Two-stage Classification System combining Model-Based and Discriminative Approaches

Jonathan Milgram, Robert Sabourin and Mohamed Cheriet
Laboratoire d'Imagerie, de Vision et d'Intelligence Artificielle
École de Technologie Supérieure de Montréal
milgram@livia.etsmtl.ca, {robert.sabourin, mohamed.cheriet}@etsmtl.ca

Abstract

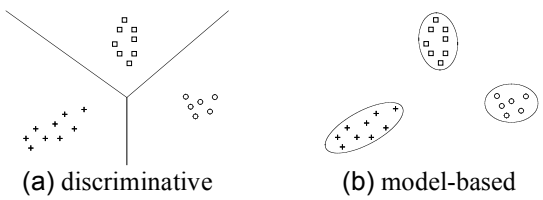
For the tasks of classification, two types of patterns can generate problems: ambiguous patterns and outliers. Furthermore, it is possible to separate classification algorithms into two main categories. Discriminative approaches try to find the better separation among all classes and minimize the first type of error. But, in general they cannot deal with outliers. Besides, model-based approaches make the outlier detection possible but are not sufficiently discriminative. Thus, we propose to combine a model-based approach with support vectors classifiers (SVC) in a two-stage classification system. Another advantage of this combination is to reduce the principal burden of SVC: the processing time necessary to make a decision. Finally, the experiments on handwriting digit recognition have shown that it is possible to maintain the accuracy of SVCs, while decreasing complexity significantly.

1. Introduction

The principal objective of a pattern recognition system is to minimize classification errors. However, another important factor is the capability to estimate a confidence measure in the decision made by the system. Indeed, this type of measure is essential to be able to make no decision when the result of classification is uncertain.

From this point of view, it is necessary to distinguish two categories of problematic patterns. The first one relates to ambiguous data which may cause confusion between several classes and the second category consists of data not belonging to any class: the outliers.

Furthermore, most classification algorithms can be divided into two main categories denoted as discriminative and model-based approaches.



(a) discriminative (b) model-based

Figure 1: Two types of classification approaches

The former tries to split the feature space into several regions by decision surfaces, whereas the latter is based on the development of a model for each class along with a similarity measure between each of these models and the unknown pattern (see Figure 1).

Thus, as is shown in [4], the discriminative classifiers are more accurate in classifying ambiguous data, but not suitable for outlier detection, whereas model-based approaches are able to reject outliers but not effective in classifying ambiguous patterns.

Considering this, the authors propose to hybridize the two types of approaches internally or to combine them externally. In a more recent paper [5], the same authors have tested an internal fusion of the two approaches. Their method improves the accuracy of the model-based approach by using discriminative learning. However, even though their classifier is more accurate, it is not as accurate as the best discriminative approaches such as support vector classifiers.

Hence, in this paper, we propose to combine a model-based approach with support vector classifiers (SVC). This classification system should give high accuracy and strong outlier resistance. The idea is to develop a two-stage classification system. At the first stage, a model-based approach can directly classify patterns that are recognized with high confidence, reject outliers or insulate those classes in conflict. Then, if conflict is detected, the appropriate SVCs will make better decision at the second stage. Another advantage of this combination is to reduce the main burden of SVC: the processing time necessary to make a decision.

Although a number of similar ideas related to two-stage classification were introduced in recent papers [1][6][7], our classification system remains different and original. Indeed, the combination of model-based and discriminative approaches is proposed in [6], but the authors use only a few MLPs to improve the accuracy of the first classifier and they do not take into account outlier rejection. On the other hand, the use of SVCs in a second stage of classification to improve the accuracy is presented in [1][7]. In [1], the authors take into account the problem of complexity of SVCs, but in the first-stage they use MLP which is another discriminative approach. Furthermore, their system does not make decisions at the first-stage and always uses one SVC, and never more than

one, which limits the performance of the system. In [7], the authors propose several elaborate strategies for detecting conflicts. However, they do not take into account the problem of complexity. Indeed, the first-stage uses a complex ensemble of classifiers. Moreover, the results of their two-stage system are not compared to a full SVC system. Thus, if the use of SVCs can improve the accuracy of the ensemble of classifier used in the first stage, would it then be better to use a full SVC system?

