Multi-Objective Optimization for SVM Model Selection

C. Chatelain, S. Adam, Y. Lecourtier, L. Heutte, T. Paquet
Laboratoire LITIS, Université de Rouen, Avenue de l'université,
76800 Saint Etienne du Rouvray, FRANCE

Abstract

In this paper, we propose a multi-objective optimization method for SVM model selection using the well known NSGA-II algorithm. FA and FR rates are the two criteria used to find the optimal hyperparameters of a set of SVM classifiers. The proposed strategy is applied to a digit/outlier discrimination task embedded in a more global information extraction system that aims at locating and recognizing numerical fields in handwritten incoming mail documents. Experiments conducted on a large database of digits and outliers show clearly that our method compares favorably with the results obtained by a state-of-the-art mono-objective optimization technique using the classical Area Under ROC Curve criterion (AUC).

1 Introduction

Tuning the hyperparameters of a Support Vector Machine (SVM) classifier is a crucial step in order to establish an efficient classification system. Generally, at least two parameter values have to be chosen carefully in advance. They concern respectively the regularization parameter (usually denoted as C), which sets the tradeoff cost between the training error and the complexity of the model, and the kernel function parameter(s), reduced to the bandwidth in the classical case of a Radial Basis Function kernel (usually denoted as γ). The problem of choosing these parameters values is called model selection in the literature and its results strongly impact the performance of the classifier.

During a long time, SVM model selection has been tackled using grid search. In such a case, the parameter space is explored with a fixed step size through a wide range of values and the obtained performance are assessed at each trial. It has been shown that such an approach is time consuming and does not perform well ([10, 12]).

Recently, model selection has been considered as an optimization task. In such a context, an optimization algorithm is used in order to find the hyperparameter set that reaches the best classification performance. Among existing optimization algorithms, the gradient descent method has been widely used for SVM model selection (see [2, 3] for example). However, such a method implies that both the score functions for assessing the performance of the hyperparameters and the SVM kernel have to be differentiable with respect to the parameters to be optimized. Moreover, the performance of gradient-based methods may strongly depend on the initialization.

Evolutionary algorithms have also been used for SVM model selection in order to overcome the above-mentioned problem. One can cite for example works described in [11] or in [18] which are based on the use of Genetic Algorithm (GA), or the approach proposed in ([9]) which is based on the use of Evolution Strategies. In both cases, the optimization algorithm is used to optimize C and γ regarding a classification performance indicator such as the predictive accuracy or the generalization error. One can note that in [11], the GA approach also aims at selecting relevant features.

In all the optimization-based approaches mentioned above, a single criterion is used as objective during the optimization process. However, it is well known that a single criterion is not always a good performance indicator. More precisely, a single criterion is unsuitable in the case of unbalanced classes or in the case of asymmetric misclassification costs, which are situations that arise very frequently in real-world problems. In such cases, the *a priori* probabilities of the classes and the misclassification costs must be considered together in order to characterize classification performance. However, the misclassification costs are often difficult to estimate, for example when the classification process is embedded in a more complex system. Within the context of a two-class problem, the Receiver Operating Characteristics (ROC) curve is known to be a better performance criterion. It represents the tradeoff between False Rejection (FR) and False Acceptance (FA) rates, also known as sensitivity *vs*. specificity tradeoff. Thus, for the optimization of a two-class classification problem, two criteria have to be minimized instead of the single predictive accuracy criterion.

In this paper, SVM model selection is considered as a Multi-Objective Optimization (MOO) problem. An Evo-

lutionnary Multi-Objective Optimization (EMOO) algorithm called Non dominated Sorting Genetic Algorithm II (NSGA-II) is applied to optimize the SVM hyperparameters using both FA and FR as criteria. Such a strategy enables to obtain in a single run a set of distinct classifiers which are optimal from the FA/FR rates criteria point of view. The performance of these classifiers cover a wide range of optimal FA/FR values. Consequently, it is possible to choose the one that best fits the application constraints at the end of the optimization process.

The proposed strategy is applied to a digit/outlier discrimination problem which takes place in a global information extraction system aiming at localizing and recognizing numerical fields in a handwritten incoming mail document [HIDDEN-REF]. Since the digit/outlier discrimination process is embedded in this more complex system, misclassification costs can not be estimated *a priori* and the best FA/FR tradeoff from the global system performance point of view (*i.e.* recall *vs.* precision of numerical field extraction) is unknown. The proposed strategy enables thus to overcome this problem as the choice of the parameter values is postponed after a single run of the optimization process.

