A Hierarchical Approach for Multi-task Logistic Regression

Àgata Lapedriza[1], David Masip[2], and Jordi Vitrià[1]

[1] Computer Vision Center-Dept. Informàtica
Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain
{agata,jordi}@cvc.uab.es
[2] Universitat de Barcelona (UB), 08007 Barcelona, Spain
davidm@maia.ub.es

Abstract. In the statistical pattern recognition field the number of samples to train a classifier is usually insufficient. Nevertheless, it has been shown that some learning domains can be divided in a set of related tasks, that can be simultaneously trained sharing information among the different tasks. This methodology is known as the multi-task learning paradigm. In this paper we propose a multi-task probabilistic logistic regression model and develop a learning algorithm based in this framework, which can deal with the small sample size problem. Our experiments performed in two independent databases from the UCI and a multi-task face classification experiment show the improved accuracies of the multi-task learning approach with respect to the single task approach when using the same probabilistic model.

1 Introduction

Automatic pattern classification is one of the most active research topics in the machine learning field. This problem consists in assigning a given instance to a predefined group or class after observing different samples of this group. Examples of these frameworks in scientific areas are medical diagnosis, speech recognition or image categorization.

Statistical procedures have been shown to be a powerful tool to treat these classification problems, where an underlying probability model is assumed in order to calculate the posterior probability upon which the classification decision is made. Nevertheless, in these classical approaches a considerable number of training examples is needed to correctly learn the parameters of the model. For this reason, their application can be not appropriate when the obtention of training samples is difficult.

There are some situations where the estimation of a predictive model can take benefit from the estimation of other related ones. For instance, in a multiple speech recognition problem, we can share information from modelling the speech of different subjects, in handwritten text classification from different writers we could also take benefit from the several related classification tasks. Other examples in the computer vision field are identity verification problems, or related

J. Martí et al. (Eds.): IbPRIA 2007, Part II, LNCS 4478, pp. 258–265, 2007.

tasks in automatic drive guiding problems such as road lane tracking, broken or solid line classification, or direction marks identification. In these examples, each of the considered tasks belong to different problems. Nevertheless it seems clear that they belong to a related domain, where they share common information that can be used to improve the classification accuracies obtained in the single task learning framework.

One of the most important open problems in the statistical classification approach, is the lack of learning samples necessary to properly estimate the parameters of the classifier. Usually, in classification problems the data lays on high dimensional subspaces, being the theoretical number of samples needed exponential in terms of the data dimensionality (known as the curse of dimensionality problem [1]). Recently, it has been proposed a new learning paradigm, the multi-task learning (MTL) [2], that has been shown to mitigate this small sample size problem [3,4]. The MTL approach is based on simultaneously learning a set of related tasks, sharing the hypothesis space of classifiers or assuming some common generative process in the data from each tasks [5,6]. The advantages of MTL have been proved in the recent theory, and can be summarized in: (i) the bias learned in a multiple related task environment is less specific than in a single task problem, resulting in classifiers with less generalization error; (ii) the number of samples needed to simultaneously learn several related tasks sub-linearly decreases as a function of the number of tasks [4]. More recently the idea of multi-task learning has been extended to some of the state of the art classifiers: Evgeniou et al. applied MTL to the SVM [7] and Torralba et al. extended the Adaboost algorithm to the MTL case by sharing the feature space where each weak learned is trained [8].

In this work we propose a hierarchical Multi-task learning approach for the logistic regression model and also extend this idea to the multinomial logistic regression case. Once the model is presented we develop a learning algorithm according to this framework. The paper is organized as follows: in the next section the hierarchical multi-task logistic regression approach is explained in detail as well as the corresponding algorithm and its extension to the multinomial logistic regression case, section 3 describes the performed experiments and section 4 includes the discussion of the results. Finally, section 5 concludes this work.

2 A Hierarchical Learning Approach for Multi-task Logistic Regression

Let be $T_1, ..., T_M$ a set of related binary tasks and $D = \{S_1, ..., S_M\}$ the set of corresponding training data, $S_i = \{(x_n^i, y_n^i)\}_{n=1,..,N(i)}$ such that $x_n^i \in \mathbb{R}^d$, $y_n^i \in \{-1, 1\}$. Consider for each task a logistic regression model, that is, for each T_i we learn a classifier f_i, that will give the probability of the output $y = 1$ according to the i-th task for the input x,

$$f_i(x) = P(y = 1|x, T_i) = \frac{1}{1 + exp(-\mathbf{w}^{(i)} x^T)} \tag{1}$$

where $\mathbf{w}^{(i)} = (w_1^i, ..., w_d^i)$ is the parameters vector of the i-th task. Let be W the parameters matrix, considering all the tasks,

$$W = \begin{pmatrix} w_1^{(1)} & \dots & w_1^{(M)} \\ \vdots & \vdots & \vdots \\ w_d^{(1)} & \dots & w_d^{(M)} \end{pmatrix}$$

