

# Solving Particularization with Supervised Clustering Competition Scheme

Oriol Pujol and Petia Radeva

Computer Vision Center, Campus UAB, 08193 (Bellaterra) Barcelona, Spain  
oriol@cvc.uab.es

**Abstract.** The process of mixing labelled and unlabelled data is being recently studied in semi-supervision techniques. However, this is not the only scenario in which mixture of labelled and unlabelled data can be done. In this paper we propose a new problem we have called particularization and a way to solve it. We also propose a new technique for mixing labelled and unlabelled data. This technique relies in the combination of supervised and unsupervised processes competing for the classification of each data point. Encouraging results on improving the classification outcome are obtained on MNIST database.

## 1 Introduction

The process of mixing labelled with unlabelled data to achieve classification improvement is being recently addressed by the community in the form of semi-supervision processes. This is a general denomination for a recently novel problem of pattern recognition based on the improvement of classification performance in the presence of very few labelled data. This line of work become very active since several authors point out the beneficial effects that unlabelled data can have [4–7].

However this is not the only scenario in which we can apply this mixture. Consider the following examples: Imagine a problem of handwritten character recognition, in which we know that all the characters in the document are written by the same author; consider an architect drawing a technical plane with its own symbols; a medical application in which we know that the data we need to classify proceeds from a single patient; the problem of face recognition, with a significative data set representing what a face is, and a huge set of data representing what a face is not. In all these problems there is a common fact, the knowledge of the fact that our test data set is a *particular subset* of the general training data set.

On the other hand, this problem is widely recognized in other domains, such as human learning theory [8] or *natural language processing* [9]. In those domains, research is done in the line of finding the context of the application given a wide training set. In particular, a general language training corpus has to be changed for domain-specific tasks. This task is called *adaptation*.

On the image domain, the work of Kumar et al. [10] use an EM based refinement of a generative model to infer the particularities of the new data, in what they call *specification*.

Here, the *particularization* problem refers to all those problems in which we can assume the test set to be intrinsically highly correlated and that it is not represented by the overall training set. A common descriptor in the examples described earlier is the fact that the intra-variability of the particular data is smaller than the inter-variability of the overall training set. This *a priori* knowledge is the basic premise for the problem we propose and aim to solve.

In order to exploit this knowledge we will use a mixed approach of supervised and unsupervised processes. The supervised process takes into account the general decision rule (the inter-variability of the complete training set), while the unsupervised process tries to uncover structure of the data (the intra-variability of the test set). The strategy used will be to express both processes using the same framework. The supervised rule will be used to deform the classification space, while the unsupervised process gathers data together in the deformed space. This process is called **Supervised Clustering Competition Scheme**, *SCCS* from now on.

## 2 Supervised Clustering Competition Scheme

### 2.1 General Framework and Main Goal

We aim to find an integrated framework for supervised and unsupervised classification processes so that both can compete for the classification of a data point. In this sense, the clustering methods will lead the data based on the minimization of a dissimilarity measure or maximization of a similarity measure, and the supervised classification process has to guide the data according to the classification borders. Since both techniques has to compete, we cast both processes in the same minimization framework. Though, the clustering scheme can be casted into it straightforwardly, the same does not happen to the supervised process. This last process must be reformulated. Therefore, in order to blend both processes we use the following equation,

$$L(\mathbf{x}) = h(\min_{\mathbf{x}}(\alpha \cdot SF(\mathbf{x}) + (1 - \alpha) \cdot UF(\mathbf{x}))) \quad (1)$$

where  $SF$  stands for *supervised functional*, a functional the minimums of which are close to the centers of each class and that has an explicit maximum on the border between classes;  $UF$  stands for *unsupervised functional*, expressing a dissimilarity measure;  $\alpha$  is the mixing parameter; and the function  $h(\cdot)$  is a decision scheme that allows the classification of each data sample.