The remainder of the paper is organized as follows. In section 2, we discuss the problem of SVM model selection. A brief introduction to support vector machines is proposed before discussing the choice of the criteria to be optimized in the model selection process. Section 3 describes the application of NSGA-II to SVM model selection. Then, in section 4, the application is described and some experimental results are given. In order to assess the performance of our multi-objective approach, we compare it with a state-of-the-art mono-objective one, *i.e.* using as criterion the Area Under ROC Curve (AUC). We demonstrate that our approach is a better solution for selecting the best SVM models. Finally, a conclusion and future works are drawn in section 5.

2 Problem Statement

2.1 SVM classifiers and their hyperparameters for model selection

As stated in [15], classification problems with asymmetric and unknown misclassification costs can be tackled using SVM through the introduction of two distinct penalty parameters C_- and C_+. In such a case, given a set of m training examples x_i in $\Re^n$ belonging to the class y_i :

$$(x_1, y_1) \ldots (x_m, y_m), x_i \in \Re^n, y_i \in \{-1, +1\}$$

the maximisation of the dual lagrangian with respects to the α_i becomes :

$$Max_\alpha \Big\{ \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \Big\}$$

$$\text{subject to the constraints: } \begin{cases} 0 \leq \alpha_i \leq C_+ \ \ for \ \ y_i = -1 \\ 0 \leq \alpha_i \leq C_- \ \ for \ \ y_i = +1 \\ \sum_{i=1}^{m} \alpha_i y_i = 0 \end{cases}$$

where α_i denote the Lagrange multipliers and $K(.)$ denotes the kernel. In the case of a Gaussian (RBF) kernel, $K(.)$ is defined as :

$$K(x_i, x_j) = exp\left(-\gamma \times \|x_i - x_j\|^2\right)$$

Hence, in the case of asymmetric misclassification costs, three parameters have to be determined to perform an optimal learning of the SVM classifier:

- The kernel parameter of the SVM-rbf: γ.
- The penalty parameters introduced above: C_- and C_+.

In the following, we call "hyperparameter" set a set of three given values for γ, C_- and C_+.

2.2 Criterion(a) for model selection

Considering model selection as an optimization process requires the choice of an (several) efficient criterion(a) to be optimized. As stated in the introduction, the ROC curve of a given classifier is a better performance indicator than a single predictive accuracy in the case of a two-class classification problem with unknown misclassification costs. However, using ROC curve as indicator in a model selection process involves to optimize two criteria, which is a much more complex task than optimizing a single criterion.

In the literature, some approaches have already been proposed in order to optimize the ROC curve, in the context of classifier learning (*i.e.* in order to optimize the classifier intrinsic parameters). Usually, such a problem is tackled using a reduction of the FR and FA rates into a single criterion such as the Area Under Curve (AUC) or the F-Measure (FM). It is the case for example in [17], where an AUC criterion is used to train the SVM classifier. In this work, the set of support vectors and the related α_i are found through the minimization of the AUC. This method is used in section 4 in comparison with the proposed approach. We refer to [17] for the details concerning the optimization process and the AUC computation. One can note that AUC based approaches were also proposed in [7] and in [13] in the case of non-SVM classifiers and that a similar approach based on F-measure is proposed in [14].

In all of these works, the aim is to design a classifier which is optimal with respect to the chosen performance indicator (AUC or FM). However, the performance indicator which is used is a resume of the ROC curve taken as a whole. Consequently, given one specific value for FA rate (resp. FR rate), such methods are enable to provide the classifier with the optimal value for the FR rate (resp. FA rate). This means that one single classifier optimizing the ROC curve is not guaranted to be the optimal classifier for any specific desired value of FA (resp. FR).

The proposed approach encompasses such traditional ROC curve optimization methods by searching for the set of optimal classifiers over the parameter space (FA,FR), using a real multi-objective optimization framework. This involves to use a multi-objective optimization algorithm in order to search for a set of hyperparameter sets, each hyperparameter set optimizing a given FA/FR tradeoff. Since the objective space dimension is greater than one, the dominance concept used in the MOO field has to be introduced to compare the performance of two classifiers.