2. Model-based approach

This type of approach is based on the development of a model for each class and a measure of the similarity between each of these models and the unknown pattern.

2.1. Characterization of the recognition problem

Although this approach is not very discriminative, it can be used to characterize the problem of pattern recognition. Indeed, as we can see in Figure 2, by the combination of similarity measures d_i shown in (a) and (b) we can detect outliers (c) and conflicts (d).

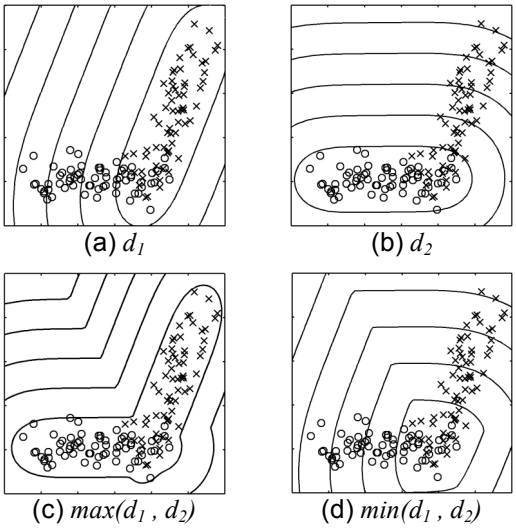


Figure 2 : Use of model-based approach to characterize the pattern recognition problem

Thus, we propose to use a model-based approach as a first-stage of decision making. At this level we can make a decision for unambiguous patterns, reject outliers or detect conflicts.

2.2 Modeling with hyperplanes

We use a simple method to model each class ω_i with a hyperplane defined by the mean vector μ_i, and the k eigenvectors of the covariance matrix Σ_i with the largest eigenvalues.

$$d_i(x) = \|x - P_i(x)\| \quad (1)$$

$$P_i(x) = (x - \mu_i)\Psi_i\Psi_i^T + \mu_i \quad (2)$$

where ψ_i denotes the matrix of the k eigenvectors.

Thus, given a data point x of the feature space, the class membership can be evaluated by the distance d_i from the point x to its projection $P_i(x)$ on the hyperplane. Figure 3 shows an example of projection distance.

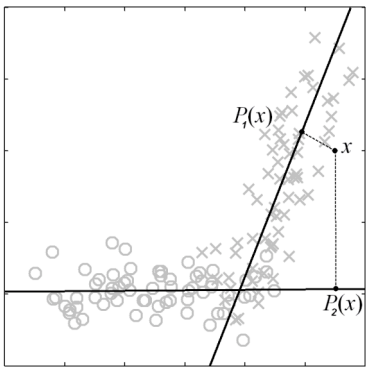


Figure 3: 2D example of projection distance

Furthermore, this method required the optimization of only one parameter: the number k of eigenvectors used. But, as we can see below, this parameter is crucial for classification.

3. Combination with discriminative approach

The second level of decision making of our system aims to process the ambiguous patterns. In this case, the final decision will be made by discriminative experts.

3.1. Conflict detection

The first step is to detect the patterns that may cause confusion. In [1] and [6], the authors consider that conflict involves only two classes and they use pairwise expert to reprocess all samples [1], or just the samples rejected by the first classifier [6]. However, we consider that it is preferable to use a dynamic number of classes in conflict. Thus, to determine the list L_ω of the p classes in conflict, we normalize the projection distances d_i to obtain membership measures s_i related to the class ω_i.

$$s_i = \frac{e^{-\alpha d_i}}{\sum_j e^{-\alpha d_j}} \quad (3)$$

Then, we use the values s_i to rank all classes ω_i in descending order and we determine the minimum number

p of classes necessary to verify the criterion:

$$1 - \sum_{i=1}^{p} s_i < \varepsilon \qquad (4)$$

The smaller the threshold ε is, the larger the number p will tend to be. Thus, this parameter controls the tolerance level of the first stage of classification. If ε is too large, then we never use the second stage of classification. But, if ε is too small, then the system uses unnecessary discriminative classifiers.

