Non-cost-sensitive SVM training using Multiple Model Selection

Clément Chatelain, Sébastien Adam, Yves Lecourtier, Laurent Heutte, T. Paquet

Non-cost-sensitive SVM training using Multiple Model Selection

Clément Chatelain, Sébastien Adam, Yves Lecourtier, Laurent Heutte, and Thierry Paquet

Laboratoire LITIS EA 4108
UFR des Sciences, Université de Rouen, France.
{firstname.lastname}@univ-rouen.fr

Abstract. In this paper, we propose a multi-objective optimization framework for SVM hyperparameters tuning. The key idea is to manage a population of classifiers optimizing both False Positive (FP) and True Positive (TP) rates rather than a single classifier optimizing a scalar criterion. Hence, each classifier in teh population optimizes a particular trade-off between the objectives. Within the context of two-class classification problems, our work introduces the "ROC front concept" depicting a population of SVM classifiers as an alternative to the ROC curve representation. The comparison with a traditional scalar optimization technique based on an AUC criterion shows promising results on UCI datasets.

Key words: ROC front, multi-model selection, multi-objective optimization, ROC curve, SVM.

1 Introduction

Optimizing the hyperparameters of SVM classifiers is a complex challenge since it is well known that the choice of their values can dramatically affect the performance of the classification system. In the literature, many contributions in this field have focused on the computation of the model selection criterion, *i.e.* the value which is optimized with respect to the hyperparameters. These contributions have led to efficient scalar criteria and strategies used to estimate the expected generalization error. One can cite Xi-Alpha bound of [14], the Generalized Approximate Cross-Validation of [20], the empirical error estimate of [3], the radius-margin bound of [7] or the maximal-discrepancy of [2]. Based on these criteria, hyperparameters are usually chosen using a grid search, generally coupled with a cross-validation procedure. In order to decrease the computational cost of grid search, some authors suggest to use gradient-based techniques (e.g. [4], [15]). In these works, the performance validation function is adapted in order to be differentiable with respect to the parameters to be optimized.

All the approaches mentioned above, though efficient, use a single criterion as the objective during the optimization process. Now, it is well known that a single criterion is not always a good performance indicator. Indeed, in many

real-world pattern recognition problems (medical domain, road safety, biometry, etc...), the misclassification costs are (i) asymmetric as error consequences are class-dependant; (ii) difficult to estimate, for example when the classification process is embedded in a more complex system. In such cases, a single criterion might be a poor performance indicator.

One remedy to tackle this problem is to use as performance indicator the Receiver Operating Characteristics (ROC) curve which offers a synthetic representation of the trade-off between the True Positive rate (TP) and the False Positive rate (FP), also known as sensitivity vs. specificity trade-off. One way to take into account both FP and TP in the model selection process is to resume the ROC curve into a single criterion, such as the F-Measure (FM), the Break-Even Point (BEP) or the Area Under ROC Curve (AUC). However, we will show in the following that we can get more advantages in formulating the model selection problem as a true 2-D objective optimization task.

In this paper, our key idea is to turn the problem of the search for a global optimal SVM, (*i.e.* the best set of hyperparameters) using a single criterion or a resume of the ROC curve into the search for a pool of locally optimal SVM's (*i.e.* the pool of the best sets of hyperparameters) w.r.t. FP/TP rates. The best classifier among the pool can then be selected according to the needs of some practitioner. Consequently, the proposed framework can be viewed as a multiple model selection approach (rather than a model selection problem) and can naturally be expressed in a Multi-Objective Optimization (MOO) framework. Under particular conditions, we assume that such an approach could lead to better results than a more traditionnal single model selection approach based on a scalar optimization. Figure 1 depicts our overall multi-model selection process. The resulting output is a pool of classifiers, each one optimizing some FP/TP rate tradeoff. The set of trade-off values constitutes an optimal front we call "ROC front" by analogy with MOO field.

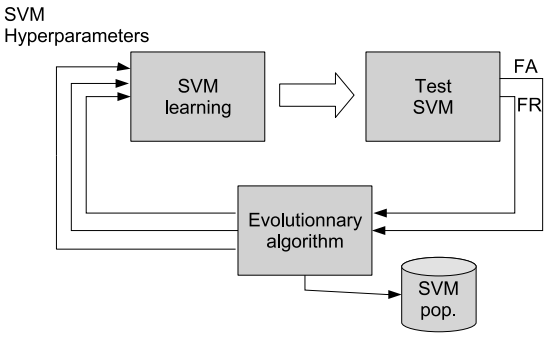

Fig. 1. SVM Multi-model selection framework

The remainder of the paper is organized as follows. In section 2, we detail the rationale behind the ROC front concept and illustrate how our multi-model selection approach can outperform traditional approaches in a MOO framework. In section 3, the SVM multi-model selection problem is adressed. Then, section 4 shows that our method compares favourably with traditionnal model selection techniques on standard benchmarks (UCI datasets). Finally, a conclusion and future works are drawn in section 5.

