ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Selecting from an infinite set of features in SVM

Rémi Flamary, Florian Yger and Alain Rakotomamonjy*

LITIS EA 4108 - Université de Rouen,
76800 Saint Etienne du Rouvray - France

Abstract. Dealing with the continuous parameters of a feature extraction method has always been a difficult task that is usually solved by cross-validation. In this paper, we propose an active set algorithm for selecting automatically these parameters in a SVM classification context. Our experiments on texture recognition and BCI signal classification show that optimizing the feature parameters in a continuous space while learning the decision function yields to better performances than using fixed parameters obtained from a grid sampling.

1 Introduction

The choice or the design of a kernel plays a primary role on the performance of a kernel machine such as an SVM or a KFDA. In order to overcome the difficulty that such a choice may bring, several recent works have investigated approaches for combining kernels. Multiple kernel learning (MKL) helps in designing kernels functions by jointly learning a decision function and a combination of bases kernel [1]. Another interesting point of MKL is that, when considered kernels have parameters, it can be used for model selection by including in the combination several instances of the same kernel but with different parameter values. This framework of MKL has been recently extended so as to handle an infinite number of kernels or kernels with continuous parameters [2].

In this work, we consider a specific case of the kernel method framework where feature maps are defined explicitly instead of being defined implicitly through the kernel function. There are many practical situations in which such a situation occurs. For instance, in face recognition problems, Gabor based features are extracted using filter convolutions [3]. The problem we address is thus the problem of automated selection of feature maps in a supervised learning framework using Support Vector Machines. We consider a situation where the feature maps involve some parameters (which can be continuous). In such a situation, the common approach is either to optimize the parameters by cross-validation (if their number is relatively small) or to fix the parameter values beforehand. The latter strategy is for instance very common when extracting Gabor features where scale and frequency are sampled from predefined intervals.

We deal with this problem of feature map parameter selection by fitting the problem into the framework of feature selection through sparsity-inducing norm. The main particularity of our problem compared to classical ones [4] is that the

*This work is funded in part by the FP7-ICT Programme of the European Community, under the PASCAL2 Network of Excellence, ICT-216886 and by the French ANR Project ASAP ANR-09-EMER-001 and by the INRIA ARC MaBi project.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

number of features we have to deal with can be potentially infinite since the feature parameters are continuous. In some ways, we follow recent works that learn from an infinite set of kernels [2]. However, by considering feature maps, we provide an optimization framework that consider the problem in its primal version from which we derive a simple and efficient active constraint algorithm. Furthermore, since the problem is still linear in the feature maps, we are able to deal with a large number of training examples. We show in our experiments on texture recognition and BCI electro-encephalogram signal classification that considering features with continuous parameters can yield to better classification performances than feature with sampled parameters.

2 Selecting from an infinite set of features

We detail in this section the algorithm we develop for dealing with an infinite set of features as well as examples of such infinite sets.

2.1 The algorithm

Let us formally define the framework of our problem. Consider a set of n training examples $\{\mathbf{x}_i, y_i\}_{i=1}^n$ with the data $\mathbf{x}_i$ belonging to some space $\mathcal{X}$ and the labels $y_i \in \{-1, 1\}$. For instance, in a context of signal or image classification, $\mathcal{X}$ would respectively be $\mathbb{R}^d$ or $\mathbb{R}^{d\times d}$. We define a θ-parametrized function $\phi_\theta(\cdot)$ that maps an element of $\mathcal{X}$ into $\mathcal{X}_\theta$, this function being explicitly known. In this context, we are looking for a decision function of the form

$$f(\mathbf{x}) = \sum_{j=1}^{N} \langle \mathbf{w}_j, \phi_{\theta_j}(\mathbf{x}) \rangle_{\mathcal{X}_{\theta_j}} \tag{1}$$

that is able to predict most accurately as possible the label of a novel example $\mathbf{x}$. Note that the decision function considers only a finite number N of feature maps each of which has a parameter value θ_j. According to Bach et al. [5], this decision function can be understood as a feature map version of a MKL one. For performing feature selection, we want some of the $\mathbf{w}_i$ to vanish. This can be enforced by a sparsity inducing norm while learning the weight vectors $\mathbf{w}_i$ through the minimization of a regularized empirical risk.