To learn the parameters of the model we can apply a negated log-likelihood estimator $L(D, W)$ and impose a prior distribution on the elements of W as a regularization therm, $R(W)$. In that case, the negated log-likelihood estimator for all the tasks T_i is

$$L(D, W) = -log[\prod_{i=1}^{M}[\prod_{n=1}^{N(i)} P(y_i^n | x_i^n, W)]] = -[\sum_{i=1}^{M}[\sum_{n=1}^{N(i)} log(P(y_i^n | x_i^n, W))] \quad (2)$$

and regarding to the regularization therm, most of the current methods use centered Gaussian priors. Then, the elements of the matrix W are obtained by the minimization of the following loss function

$$H(W) = L(D, W) + \frac{1}{\sigma^2}\|W\|_2 \quad (3)$$

where $\sigma \in \mathbb{R}^+$ is the variance of the imposed regularization distribution. This optimization problem can be solved applying any appropriated method, for example a gradient descent algorithm [9].

This method has shown to be efficient in many situations. However, observe that in this presented framework there is no transit of information between the models of the different tasks. Suppose that we want to learn the parameters of the logistic regression for this classification scenario enforcing the different classes to share information, following the principles of MTL. For this purpose, we can impose prior distributions on each row of W in a hierarchical way as follows. Consider the mean vector $\bar{\mathbf{w}} = (\bar{w_1}, ..., \bar{w_d})$ where

$$\bar{w_j} = \frac{\sum_{i=1}^{M} w_j^{(i)}}{M} \quad (4)$$

First, we can impose a Gaussian centered prior to the mean vector $\bar{\mathbf{w}}$ and after that we can enforce that each row of W follows a Gaussian distribution with $\bar{w}_d$ mean. In short, this can be obtained by the minimization of the loss function

$$G(W) = L(D, W) + \frac{1}{\sigma_1^2}\|\bar{\mathbf{w}}\|_2 + \frac{1}{\sigma_2^2}\sum_{i=1}^{M}\|\mathbf{w}^{(i)} - \bar{\mathbf{w}}\|_2 = L(D, W) + R(W) \quad (5)$$

where $L(D, W)$ is again the negated log-likelihood estimator and σ_r^2 are the corresponding variances of the imposed priors, $r = 1, 2$.

2.1 Training Algorithm

Any optimization method that allows to minimize G will yield a training algorithm for our purpose. In this case we can apply a gradient descent algorithm to optimize it given that the loss function in equation 5 is differentiable. More concretely, we have used the BFGS gradient descent method. The principal idea of the method is to construct an approximate Hessian matrix of second derivatives of the function to be minimized, by analyzing successive gradient vectors. This approximation of the function's derivatives allows the application of a quasi-Newton fitting method in order to move towards the minimum in the parameter space.

Thus, we need to compute the partial derivatives

$$\frac{\partial G(W)}{\partial w_k^{(s)}} = \frac{\partial L(W,D)}{\partial w_k^{(s)}} + \frac{\partial R(W)}{\partial w_k^{(s)}} \tag{6}$$

Observe that $R(W)$ can rewritten as follows

$$R(W) = \sum_{j=1}^{d} [\frac{\bar{\mathbf{w}}_j^2}{\sigma_1^2} + \frac{1}{\sigma_2^2} \sum_{i=1}^{M} (w_j^i - \bar{\mathbf{w}}_j)^2] \tag{7}$$

and this is the only part of $G(W)$ that depends on $\bar{\mathbf{w}}$. Thus, given that we want to minimize this function, we can get an expression for $\bar{w}_j$ depending on W by

$$\bar{w}_j = \arg\min_w (\frac{w^2}{\sigma_1^2} + \frac{1}{\sigma_2^2} \sum_{i=1}^{M} (w_j^i - w)^2) \tag{8}$$

that yields

$$\bar{\mathbf{w}}_j(W) = \frac{\sigma_1^2 \sum_{i=1}^{M} w_j^i}{\sigma_2^2 + M\sigma_1^2} \tag{9}$$

and consequently

$$\frac{\partial \bar{\mathbf{w}}_j(W)}{\partial w_k^{(s)}} = \begin{cases} \frac{\sigma_1^2}{\sigma_2^2 + M\sigma_1^2} & \text{if } j = k \\ 0 & \text{if } j \neq k \end{cases}$$

Moreover,

$$\frac{\partial R(W)}{\partial w_k^{(s)}} = \frac{2\bar{w}_k}{\sigma_1^2} \frac{\partial \bar{\mathbf{w}}_k}{\partial w_k^{(s)}} + \frac{2}{\sigma_2^2} \sum_{i=1}^{M} [(w_k^{(i)} - \bar{\mathbf{w}}_k) \frac{\partial \bar{\mathbf{w}}_k}{\partial w_k^{(s)}})] \tag{10}$$

and substituting by the functions in equations 9 and the corresponding derivatives we obtain the final expression for the partial derivatives of $R(W)$.