To minimize this process we can use gradient descent,

$$\frac{\partial \mathbf{x}}{\partial t} = -\nabla F(\mathbf{x})$$

The iterative scheme is,

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \Delta t \cdot \nabla F(\mathbf{x})$$

where  $\nabla F(\mathbf{x}) = \{\partial F(\mathbf{x})/\partial x_i\}$ . Therefore,

$$\frac{\partial \mathbf{x}}{\partial t} = -\alpha \frac{\nabla SF(\mathbf{x})}{\|\nabla SF(\mathbf{x})\|} - (1 - \alpha) \frac{\nabla UF(\mathbf{x})}{\|\nabla UF(\mathbf{x})\|} \quad (2)$$

The next step is to define the minimization process that represent the supervised classification and the clustering process.

**Similarity Clustering.** When considering the unsupervised functional, we are interested in tools for robust clustering related to invariability of the initial state, capability to differentiate among different volumes of different sizes and toleration to noise and outliers [1–3].

In the approach we have chosen [1] a *similarity* measure between two points  $S(z_j, x_i)$  is used in a maximization framework. Our goal is to maximize the total similarity measure  $J_s(\mathbf{x})$  defined as:

$$J_s(\mathbf{x}) = \sum_{i=1}^c \sum_{j=1}^n f(S(z_j, x_i))$$

where  $f(\cdot)$  is a monotone increasing function,  $\mathbf{x}$  represents the centers of each cluster and  $\mathbf{z}$  is the original data set (where  $\mathbf{z} = \{z_1, \dots, z_n\}$  and  $z_i$  is a  $D$ -dimensional data point). As a similarity relation  $S(z_j, x_i)$ , we use,

$$S(z_j, x_i) = e^{-\left(\frac{\|z_j - x_i\|^2}{\beta}\right)}$$

where  $\beta$  is a normalization term. Let the monotone increasing function  $f(\cdot)$  be,

$$f(\cdot) = (\cdot)^\gamma \quad , \quad \gamma > 0$$

Therefore, the complete similarity measure  $J_s(\mathbf{x})$  is,

$$J_s(\mathbf{x}) = \sum_{i=1}^c \sum_{j=1}^n \left( e^{-\frac{\|z_j - x_i\|^2}{\beta}} \right)^\gamma \quad (3)$$

The parameter  $\beta$  is superfluous in this scheme and can be defined as the sample variance,

$$\beta = \frac{\sum_{j=1}^n \|z_j - \bar{z}\|^2}{n} \quad \text{where} \quad \bar{z} = \frac{\sum_{j=1}^n z_j}{n}$$

The parameter  $\gamma$  gains a considerable importance in this scheme since a good  $\gamma$  estimate induces a good clustering result. The process of maximizing the total similarity measure is a way to find the peaks of the objective function  $J_s(\mathbf{x})$ . It is shown in [1] that the parameter  $\gamma$  is used as a neighboring limiter, as well as a local density approximation. To find  $\gamma$  one can use an exhaustive search of the correlation of the similarity function for each point when changing the parameter. If the correlation value is over a certain threshold, we can consider that the similarity measure represents the different variability and volumes of

the data set accurately. The authors set this threshold experimentally to 0.97 but can change according to the application.

The similarity clustering approach uses the same similarity function but, as it is a self-organizing approach, we define the initial data and centers by the unlabelled data points  $\mathbf{z}^0 = \mathbf{x}^0$ ,

$$UF(\mathbf{x}) = J_s(\mathbf{x}) = \sum_{j=1}^n \left( e^{-\frac{\|z_j - x_k\|^2}{\beta}} \right)^\gamma, \quad k = 1 \dots n$$

Getting its gradient, we obtain,

$$\nabla UF(\mathbf{x}) = -2\frac{\gamma}{\beta} \sum_{j=1}^n \left( e^{-\frac{\|z_j - x_k\|^2}{\beta}} \right)^\gamma (z_j - x_k), \quad k = 1 \dots n \quad (4)$$

## 2.2 Supervised Classifier Functional

The definition of the supervised classifier functional should be made so that both processes can interact with each other. Therefore, we must reformulate the supervised classifier process as a self-organizing iterative process.

