy25 methods.

In order to evaluate the quantification methods CC, AC, PP and HDy, the neural network described in Sect. 4.2 has been used as classifier. It has been trained with sets which have 25% class-1 elements – the center of our interval of interest –, so they will be referred to as CC25, AC25, PP25, and HDy25. The number of bins taken in the Hellinger-based methods has been fixed to 60. [3]

[3] This parameters was not very sensitive, as long as it was not too high so that there were few or none examples in each bin.

Results using the Hellinger-based methods (HDx, Median HDx and HDy25) are gathered and compared in Table 1. As it can be seen, the estimations given by HDy25 achieved the lowest error, in terms of MAE and MRE no matter the test class distribution. HDy25 allows to estimate the proportions with a MAE at the most of 0.023. Experimental results corroborate that computing the divergence between the output distributions yields better estimations than directly measuring raw data distribution divergences. A problem of data scarceness in the twenty-dimensional input space that does not appear in the one-dimensional output space may be the main reason.

With regard to CC, AC, PP and HDy25 (compared in Table 2), all methods provided satisfactory estimations in terms of absolute errors. Thus, for a test set with a ratio of damaged cells of 0.30, HDy achieves a MRE of 3,43% whereas PP and AC get a value around 4% what sounds very reasonable for the application field. It is remarkable that the absolute errors are quite stable across the different test distributions. It is also noticeable that the PP method works better for low probabilities, while AC method performs better in balanced distributions, and the HDy method is the best in the intermediate range of the interest interval.

	MAE				MRE (%)			
Ptest	HDy25	PP25	AC25	CC25	HDy25	PP25	AC25	CC25
0.05	0.009	**0.006**	0.008	0.018	18.60	**12.45**	16.90	36.33
0.10	0.010	**0.008**	0.009	0.013	10.00	**7.64**	8.88	12.73
0.15	**0.007**	0.009	0.010	0.009	**4.92**	6.13	6.56	6.21
0.20	**0.008**	0.011	0.011	0.009	**4.01**	5.39	5.43	4.33
0.25	**0.011**	0.012	0.012	**0.011**	4.56	4.72	4.66	**4.36**
0.30	**0.010**	0.012	0.012	0.015	**3.43**	4.06	4.00	4.91
0.35	**0.014**	0.015	**0.014**	0.021	4.09	4.21	**3.87**	6.06
0.40	0.021	**0.017**	**0.017**	0.027	5.30	4.26	**4.22**	6.80
0.45	0.017	0.018	**0.016**	0.034	3.82	4.04	**3.61**	7.44
0.50	0.023	0.018	**0.017**	0.040	4.66	3.64	**3.49**	7.93

Table 2. MAE and MRE of the HDy25, CC25, AC25 and PP25 methods.

5 Conclusions and Future Work

This work tackles the problem of automatically quantifying the proportion of sperm cells with damaged acrosome in a given boar semen sample. We follow a computer vision-based approach that describes each cell by texture features and afterwards they are classified by a neural network.

Our novel proposal to estimate the class distributions is based on measuring the discrepancy between probability distributions (data and classifier output distributions) with the Hellinger distance. Satisfactory results for this application have been achieved with a mean absolute error lower than 0.023 for test semen samples with a proportion of damaged cells that vary from 0.05 to 0.50.

Comparisons of HD with AC,MS and PP show that: (a) all techniques get a significant improvement with respect to the baseline approach CC, (b) there is no single method that outperforms all the others in the whole probability range, (c) PP appears to be the best option for very low probabilities and AC when the test data set is balanced, whereas HD should be the choice for proportions in the middle of the interval.

Results suggest that a hybrid solution that either selects or combines the different estimation methods would allow to improve the total performance. Future work will follow this line as well as the study of visual tools that enable the analysis in a two dimensional space of the vast amount of performance results.

References

1. R. Alaiz-Rodríguez, E. Alegre, V. González-Castro, and L. Sánchez. Quantifying the proportion of damaged sperm cells based on image analysis and neural networks. In *SMO'08: Proc. of the 8th conf. on Simulation, modelling and optimization*, pages 383–388, 2008.
2. E. Alegre, V. González-Castro, S. Suárez, and M. Castejón. Comparison of supervised and unsupervised methods to classify boar acrosomes using texture descriptors. In *Proceedings ELMAR-2009*, pages 65–70, September 2009.
3. S. Arivazhagan and L. Ganesan. Texture classification using wavelet transform. *PATTERN RECOGN LETT*, 24(9–10):1513–1521, June 2003.
4. C.M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, 1996.
5. Y. S. Chan and H. T. Ng. Estimating class priors in domain adaptation for word sense disambiguation. In *ACL-44: Proc. of the 21st Int. Conf. on Computational Linguistics*, pages 89–96, 2006.
6. D. Cieslak and N. Chawla. A framework for monitoring classiffiers performance: When and why failure occurs? *KNOWL INF SYST*, 18(1):83–108, January 2009.
7. I. Csiszar and P. Shields. *Information Theory and Statistics: A Tutorial (Foundations and Trends in Communications and Information The)*. Now Publishers Inc, December 2004.
8. G. Forman. Quantifying counts and costs via classifcation. *DATA MIN KNOWL DISC*, 17(2):164–206, October 2008.
9. M. González, E. Alegre, R. Alaiz, and L. Sánchez. Acrosome integrity classification of boar spermatozoon images using dwt and texture techniques. In *International Conference VipIMAGE 2007*. Taylor and Francis, 2007.
10. M. Saerens, P. Latinne, and C. Decaestecker. Adjusting a classifier for new a priori probabilities: A simple procedure. *NEURAL COMPUT*, 14:21–41, January 2002.
11. J. C. Xue and G. M. Weiss. Quantification and semi-supervised classification methods for handling changes in class distribution. In *Proc. of the 15th ACM SIGKDD int. conf. on Knowledge discovery and data mining*, pages 897–906, 2009.
12. C. Yang and J. Zhou. Non-stationary data sequence classification using online class priors estimation. *PATTERN RECOGN*, 41(8):8, August 2008.