2.2 Extension to the Multinomial Logistic Regression Model

Multinomial Logistic Regression model is a statistical model suitable for probabilistic multi-class classification problems. Formally, given M classes $C_1, ..., C_M$, any element x in the input space $\mathbb{R}^d$ is categorized according to the criterion

$$class(x) = \arg \max_{C_i, i=1..M} \frac{P(x \in C_i)}{\sum_{k=1}^{M} P(x \in C_k)} \tag{11}$$

where

$$P(x \in C_i) = \frac{1}{1 + exp(-\mathbf{w}^{(i)} x^T)} \tag{12}$$

and each $\mathbf{w}^{(i)}$ is the parameters vector corresponding to the ith-class, that is the ith-column of the parameters matrix

$$W = \begin{pmatrix} w_1^{(1)} & \dots & w_1^{(M)} \\ \vdots & \vdots & \vdots \\ w_d^{(1)} & \dots & w_d^{(M)} \end{pmatrix}$$

Assuming that we have a training set of samples $D = \{(x_n, y_n)\}_{n=1,..,N}$, where each $x_n \in \mathbb{R}^d$ and $y_n \in \{C_1, ..., C_M\}$, we can consider the loss function described above (see 5) to fix the parameters supposing that $L(W, D)$ is now the negated log-likelihood estimator for this new situation, according to equation 11 and 12.

3 Experiments

To test the presented model for both multi-task and multi-class problems we have performed different experiments. For the multi-task case we have learned different face verification tasks and have used images from the public ARFace Database [10]. For the multi-class case, we have performed classification experiments in two databases from the UCI Machine Learning Repository [11].

3.1 Multi-task Experiments

To test the algorithm in the case of multiple binary related tasks we have performed a set of face verification experiments using the public database AR Face (*http://rvl.www.ecn.purdue.edu/RVL/*). Here we consider that a verification task is a binary problem consisting on decide whether a new unseen face image belongs to the learned subject or not.

The AR Face database contains 26 frontal face images from 126 different subjects. The data set has from each person 1 sample of neutral frontal images, 3 samples with strong changes in the illumination, 2 samples with occlusions (scarf and glasses), 4 images combining occlusions and illumination changes, and 3 samples with gesture effects. Images where taken in two separately periods of time (two samples from each type). Some examples of images in the AR Face database are shown in figure 1.

We have performed the experiments considering from 2 to 10 verification problems. In this experiments we have used 2 positive samples and 4 negative samples to train the system, and the test set includes 20 positive images and 40 negatives. We have performed 10 experiments for each case, and both train and test samples have been randomly selected. The parameters of the method that we have used in multi-task case are $\sigma_1 = 2$ and $\sigma_2 = 6$. In single task case we used $\sigma = 2$.

Table 1 includes the mean error obtained in each case and the corresponding confidence intervals.

Fig. 1. Some samples of images in the AR Face database

Table 1. Obtained error and 95% confidence intervals for the logistic regression method trained separately (first row) and for our shared logistic approach (second row). When more than 4 verification tasks are simultaneously trained, the error rates of the shared approach become lower. No mean error is shown in the case of multi-task logistic regression when only one task is considered.

	1	2	3	4	5
Logistic	32.9 ± 8.2	34.5 ± 6.4	30.5 ± 5.3	31.8 ± 4.2	30.2 ± 3.9
Multi-task Logistic	-	41.6 ± 4.2	35.8 ± 5.4	32.1 ± 5.2	28.7 ± 5.2
	6	7	8	9	10
Logistic	31.8 ± 3.6	31.4 ± 3.3	29.6 ± 3.3	30.2 ± 3.0	29.6 ± 2.9
Multi-task Logistic	27.2 ± 4.2	23.6 ± 3.1	21.8 ± 2.8	17.5 ± 2.4	15.4 ± 2.3

3.2 Multi-class Classification Experiments

We have used Balance and Iris databases from the UCI Machine Learning Repository to perform multi-class classification experiments. In table 2 are detailed the characteristics of these databases.