3.2. Use of Support Vector Classifiers

The objective of the second level of classification is to reprocess the ambiguous patterns with the discriminative approach in order to make decisions between the p classes which are in conflict. It seems preferable to use a modular approach like pairwise strategy which consists of decomposing a n class problem in $n(n\text{-}1)/2$ binary sub-problems. In this context, it is interesting to use SVCs because they are good discriminative classifiers. Thus, we train the SVCs concerning all pairs of classes. For classification, we use only the $p(p\text{-}1)/2$ classifiers defined by the list L_ω with a voting strategy: each of the $p(p\text{-}1)/2$ SVCs votes for one of the p classes and finally the unknown pattern is classified to the class with the maximum number of votes. Also, concerning equality, we select the class with the smaller d_i.

4. Experimental results

We have applied the proposed approach to a classical pattern recognition problem: handwritten digit recognition.

4.1. Database

In the experiments, we used a well-known benchmark database. The MNIST (Modified NIST) dataset [3] was extracted from the NIST special database SD3 and SD7. The original binary images were normalized into 20 × 20 grey-scale images with aspect ratio preserved and the normalized images were centered by center of mass in 28 × 28 images. The learning dataset contains 60 000 samples and 10 000 others are used for testing. Moreover, we have divided the learning database into two subsets. The first 50 000 samples have been used for training and the next 10 000 for validation.

4.2. Model-based approach

Initially, we must fix the dimensionality of the hyperplane models. For this purpose, we use the validation dataset. The results are plotted in Figure 4. Thus, we can see that the parameter k strongly influences the accuracy of the classification. Consequently, we use $k = 25$ and we obtain an error rate of 4.09 % on the test dataset.

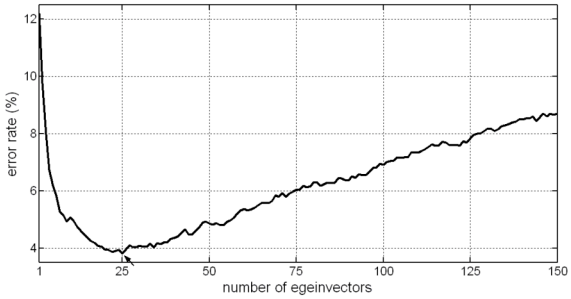


Figure 4: Effect of *k* on classification results

But, as we can see on Table 1, the true label of a test data is not always in the first two classes. Thus, in this case, conflicts might be related to more than two classes.

Table 1: Accuracy of the model-based approach

position of the true label	1	2	3	>3
% of the test dataset	95.91	2.82	0.72	0.55

Finally, even though the reliability of the proposed model-based approach is not very high, it should probably be able to characterize the classification problem.

4.3. Support Vector Classifiers

The training and testing of all SVCs are performed with the LIBSVM Software [2]. We use the RBF kernel $K(x,y) = e^{-\gamma\|x-y\|^2}$. The penalty parameter C and the kernel parameter γ are empirically optimized. We choose parameters that minimize the error rate on the validation dataset. Thus, for subsequent tests we use $C = 10$ and $\gamma = 0.0185$ for which we have obtained an error rate of 1.47 % on validation. This ensemble of SVCs uses 11 118 support vectors and makes it possible to obtain an error rate of 1.54 % on the test dataset. This result is comparable with the best results obtained on this database without use of prior knowledge (see benchmark in [5]). Notice that the number of support vectors used is an important factor, because it is proportional to the classification complexity and thus to the decision time.

4.4. Two-stage classification system

According to the application constraints, it is necessary to make a compromise between accuracy and complexity. The threshold ε of equation (4) controls this compromise, as we can see in Figure 5. Thus, the validation dataset can be used to fix this parameter according to the objective fixed by the application.

On the other hand, the coefficient α of the normalization function (3) is chosen to minimize the mean square error (MSE) on the validation dataset. We obtain the best result with α = 6.0.