Before delving into the details on how we deal with an infinite set of features, we first describe an efficient active set approach for addressing the finite number of features case. Let us define $\mathbf{w}$ the vector of stacked $\mathbf{w}_j$, $\mathbf{\Phi}_{\theta_j}$ the matrix which rows i are $\phi_{\theta_j}(\mathbf{x}_i)$ and $\mathbf{\Phi}$ the matrix of feature maps, resulting from the concatenation of the N matrices $\{\mathbf{\Phi}_{\theta_j}\}$. Each column of $\mathbf{\Phi}$ is normalized to unit norm and $\widetilde{\mathbf{\Phi}} = \mathrm{diag}(\mathbf{y})\mathbf{\Phi}$, with $\mathbf{y}$ being the vector of labels $\{y_i\}$. We define $\mathbf{w}$ as the solution of the following learning problem where the loss function is a square hinge loss

$$\min_{\mathbf{w},b} \quad J(\mathbf{w}) = \frac{C}{2n}(\mathbb{1} - \widetilde{\mathbf{\Phi}}\mathbf{w})_+^T(\mathbb{1} - \widetilde{\mathbf{\Phi}}\mathbf{w})_+ + \Omega(\mathbf{w}) \tag{2}$$

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

Algorithm 1 Active set method algorithm for SVM finite/infinite feature selection

Set $\mathcal{A} = \emptyset$ initial active set
Set $\mathbf{w} = \vec{0}$
repeat
  $\mathbf{w} \leftarrow$ solve problem (2) with features from $\mathcal{A}$
  $\theta, i \leftarrow \max_{i \in \mathcal{A}^c} \quad ||\mathbf{r}_i||_2$
  if $\theta > 1$ **then**
    $\mathcal{A} = \mathcal{A} \cup i$
  end if
until $\theta \leq 1$

where $[\widetilde{\mathbf{\Phi}}\mathbf{w}]_i = f(\mathbf{x}_i)$, $\mathbb{1}$ is a unitary vector, $(\cdot)_+ = \max(0, .)$ is the element-wise positive part of a vector, Ω is a regularization term that induces sparsity and C is a trade-off parameter that balances training error and regularization. Typical sparsity inducing norms are the ℓ_1 norm defined as $\Omega_1(\mathbf{w}) = \|\mathbf{w}\|_1$ and the mixed ℓ_{1-2} norm $\Omega_{1,2}(\mathbf{w}) = \sum_i \|\mathbf{w}_i\|_2$ which induces sparsity on groups of features. The optimality condition of problem (2) is $-\frac{C}{n}\widetilde{\mathbf{\Phi}}^T(\mathbb{1} - \widetilde{\mathbf{\Phi}}\mathbf{w})_+ + \mathbf{c} = \vec{0}$ where $\mathbf{c}$ is a subgradient of the norm Ω. We focus in the sequel on the the mixed-norm $\Omega_{1,2}(\mathbf{w})$. By definition of the subgradient of the $\|\mathbf{w}_i\|_2$ norm and because $\Omega_{1,2}(\mathbf{w})$ is group-separable, it can be shown that the optimality conditions are :

$$\begin{array}{rcl} -\mathbf{r}_i + \frac{\mathbf{w}_i}{||\mathbf{w}_i||_2} = \vec{0} & \forall i & \mathbf{w}_i \neq \vec{0} \\ ||\mathbf{r}_i||_2 \leq 1 & \forall i & \mathbf{w}_i = \vec{0} \end{array} \tag{3}$$

with $\mathbf{r}_i = \frac{C}{n}\widetilde{\mathbf{\Phi}}_i^T(\mathbb{1} - \widetilde{\mathbf{\Phi}}\mathbf{w})_+$. These conditions suggest an active-set method for solving the unconstrained optimization problem. Indeed, since the problem is supposed to have a sparse solution, many $\mathbf{w}_i$ are expected to be 0 at optimality. Then, in a block-coordinate descent approach, it seems reasonable to optimize on variables that violate their constraints while keeping the other fixed and repeat these procedures as long as some $\mathbf{w}_i$ violate their optimality constraints. Hence the algorithm runs as follows : at some given iteration, we have a subset $\mathcal{A}$ of features defined as the active ones and optimize the problem only over $\mathbf{w}_\mathcal{A}$. At this point, depending on the value of $\mathbf{r}_i$, some of the $\mathbf{w}_i$ that do belong to the complementary set $\mathcal{A}^c$ may not satisfy their own optimality condition. Such a violating-constraint feature is added to the active set and optimization over $\mathcal{A}$ is repeated again until all $\mathbf{w}_i$ satisfy their optimality conditions. Note that if several $\mathbf{w}_i$ violate their constraints, the choice of the one to integrate to the active set impacts on the speed of convergence of the algorithm. While adding any violating feature or group of features into the active set leads to decrease the objective value, it is more appropriate to add the most violating one. This consists in solving the problem $\max_{i \in \mathcal{A}^c} ||\mathbf{r}_i||_2$. The overall procedure is described in Algorithm 1.