2 The "ROC front" concept

As stated in the introduction, a model selection problem may be seen from a multiobjective point of view, turning thus into a multi-model selection approach. In the literature, some multi-model selection approaches have been proposed. However, these approaches aim at designing a single classifier and thus cannot be considered as real multi model selection approaches. Caruana for example proposes in [6] an approach for constructing ensembles of classifiers, but this method aims at combining those classifiers in order to optimize a scalar criterion (accuracy, cross entropy, mean precision, AUC). Bagging, boosting or Error-correcting-output-codes (ECOC) ([10]) also aim at combining different classifiers of an ensemble in order to produce a single classifier efficient with respect to a scalar performance metric. In [17], an EA-based approach is applied to find the best hyperparameters of the set of binary SVM classifiers combined to produce a multiclass classifier.

The approach which is proposed in this paper is different since our aim is not to build a single classifier but a pool of classifiers using a real multi-objective framework. In such a context, let us recall that a problem arising when ROC curves are used to quantify classifier performance is their comparison in a 2-D objective space : a classifier may be better for one of the objectives (e.g. FP) and worse for the other one (e.g TP). Consequently, the strict order relation that can be used to compare classifiers when a single objective is considered become unusable and classical mono-objective optimization strategies can not be applied.

Usually, this problem is tackled using a reduction of the FP and TP rates into a single criterion such as the Area Under ROC Curve (AUC) ([19]). However, such performance indicators are a resume of the ROC curve taken as a whole and do not consider the curve from a local point of view. The didactic example proposed in figure 2 illustrates this statement. One can see on this figure two synthetic ROC curves. The curve plotted as continuous line has a better AUC value, but the corresponding classifier is not better for any specific desired value of FP rate (resp. TP). Consequently, optimizing such a scalar criterion to find the best hyperparameters could lead to solutions that do not fit the practitioner needs in certain context. It seems a better idea to simultaneously optimizing FP and TP rates using a MOO framework and a dominance relation to compare classifier performance.

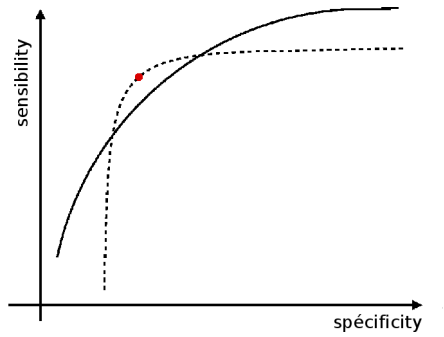

Fig. 2. Comparing ROC curves: the continuous ROC curve provides a better AUC than the dashed ROC curve, but is not locally optimal for a given range of specificity (False Positive Rate).

Let us recall that the dominance concept has been proposed by Vilfredo Pareto in the 19th century. A decision vector $\overrightarrow{u}$ is said to dominate another decision vector $\overrightarrow{v}$ if $\overrightarrow{u}$ is not worse than $\overrightarrow{v}$ for any objective functions 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 Multi-Objective Optimization 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 :

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

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. The corresponding values in the objective space constitute the so-called Pareto Front in the 2 dimension context.

From our model selection point of view, the POS corresponds to the pool of non-dominated classifiers (the pool of the best sets of hyperparameters). In this pool, each classifier optimizes a particular FP/TP trade-off. The resulting set of FP/TP points constitutes an optimal front we call "ROC front". This concept is illustrated with a didactic example as shown in figure 3: let us assume that ROC curves have been obtained from three distinct hyperparameter sets. This could lead to the three synthetic curves plotted as dashed lines. One can see on this example that none of the classifiers dominates the others on the whole range of FP/TP rates. An interesting solution for a practitioner is the "ROC front" (the red curve), which is made of non-dominated parts of a set of classifiers

ROC curves. The method proposed in this paper aims at finding this "ROC front" (and the corresponding *POS*), using an Evolutionnary Multi-Objective Optimization Algorithm called NSGA-II. This class of optimization algorithm has been chosen since Evolutionary Algorithms (EA's) are known to be well-suited to search for multiple Pareto optimal solutions concurrently in a single run, through their implicit parallelism.