Without loss of generality, we restrict our classifier design to a two class supervised classification process using a Bayesian framework with known class probability density functions. Assuming that we can estimate each class probability density function  $f_A(\mathbf{x}|c = A)$  and  $f_B(\mathbf{x}|c = B)$ , the optimal discriminative rule using a Bayesian approach is given by,

$$h(\mathbf{x}) = \begin{cases} A & f(\mathbf{x}|c = A)P(A) > f(\mathbf{x}|c = B)P(B) \\ B & f(\mathbf{x}|c = A)P(A) \leq f(\mathbf{x}|c = B)P(B) \end{cases}$$

If *a priori* knowledge of the probability appearance is not known, we assume  $P(A) = P(B)$ .

The manifold we are looking for, must maintain the optimal borderline as a maximum, since we want the minimums to represent each of the classes. It can be easily seen that the following family of functions satisfies the requirement,

$$SF = -(f(\mathbf{x}|c = A) - f(\mathbf{x}|c = B))^{2N}, \quad \forall N \in \{1, \dots, \infty\} \quad (5)$$

In order to obtain a feasible closed form of the functional we can restrict the density estimation process to a Gaussian mixture model,

$$SF(\mathbf{x}) = \mathbf{f}_K(\mathbf{x}) = \sum_{i=1}^{M_k} \pi_i \mathbf{g}_i(\mathbf{x}, \theta_i)$$

where  $M_k$  is the model order and  $g_i(\mathbf{x}, \theta_i)$  is the multidimensional gaussian function,

$$g_i(\mathbf{x}, \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2} |\Sigma_i|} e^{-\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i)}$$

where  $\theta_i = \{\mu_i, \Sigma_i\}$  are the parameters for the gaussian, the covariance matrix and the mean, and  $d$  is the dimensionality of the data.

### 2.3 The Procedure

Summarizing the overall procedure, we begin the process with three data sets, the labelled data, the unlabelled data and the test set.

Labelled data is used to create the model for the supervised functional as well as used for the final decision step. Unlabelled data feeds the competition scheme and it is the data that will be labelled according to the balance of the supervised and unsupervised processes. The final algorithm is as follows:

- Step 1:** Determine a supervised set  $L$  of data among labelled data and a set  $U$  of unlabelled data.
- Step 2:** Estimate  $SF(\mathbf{x})$   $x \in L$  from (5).
- Step 3:** Apply the CCA<sup>a</sup> step to set the parameter  $\gamma$  of  $UF(x)$   $x \in U$ .
- Step 4:** Feed the competition scheme by evolving the clusters (that come from unlabelled data  $U$ ) according to (2) and let them converge.

---

<sup>a</sup> Please refer to [1] for further information on this process.