When dealing with the infinite set ϕ_θ of features, θ being a continuous parameter, we cast the problem as $\min_{|\Theta| < \infty} \min_{\mathbf{w}} J(\mathbf{w})$ where $\Theta = \{\theta_j\}$ is the set

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

of feature map parameter values involved in $\mathbf{\Phi}$ and $|\Theta|$ is the cardinality of this set. This optimization problem searches for a finite number of features $\{\phi_{\theta_j}\}$ which achieves the lowest objective value of $J(\mathbf{w})$. The optimality conditions of (3) can be readily extended to this infinite case as

$$\begin{array}{l} -\mathbf{r}_i + \frac{\mathbf{w}_i}{||\mathbf{w}_i||_2} = \vec{0} \quad \forall i \ \mathbf{w}_i \neq \vec{0} \quad , \quad ||\mathbf{r}_i||_2 \leq 1 \quad \forall i \ \mathbf{w}_i = \vec{0} \\ \|\widetilde{\mathbf{\Phi}}_{\theta_s}^T(\mathbb{I} - \widetilde{\mathbf{\Phi}}\mathbf{w})_+\|_2 \leq 1 \quad \forall \ \theta_s \notin \Theta \end{array} \tag{4}$$

Indeed, it is easy to show that if a feature $\mathbf{\Phi}_{\theta_s}$ violates the third condition then adding that feature into the finite feature subset leads to a decrease of $J(\mathbf{w})$ objective value. In the same way, adding a feature $\mathbf{\Phi}_{\theta_s}$ which satisfies this third condition would not improve the objective value since it would get a zero weight.

From an algorithmic point of view, there is only one major difference when dealing with finite or infinite set of features. While in the former, it may be always possible to find the most violating constraint by sweeping all over them, in the latter case, finding this constraint is a difficult problem since we have to solve the maximization problem given Line 5 of Algorithm 1. Hence, the strategy we adopt consists in randomly sampling a given number of feature parameters $\{\theta_s\}$, so as to build a set of candidate features $\{\mathbf{\Phi}_{\theta_s}\}$, and then in adding to the finite set the one among those that violates the most its constraints. Similarly to the finite case, while not optimal this step always yields to a decrease in the objective value. Checking for the full optimality of the problem is also difficult since again one has to verify the third condition of Equation (4). In practice, we stop the algorithm when the maximal number of iterations is reached or when the sampling strategy does not return any violating feature.

Either in finite or infinite feature cases, solving Line 4 of the algorithm involves the resolution of problem (2) over the set of active features. Problem (2) is an unconstrained optimization problem where the loss function is differentiable with $L-$Lipschitz gradient and the mixed norm $\Omega_{1,2}$ admits a simple and closed-form proximal operator. Hence, the problem nicely fits into the class of problem that can be solved efficiently through fast iterative shrinkage thresholding algorithm (FISTA) [6] with some guarantee of convergence.

2.2 Examples of infinite set of features

There exists several feature extraction methods that need parameters tuning. In this paragraph, we review two of them that are of interest for our experimental analysis.

Gabor functions based features are typically features that need parameter tuning. Gabor features are usually built by convolving an image with a Gabor filter with angle φ, frequency f and Gaussian shape σ_1, σ_2 as parameters. Unlike usual approaches, instead of fixing these parameters to some predefined values, our approach is able to select them automatically during the learning process.

In BCI application such as motor imagery signal discrimination task, a common feature extraction consists in extracting the Power Spectral Density in a frequency band. Usually the common choice is [8,30] Hz for a motor imagery

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

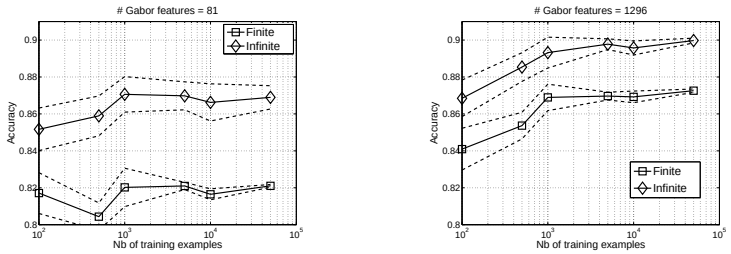

Fig. 1: Examples of accuracy performance in the finite and infinite feature cases with different numbers of sampled features (left) 81. (right) 1296. For these experiments, C has been set to 10.

task [7] but the choice of these frequency bands might be seen as parameter tuning in a continuous space.