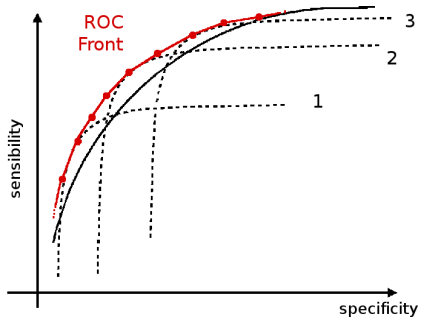

Fig. 3. Illustration of the ROC front concept : the ROC front depicts the FP/TP perfomance corresponding to the pool of non dominated operating points.

3 SVM multi-model selection

As explained in the preceding sections, the proposed framework aims at finding a pool of SVM classifiers, optimizing simultaneously FP and TP rates.

As stated in [18], 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 :

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

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

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

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

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, the proposed framework (see figure 1 is used in order to tune these three hyper-parameters values. For that, two particular points have to be specified for the application of NSGA-II to SVM multi-model selection :

- 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 FP and TP rates as performance criteria.

Let us now present some experimental results on several UCI databases.

4 Experimental results on UCI datasets

In this subsection, the proposed multi-model selection approach based on the ROC front concept is evaluated and compared with other approaches on publicly available benchmark datasets. First, the experimental protocol of our tests is described. Then, the results are shown and compared with some reference works, and finally several comments on these results are proposed.

Our approach has been applied on several 2-class benchmark datasets publicly available in the UCI Machine Learning repository on which state-of-the-art results have been published. The number of samples and the number of attributes for each problem are reported in table 1.

As we propose a real multi objective approach, the result of our experiment is a pool of classifiers describing the ROC Front. Thus, the evaluation of our approach is not easy since as mentioned in the introduction, comparing some results in a multi-dimensional space is a difficult task. Nevertheless, there exists some dedicated measures such as the Set Coverage Metric proposed in [22]. However, to the best of our knowledge, the other methods in the literature always

problem	# samples	# attributes
australian	690	14
wdbc	569	30
breast cancer	699	10
ionosphere	351	34
heart	270	13
pima	768	8

Table 1. Number of samples and number of attributes of the considered 2-class UCI problems.

consider a single classifier as a solution for a classification problem, which makes it difficult to compare our results with those in the literature.

Thus, the only way to evaluate our approach is to reduce the ROC front to a scalar criterion. For that, an Area Under the ROC Front (AUF) is calculated and compared with the Area Under the ROC Curve (AUC) of several approaches in the literature. We know that this comparison is not theoretically correct since the best results of a pool of classifiers is compared with a curve obtained by varying the threshold of a unique classifier. However, the aim of this comparison is only to highlight the fact that more interesting trade-offs may be locally reached with the ROC front approach. This comparison may also be justified by the fact that finally, in both cases only one classifier with a unique threshold will be retained for a given problem. Note also that the results of our approach are compared with several works based on the optimization of a scalar criterion for various classifiers. We emphasize that our comparison is based on the selection for each database of the best results found in the literature up to now: [5] (Decision lists and rules sets), [8] (Rankboost), [12] (Decision trees), [19] (SVMs) and [21] (five models : naive Bayes, logistic, decision tree, kstar, and voting feature interval). We refer to these papers for more explanation of the criterion and the model used.

Concerning the application of our multiobjective strategy, a cross validation procedure has been performed with 5 folds for each dataset. The results are presented in table 2, where the first column is the best AUC among the precited works based on the optimization of a scalar criterion, and the second one is the AUF of our approach.

As expected, one can see that for every dataset the ROC front yielded by the pool of classifiers leads to a higher area than the area under the ROC curve of the other single classifiers. As said before, it is important to emphasise that the AUF cannot theoretically be compared with AUC since the different operating points of the ROC front cannot be reached by a single classifier. However, this comparison with methods which explicitely optimize AUC clearly shows that our approach enables to reach very interesting local operating points which cannot be reached by the AUC-based classifiers. Hence, we claim that if the good model can be selected among the pool of classifiers, our approach can lead to better results than AUC-based methods.

problem	ref. AUC literature	AUF
australian	[21] 90.15 ± 0.53	96.91 ± 1.8
wdbc	[12] 94.7 ± 4.6	99.82 ± 0.1
breast cancer	[5] 99.13	99.86 ± 0.1
ionosphere	[19] 98.7 ± 3.3	99.14 ± 1.2
heart	[21] 92.60 ± 0.7	95.67 ± 1.5
pima	[8] 84.80 ± 6.5	88.36 ± 3.1

Table 2. Comparison of the Area Under the ROC Curve (AUC) in the literature with the Area Under the ROC Front (AUF).