2.3 Pareto dominance concept for SVM model selection

The dominance concept has been proposed by Vilfredo Pareto in the 19th century. A decision vector $\overrightarrow{u}$ (in our case, a given (C_+, C_-, γ) set) is said to dominate another decision vector $\overrightarrow{v}$ if $\overrightarrow{u}$ is not worse than $\overrightarrow{v}$ for any objective functions (FA and FR) and if $\overrightarrow{u}$ is better than $\overrightarrow{v}$ for at least one objective function. This is denoted $\overrightarrow{u} \prec \overrightarrow{v}$. More formally, in the case of the minimization of all the objectives, a vector $\overrightarrow{u} = (u_1, u_2, \ldots, u_k)$ dominates a vector $\overrightarrow{v} = (v_1, v_2, \ldots, v_k)$ if and only if:

$$\forall i \in \{1, \ldots, k\}, u_i \leq v_i \wedge \exists j \in \{1, \ldots, k\} : u_j < v_j$$

Using such a dominance concept, the objective of a MOO algorithm is to search for the Pareto Optimal Set (*POS*), defined as the set of all non dominated solutions of the problem. Such a set is formally defined as the set :

$$POS = \left\{ \overrightarrow{u} \in \vartheta / \neg \exists \overrightarrow{v} \in \vartheta, \overrightarrow{f(v)} \prec \overrightarrow{f(u)} \right\}$$

where ϑ denotes the feasible region (*i.e.* the parameter space regions where the constraints are satisfied) and $\overrightarrow{f}$ denotes the objective function vector. From a SVM model selection point of view, this POS corresponds to the optimal set of hyperparameter sets, *i.e.* the set of parametrized classifiers that yield all the optimal FA/FR tradeoffs. In the objective space, this set of optimal tradeoffs is called the Pareto front. One can note that in the context of SVM model selection, this Pareto front can be compared to a ROC curve since it describes a set of obtained FA/FR rates. However, in our context, it corresponds to the FA/FR rates obtained using a set of distinct and parametrized classifiers whereas a ROC curve is a performance indicator for a single classifier. A discussion concerning the relation between the Pareto front and classifier ROC curves is proposed in section 4 as an interpretation of the obtained results.

The approach described in this paper aims at approximating the Pareto optimal set corresponding to the optimal hyperparameters of a two-class SVM classifier using an EMOO. In the following section, a brief review of existing EMOO algorithms is proposed, the chosen algorithm is described and the application to SVM model selection is detailed.

3 EMOO for SVM model selection

As stated earlier, our objective is to search for a set of parametrized SVM classifiers corresponding to the optimal set of FA/FR tradeoffs. From a multiobjective optimization point of view, this set can naturally be seen as the Pareto optimal set and the set of corresponding FA/FR tradeoffs is the Pareto front. To tackle such a problem of searching a set of solutions describing the Pareto front, Evolutionary Algorithms (EA's) are known to be well-suited because they are able to search for multiple Pareto optimal solutions concurrently in a single run, through their implicit parallelism. This is why we do not consider in the following the approaches that optimize a single objective using the aggregation of different objectives into a single one (e.g. the use of the AUC) or the transformation of some objectives into constraints. For more details concerning these methods, see for example [4].

In the context of SVM model selection, computation of the objective values is very time consuming since it involves learning and testing the SVM for each hyperparameter set. Moreover, a good diversity of solutions is necessary since there is no *a priori* information concerning the adequate operating point on the Pareto front. That is why we have chosen to use NSGA-II in the context of our study. For more details about the NSGA-II algorithm, we refer to [5].

Application of NSGA-II for SVM model selection

In this subsection, the application of NSGA-II to SVM model selection problem is detailed. Two particular points have to be precised in such a context:

- the solution coding : as said before, three parameters are involved in the learning of SVM for classification problems with asymmetric misclassification costs: C_+,C_- and γ. These three parameters constitute the parameter space of our optimization problem. Consequently, each individual in NSGA-II has to encode these three real values. We have chosen to use a real

coding of these parameters in order to be as precise as possible.

- the evaluation procedure : each individual in the population corresponds to some given values of hyperparameters. In order to compute the performance associated to this individual, a classical SVM learning is performed using the encoded parameter values on a learning dataset. Then, this classifier is evaluated on a test dataset with the classical FA and FR rates as performance criteria.