3 Experiments

We describe in this section some experiments that show that using infinite set of features may help improving performances by selecting the ones that are relevant. We consider two real-world applications on texture recognition and EEG signal classification for Brain-Computer Interface.

The texture recognition problem consists in classifying 16×16 patches extracted from two Brodatz textures D29 and D92. The Gabor features are obtained by computing the inner product of all Gabor functions located at some pre-defined location and the patch and then by summing the absolute value of the response. By doing so, we obtain features that are translation-invariant. When Gabor parameters are sampled, we have a number of features that depends on the number of samples used for each parameter φ,f, σ_1 and σ_2. In order to be fair, in the infinite feature case, the number of sampled feature considered for the constraints violation checking problem (Line 5) is equal to the number of features used in the finite feature Gabor approach. We can see in Figure 1 that automatically selecting the relevant features from an infinite set leads to notable increase of performances.

The BCI dataset is the dataset IIa of the BCI competition IV. It consists of EEG signals of 9 subjects performing motor imagery. The signals have been acquired over 22 channels. In this study, we want to classify EEG trials of left/right hand motor imagery movement. For each class, we have 72 trials for learning and testing. For each trial, we have extracted the time segment from 0.5 to 2.5 seconds after the cue asking the subject to perform motor imagery. As a fixed feature, we have used the band-pass power over the $[8, 30]$ Hz, while for our infinite feature algorithm, we allow slight modification of the band-pass filter since we randomly draw filter which band-pass size of at least 20 Hz included in $[8, 30]$ Hz. By doing so, we hope that the classifier is able to adjust the most discriminant frequencies for each subject. In both cases, we perform CSP on the

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

	Subjects									
Methods	S1	S2	S3	S4	S5	S6	S7	S8	S9	Avg
CSP [7]	88.89	51.39	**96.53**	70.14	54.86	71.53	81.25	93.75	**93.74**	78.01
Fixed	88.19	**53.47**	**96.53**	63.89	60.42	69.44	79.17	**97.92**	93.06	78.01
Random	**90.97**	52.78	95.14	**73.61**	**62.50**	**72.92**	**82.64**	97.22	92.36	**80.01**

Table 1: Classification accuracy on the test set for classical CSP approach, fixed and random bandpass filter for feature extraction on the BCI dataset.

filtered EEG and include in the cross-validation stage the choice of the number of CSP filters and regularization parameter C. Results are summarized in Table 1 and they show that the advantage of learning the filter cut-off frequencies. We should however note that allowing more flexibilities to these cut-off frequencies leads to overfitting.

4 Conclusion

In this work, we proposed an efficient algorithm for selecting from an infinite set of features. This approach allows the automated selection of features with continuous parameters. The optimization problem is handled by an active set algorithm. At each iteration, the algorithm adds to the active feature set a relevant feature which is determined according to the optimality conditions of the problem. The approach was tested on a texture recognition dataset and on a BCI motor imagery task providing empirical evidences that dealing with infinite set of features may enhance performances of learning algorithms. Finally, as the algorithm has linear complexity and may be parallelized, we intend in future works to test our approach on large-scale datasets.

References

[1] G. Lanckriet, N. Cristianini, L. El Ghaoui, P. Bartlett, and M. Jordan. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

[2] Peter Gehler and Sebastian Nowozin. Let the kernel figure it out: Principled learning of pre-processing for kernel classifiers. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2009.

[3] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Trans. Image Processing*, 19(6):1635–1650, 2010.

[4] P. Bradley and O. Mangasarian. Feature selection via concave minimization and support vector machines. pages 82–99. Proceedings of the fifthteen International Conference in Machine Learning, Morgan Kaufmann, 1998.

[5] F. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the 21st International Conference on Machine Learning*, pages 41–48, 2004.

[6] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2:183–202, 2009.

[7] F. Lotte and C. Guan. Regularizing common spatial patterns to improve bci designs: Unified theory and new algorithms. *IEEE Trans Biomed Eng*, to appear, 2010.