The parameters of the method have been adjusted by cross validation. The values were $\sigma_1 = 2$ and $\sigma_2 = 6$ for the multi-task case, and $\sigma = 2$ for the single task training.

Table 2. Balance and Iris databases details

Database	*Number of elements*	*Number of features*	*Number of classes*
Balance	625	4	3
Iris	150	4	3

Given that multi-task learning frameworks are specially appropriated when there are few elements in the training set, we have used 10% of the data in

Table 3. Error and confidence interval in the classification experiments using Balance and Iris databases using single-task and multi-task training processes

Database	*Single-Task*	*Multi-Task*
Balance	$36.76\% \pm 1.48\%$	$31.48\% \pm 1.29\%$
Iris	$14.45\% \pm 1.85\%$	$7.04\% \pm 0.65\%$

the training step and 90% in the test step. We have performed 10 10-fold cross validation experiments and the results are detailed in table 3.

3.3 Discussion

In the multi-task learning experiments we observe a considerable improvement of the accuracy when using the proposed multi-task logistic regression approach. On the one hand, when the single task model is used, the accuracy does not variate when we consider more tasks. However, when we use the proposed MTL approach we can observe that the accuracy increases when we consider more tasks, and this improvement is specially significant when more than 7 tasks are considered, where we do not have overlapping between the obtained results with the corresponding confidence intervals in both cases. To justify this evolution of the results, it should be taken into account that with the presented model the method can detect in a more general way features that are relevant for any subject verification task. In these experiments, the task relatedness is clear: the features that can be relevant to determine whether a face belongs to a given subject or not can be as well interesting to verify another subject.

In the multi-class learning experiments performed with Balance and Iris databases from the UCI data sets, there is also a significant improvement of the results when we use the MTL approach, although the statistical relationship of the features among the different classes is not as clear as in the face verification case.

4 Conclusion

In this paper we propose a multi-task learning approach based on sharing knowledge from the parameter space of the probabilistic model. The contribution of the information sharing among the related classification tasks is specially noticeable when only a few samples per class are available.

The experiments performed using two data sets from the UCI database, and a face classification problem using the AR Face data base suggest that the multi-task approach fares better than a single task learning of the same tasks using the same probabilistic logistic regression model. Notice that the MTL restrictions that the model assumes are strong, for this reason it can not be appropriated in general data sets. However, there are cases where these restrictions do hold and in these cases the improvement of our MTL approach is notably. Therefore we plan as a future work to develop a less restrictive version of this MTL modelling.

The probabilistic model presented in this paper suggests new lines of future research. In this formulation, we impose the knowledge sharing property by constraining the parameter space of the classifiers along the multiple tasks. However, more complex approaches based on hidden distributions on the parameters space can be considered.

Moreover, in MTL topic there are still open lines of research, for example to define formally the task relatedness concept. In our model, we impose statistical

priors on the task distribution, assuming certain feature information share among the tasks. Given that this assumption is quite restrictive, the method will be appropriated only when the data distribution is agree with this considerations.

Acknowledgments. This work is supported by MEC grant TIN2006-15308-C02-01, Ministerio de Ciencia y Tecnologia, Spain.

References

1. Bellman, R.: Adaptive Control Process: A Guided Tour. Princeton University Press, New Jersey (1961)
2. Caruana, R.: Multitask learning. Machine Learning 28(1), 41–75 (1997)
3. Thrun, S., Pratt, L.: Learning to Learn. Kluwer Academic Publishers, Dordrecht (1997)
4. Baxter, J.: A model of inductive bias learning. Journal of Machine Learning Research 12, 149–198 (2000)
5. Intrator, N., Edelman, S.: Making a low-dimensional representation suitable for diverse tasks. Connection Science 8, 205–224 (1997)
6. Zhang, J., Ghahramani, Z., Yang, Y.: Learning multiple related tasks using latent independent component analysis. In: Weiss, Y., Schölkopf, B., Platt, J. (eds.) Advances in Neural Information Processing Systems 18, MIT Press, Cambridge, MA (2006)
7. Evgeniou, T., Micchelli, C., Pontil, M.: Learning multiple tasks with kernel methods. Journal of Machine Learning Research 6, 615–637 (2005)
8. Torralba, A., Murphy, K., Freeman, W.: Sharing features: efficient boosting procedures for multiclass object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2004)
9. Madigan, D., Genkin, A., Lewis, D.D., Fradkin, D.: Bayesian multinomial logistic regression for author identification
10. Martinez, A., Benavente, R.: The AR Face database. Technical Report 24, Computer Vision Center (1998)
11. Blake, C., Merz, C.: UCI repository of machine learning databases (1998)