## 3 Experimental Results

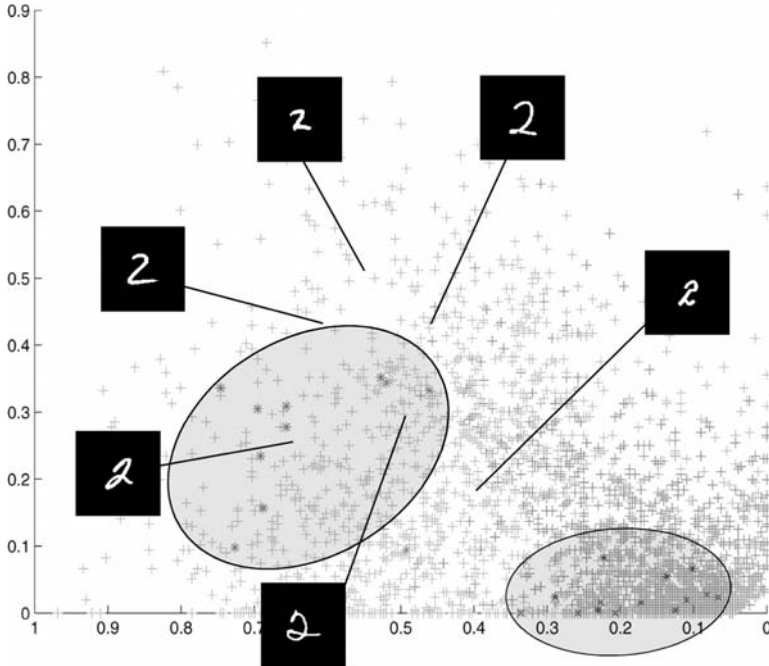
To illustrate the behavior and the advantages of the Supervised Clustering Competition Scheme in a real particular problem, we have performed our tests on data from the MNIST handwritten characters database.

### 3.1 Experiment Settings

In order to exemplify the behavior of the methodology we have chosen a full supervised classification scenario in which an OCR is used to distinguish between two hand written digits. In order to ensure structure in the data, we aim for the classification of data of a single writer at a time, in the same way as if the OCR was scanning and recognizing the digits in a single sheet of paper.

We have used the MNIST digits database with images of  $128 \times 128$  pixels. The feature extraction process has been a  $4 \times 4$  zoning procedure on the  $64 \times 64$  central values of the image. The number of white pixels is counted and stored as a feature. Therefore we have a 16 dimensional feature space. The training set is composed by 100 different hand-written sets of 120 digits each one (The first 100 writers of the HSF7). Each set corresponds to a different writer. The test set is a 500 different hand-written sets of 120 digits each one (The first 500 writers of the HSF6). We center the experiment in a two-class classification process distinguishing number ONE from number TWO. From the set of features we have selected automatically two of the most discriminative ones using discriminant analysis.

Figure 1 shows the two class classification problem. The figure represents the feature space associated to the training set of the digits “two” and “one”. As one can see, the training set is general and contains different instances of the digit “two” with different particularities. However, not always a general purpose

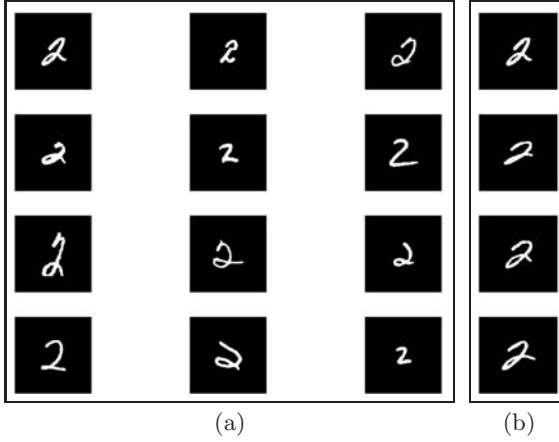


**Fig. 1.** Feature space representation of the training set. The asterisks represents the test set. The shadow areas represents a possible partition of the training set that solves the two class problem.

classifier is needed. If our particular instance of the problem is a subset of the general set, we can focus on extracting some information of the test data, so we can constrain the training space. The asterisks are the test data for the digits “one” and “two” written by the same writer. The shadowed areas represent possible subsets of the training data which characterize our test data. We can see that this subproblem has much more structure than the original one and that the intra-variability of the particular set is lower than the variability for all writers. In this scenario we can use the SCCS.

Figure 2 shows a representative set of hand-written digits. Figure 2.a shows the variability of the digit “two” written by different writers. Figure 2.b shows the variability of the digit “two” written by just one author.

**Experiment Results.** As a ground-truth unsupervised classifier we have used Expectation-Maximization with K-Means initialization, and a supervised labelling for the cluster centers. This process yields a recognition rate of 89.35%. The supervised classifier related to the whole training set achieves a recognition rate of 87.65%. This supervised classifier is the same we are mimicking with the supervised functional, so that we can compare performances of adding the unsupervised process.