4 Application and results

4.1 Digit/outlier discrimination

The work described in this paper is part of the design of a more complex system that aims at extracting numerical fields (phone number, zip code, customer code, etc.) from incoming handwritten mail document images . The main difficulty of such a task comes from the fact that handwritten digits may touch each other in the image while some textual parts sometimes are made of separated or touching characters. Figure 1 gives some examples of segmented components to deal with. In such a variable context, segmentation, detection and recognition of a digit and rejection of textual components must be performed altogether.

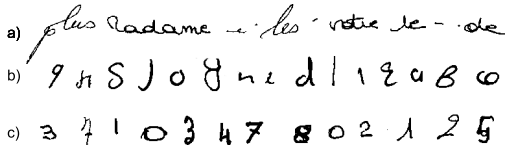

Figure 1. Examples of "obvious" (a) and "ambiguous" outliers (b), and digits (c).

In this paper, the proposed approach is applied to a particular stage of the numerical field extraction system [HIDDEN-REF].More precisely, the SVM to be optimized is used as a fast two-class classifier prior to the digit recognizer itself, aiming at filtering the "obvious outliers" (see fig 1.a) from all the other shapes (see fig 1.b and 1.c) in order to avoid a costly digit recognition when it is not necessary. The choice of the SVM classifier has been motivated by its efficiency in a two-class context. Its objective is to reject as many outliers as possible, while accepting as many digits as possible. Further stages of the system concern digit recognition and ambiguous outliers rejection. This context is a good example of a classification task with asymetric misclassification costs since the influence of a FA or a FR on the whole system results is unknown *a priori*. In the next subsection, the performance of the proposed system are assessed.

4.2 Experimental results and discussion

In this section, the experimental results obtained using the proposed approach are analysed. These results are compared to those obtained using a state-of-the-art algorithm ([17]), where a SVM classifier is trained with respect to an AUC criterion. Both NSGA and AUC-based approaches have been applied on a learning database of 7129 patterns (1/3 digit, 2/3 outliers), tested and evaluated on a test and a validation database of resp. 7149 and 5000 patterns with the same proportions of digits and outliers. In the case of NSGA-II, the range values for SVM hyperparameters are given in table 1. Concerning the NSGA-II parameters, we have used some classical values, proposed in [5]. Among them, one can note that the size of the population has been set to 40 in order to have enough points on the Pareto front. The resulting curves are presented in figure 2.

γ	C_-	C_+
$0-1$	$0-500$	$0-5000$

Table 1. Range values for γ, C_- and C_+.

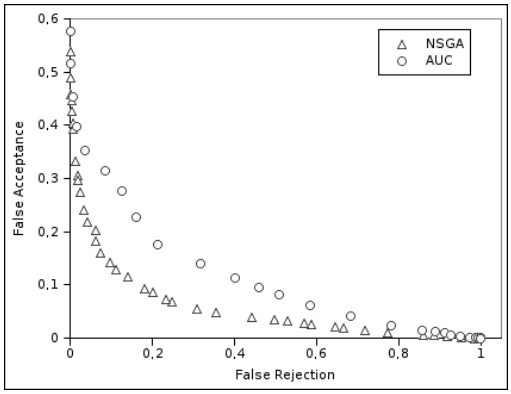

Figure 2. FA/FR curves obtained using NSGA-II and AUC.

Several comments can be made from the obtained results. First, one can remark that each point of the ROC curve obtained for a single classifier trained with AUC criterion is dominated by at least one of the point of the FA/FR curve determined by NSGA-II. Such a result stem from the fact that using an EMOO approach, FA and FR rates are minimized simultaneously through the variation of the three involved SVM hyperparameters whereas in the case of an AUC approach, a single parametrized classifier is trained

to optimize every possible FA/FR trade-offs. Consequently, one can argue that the Pareto front obtained using a multi-objective optimization can be viewed as the envelope of all the possible ROC curves. Such a result is interesting since the envelope constitutes for the practitionner an information *a priori* that can guide him for choosing a particular classifier.