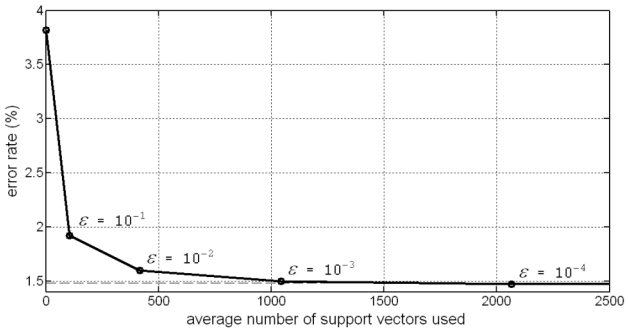


Figure 5: Results on the validation dataset

Considering the final results obtained on the test dataset (see Table 2 and Figure 6), a number of conclusions can be drawn. First, a significant accuracy improvement is achieved by the reprocess from only 10 % of the examples. Indeed, while using a tolerance threshold $\varepsilon = 10^{-1}$, the error rate on the test dataset is 2.03 % *vs.* 4.09 % with only the model-based approach. Second, it is possible to maintain the accuracy of the full pairwise ensemble by using approximately 10 % of the initial complexity. Indeed, while using $\varepsilon = 10^{-3}$, the average number of support vectors used is 1 054.3 *vs.* 11 118 with all SVCs.

Table 2: Final results on the test database

ε threshold	10^{-1}	10^{-2}	10^{-3}	10^{-4}
error rate (%)	2.03	1.62	1.53	1.51
average # SVCs	0.18	0.81	2.38	5.13
average # SVs	108.2	416.0	1054.3	2026.5

Furthermore, because the number of classes in conflict is dynamic, it is interesting to observe the distribution of the number of SVCs used to classify the test dataset (see Figure 6).

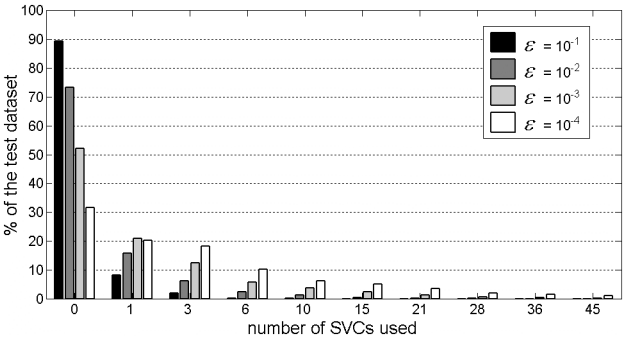


Figure 6: Distribution of number of SVCs used

Thus, we can see that to maintain accuracy of the combination of all SVCs, it is necessary to use more than one SVC to resolve conflict. This fact shows that the first level is not effective enough.

5. Conclusions and perspectives

We have presented a new classification architecture that has several interesting properties for application to pattern recognition. It combines the advantages of a model-based classifier, in particular modularity and efficient rejection of outliers, with the high accuracy of SVC. Moreover, it greatly reduces the decision time related to the SVC, which is very important in the majority of real pattern recognition systems.

In future works, the rejection capabilities of our two-stage classification system have to be implemented. The model-based approach can be used to detect outliers. On the other hand, to reject ambiguous data, the voting scheme of SVCs might be replaced by probability estimation.

Finally, our approach is limited by the accuracy of the first decision stage. Indeed, the model-based approach used is not accurate. Thus, a significant improvement would be to use at the first-stage more than one single hyperplane per class.

References

[1] Bellili, A., Gilloux, M., Gallinari, P. (2003) An MLP-SVM combination architecture for offline handwritten digit recognition, *International Journal on Document Analysis and Recognition*, 244-252.

[2] Chang, C.-C., Lin, C.-J., (2001) LIBSVM : a library for support vector machines. Software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[3] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. (1998) Gradient-based learning applied to document recognition. *Proceedings of IEEE,* 86(11), 2278-2324. MNIST database available at: http://yann.lecun.com/exdb/mnist/.

[4] Liu, C.-L., Sako, H. & Fujisawa, H. (2002) Performance evaluation of pattern classifiers for handwritten character recognition, *International Journal on Document Analysis and Recognition*, 191-204.

[5] Liu, C.-L., Sako, H. & Fujisawa, H. (2003) Handwritten digit recognition: benchmarking of state-of-the-art techniques, *Pattern Recognition*, 2271-2285.

[6] Prevost, L., Michel-Sendis, C., Moises, A., Oudot, L., Milgram, M. (2003) Combining model-based and discriminative classifiers: application to handwritten character recognition, *International Conference on Document Analysis and Recognition,* 31-35.

[7] Vuurpijl, L., Schomaker, L., Van Erp, M. (2003) Architectures for detecting and solving conflicts: two-stage classification and support vector classifiers, *International Journal on Document Analysis and Recognition,* 213-223.