**Fig. 2.** (a) Set of digit “two” written by different authors. (b) Set of digit “two” written by the same writer.

**Table 1.** Comparative table for fixed  $\alpha$  values.

Mixing value	Recognition Rate	Overall Gain
0	87.73%	0.08%
0.05	87.16%	-0.49%
0.1	87.49%	-0.16%
0.15	87.89%	0.24%
0.2	87.73%	0.08%
0.25	88.14%	0.49%
0.3	89.67%	2.02%
0.35	90.65%	3.00%
0.4	93.07%	5.42%
0.45	92.59%	4.94%
0.5	87.65%	0.00%

Table 1 shows the behavior of the process for different fixed values of the parameter  $\alpha$ . The parameter  $\gamma$  is set to adapt automatically. The second column refers to the recognition rate of the overall process. It can be clearly seen that a mixture of supervised and unsupervised improves the performance of the supervised classifier. We must take into account that when  $\alpha \geq 0.5$  the process behaves as a pure supervised classifier, thus, we only show the results for  $\alpha$  between 0 and 0.5. The third column (Overall Gain) represents the percentage of gain or degradation of the process as the difference between the error of the process and the error achieved using the supervised classifier. Positive values are gain in performance. Negative values are degradation of the process performance. Hence, for  $\alpha = 0.1$  the process performs worse than the supervised classifier. This is particularly obvious in the sense that low values of  $\alpha$  mean a nearly no contribution of the supervised process, and therefore, the classification

is made with the unsupervised process alone. As we can see, the discriminative power of the resulting classifier improves by more than a 5% the classification results obtained by the supervised classifier without the unsupervised part.

## 4 Conclusions and Future Lines

In this paper we have introduced a new competition technique for pattern recognition that combines supervised and unsupervised approaches. This technique exploits the structure of the data set to be classified and adds this feature to a classical supervised classification process based on generative models.

The method we propose has been used in an emerging problem of *particularization*. The results using real data of MNIST database with 500 writers show that the SCCS improves the performance of the supervised rule alone by more than 5%. This method has been particularized for a multidimensional two class problem, we are now developing the formulation for the basis of the multi-class supervised functional.

## References

1. Miin-Shen Yang and Kuo-Lung Wu, A Similarity-Based Robust Clustering Method. IEEE Trans. on PAMI, Vol. 26, No. 4, pp. 434-448, 2004.
2. R.N. Dave and R. Krishnapuram. Robust Clustering Methods: A Unified View. IEEE Trans. Fuzzy Systems, vol. 5, pp. 270-293, 1997.
3. P.J. Huber. Robust Statistics. Wiley, 1981.
4. A. McCallum and K. Nigam. Employing EM and pool-based active learning for text classification. Int. Conf. on Machine Learning, pp. 359-367, 1998.
5. T.J. O'Neill. Normal Discrimination with unclassified observations. J. of American Statistical Assoc, vol. 73, pp. 821-826, 1978.
6. F. Gagliardi and M. Cirelo. Semi-Supervised Learning of Mixture Models. Proc. XXth ICML, 2003.
7. A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. Proc. XIth. Annual Conference on Computational Learning Theory, Madison, WI, 1998.
8. W.J. Clancey. A Tutorial on Situated Learning. In Proc. of the International Conference on Computers and Education, pp. 49-70, 1995.
9. A.R. Coden, S.V. Pakhamov and C.G. Chute. Domain-Specific Language Models and Lexicons for Tagging. Tech. Reprt. RC23195 (W0404-146) IBM Research Division, 2004.
10. S. Kumar, A. Loui and M. Herbert. An observation-constrained generative approach for probabilistic classification of image regions. Image and Vision Computing, vol. 21, pp. 87-97, 2003.