Another comment concerning the proposed approach is that, for a given set of hyperparameters, the intrinsic parameters of the SVM classifiers (*i.e.* the position and weight of support vectors) are fixed using a mono-objective optimization algorithm well suited for such a task. Therefore, the EMOO concentrates on the choice of the hyperparameter values. This approach differs from other works using the EMOO to perform both intrinsic and hyperparameter setting. In the context of ROC curve optimization we can mention [1, 8, 6]. All these works are limited to non-complex classifiers (with a few number of intrinsic parameters) because EMOO algorithms rapidly become intractable when the size of the parameter space increases. Within a monoobjective context, such a limitation has been removed by developping specific methods for specifics problems like the Lagrangian maximisation for the SVM. Therefore, using the Lagrangian method for the tuning of SVM intrinsic parameters enables the EMOO algorithm to concentrate on a small number of hyperparameters.

Finally, let us point out that the EMOO may imply some overfitting. This should be fixed using the strategy presented in [16].

5 Conclusion

In this paper, we have presented a strategy to tackle the problem of SVM model delection with unknown misclassification costs. The approach is based on an Evolutionnary Multi-Objective Optimization of the SVM hyperparameters to depict an optimal FA/FR curve. Using such a curve, it is possible to choose the FA/FR tradeoff that best fits the application constraints.

The approach has been applied on a real classification problem, and compared favourably to a state-of-the-art approach based on the Area Under ROC Curve criterion.

The approach we propose is simple and generic. It can be applied to other parametric classifiers (KNN, Neural network, etc.). Moreover, it can be easily extended through the introduction of other parameters (kernel type) or objectives (number of support vectors, decision time). In our future works, we plan to apply a multi-objective optimization strategy to the whole numerical field extraction system, using recall and precision as criteria.

References

[1] M. Anastasio, M. Kupinski, and R. Nishikawa. Optimization and froc analysis of rule-based detection schemes using a multiobjective approach. *IEEE Trans. Med. Imaging*, 17:1089–1093, 1998.

[2] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1):131–159, 2002.

[3] K. Chung, W. Kao, C. Sun, and C. Lin. Radius margin bounds for support vector machines with rbf kernel. *Neural comput.*, 15(11):2643–2681, 2003.

[4] K. Deb. Multi-objective optimization using evolutionary algorithms. 2001.

[5] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan. A fast elitist nondominated sorting genetic algorithm for multiobjective optimization : Nsga-ii. In *Parallel problem solving from nature*, pages 849–858. 2000.

[6] R. Everson and J. Fieldsend. Multi-class roc analysis from a multi-objective optimisation perspective. *Pattern Recognition Letters*, page in press, 2006.

[7] C. Ferri, P. Flach, and J. Hernandez-Orallo. Learning decision trees using the area under the roc curve. *Proceedings of the 19th International Conference on Machine Learning*, pages 139–146, 2002.

[8] J. Fieldsend and R. Everson. Roc optimisation of safety related systems. *Proceedings of ROCAI 2004*, pages 37–44, 2004.

[9] F. Friedrichs and C. Igel. Evolutionnary tuning of multiple svm parameters. *Neurocomputing*, 64:107–117, 2005.

[10] C. Hsu and C. Lin. A simple decomposition method for support vector machine. *Machine Learning*, 46:219–314, 2002.

[11] C.-L. Huang and C.-J. Wang. A ga-based feature selection and parameters optimization for support vector machine. *Expert systems with application*, 31:231–240, 2006.

[12] S. Lavalle and M. Branicky. On the relationship between classical grid search and probabilistic roadmaps. *International Journal of Robotics research*, 23:673–692, 2002.

[13] M. C. Mozer, R. Dodier, M. D. Colagrosso, C. Guerra-Salcedo, and R. Wolniewicz. Prodding the roc curve: Constrained optimization of classifier performance. *NIPS*, pages 1409–1415, 2002.

[14] D. Musicant, V. Kumar, and A. Ozgur. Optimizing f-measure with support vector machines. *FLAIRS Conference*, pages 356–360, 2003.

[15] E. Osuna, R. Freund, and F. Girosi. *Support vector machines: Training and applications*. 1997.

[16] P. Radtke, T. Wong, and R. Sabourin. An evaluation of overfit control strategies for multi-objective evolutionary optimization. *WCCI/IJCNN 06*, pages 3327–3334, 2006.

[17] A. Rakotomamonjy. Optimizing auc with support vector machine. *European Conference on Artificial Intelligence Workshop on ROC Curve and AI*, pages 469–478, 2004.

[18] C.-H. Wu, G.-H. Tzeng, Y.-J. Goo, and W.-C. Fang. A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. *Expert systems with applications Article in Press*, 2006.