Let us also remark that compared with other EMOO approaches, 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 [16, 1, 13, 11]. 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 avoided 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.

Computational issues

Assume that o is the learning complexity of the classifier, the learning complexity of the proposed approach $\mathcal{O}$ is defined as:

$$\mathcal{O} = M.N \times o$$

where we recall that M stands for the number of generations of the evolutionary algorithm, and N is the population size. Although high, this complexity can be compared with that of a full search approach, which is $\mathcal{O} = \prod_i^K D_i \times o$ for a K-hyperparameter problem, where D_i is the number of discrete values of each parameter. It is still higher than a gradient based approach, but as shown before leads to better results. Let us notice that the decision complexity of our approach is the same as the standard decision complexity of the involved classifier.

5 Conclusions

In this paper, we have presented a framework to tackle the problem of SVM model selection with unknown misclassification costs. The approach is based on a multi-model selection strategy in which a pool of SVM classifier is trained in order to depict an optimal ROC front. Using such a front, it is possible to choose

the FP/TP trade-off that best fits the application constraints. An application of this strategy with Evolutionnary Multi-Objective Optimization have been proposed, with a validation on UCI datasets. Obtained result have shown that our approach compares favourably to a state-of-the-art approach based on the Area Under ROC Cure criterion since better operating points can be locally reached. As a conclusion, one can say that an AUC-based approach suit pattern recognition problems where the operating point may vary, whereas our approach better suits problems where the operating point is supposed to be static.

The proposed approach is simple and generic. It can be applied to other parametric classifiers (KNN, Neural network, etc.) with other optimization methods ([9]). Moreover, it can be easily extended through the introduction of other parameters (kernel type) or objectives (number of support vectors, decision time). Note that a feature selection process can been included in the optimization method in order to provide a complete model selection approach. In our future works, we also plan to extend the approach to the multiclass problem.

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. D. Anguita, S. Ridella, F. Rivieccio, and R. Zunino. Hyperparameter design criteria for support vector classifiers. *Neurocomputing*, 55(1-2):109–134, 2003.
3. N. Ayat, M. Cheriet, and C. Suen. Automatic model selection for the optimization of svm kernels. *Pattern Recognition*, 30:1733–1745, 2004.
4. Y. Bengio. Gradient-based optimization of hyperparameters. *Neural Computation*, 12:1889–1900, 2000.
5. H. Bostrm. Maximizing the area under the roc curve using incremental reduced error pruning. *Proceedings of ROCML*, 2005.
6. R. Caruana, A. Niculescu-Mizil, G. Crew, and A. Ksikes. Ensemble selection from libraries of models. *proceedings of ICML*, 2004.
7. O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1):131–159, 2002.
8. C. Cortes and M. Mohri. Auc optimization vs. error rate minimization. *Advances in NIPS*, 2004.
9. B. de Souza, A. de Carvalho, R. Calvo, and R. P. Ishii. Multiclass svm model selection using particle swarm optimization. *Proceedings of HIS*, page 31, 2006.
10. T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.
11. R. Everson and J. Fieldsend. Multi-class roc analysis from a multi-objective optimisation perspective. *Pattern Recognition Letters*, pages 918–927, 2006.
12. C. Ferri, P. Flach, and J. Hernandez-Orallo. Learning decision trees using the area under the roc curve. *Proceedings of ICML*, pages 139–146, 2002.
13. J. Fieldsend and R. Everson. Roc optimisation of safety related systems. *Proceedings of ROCAI*, pages 37–44, 2004.
14. T. Joachims. Making large-scale support vector machine learning practical. *Advances in Kernel Methods: Support Vector Machines*, pages 169–184, 1998.
15. S. Keerthi, V. Sindhwani, and O. Chapelle. An efficient method for gradient-based adaptation of hyperparameters in svm models. *Advances in Neural Information Processing Systems 19*, 2007.
16. M. Kupinski and M. Anastasio. Multiobjective genetic optimization of diagnostic classifiers with implications for generating receiver operating characterisitic curves. *IEEE Trans. Med. Imaging*, 8:675–685, 1999.

17. G. Lebrun, O. Lezoray, C. Charrier, and H. Cardot. An ea multi-model selection for svm multiclass schemes. *Proceedings of IWANN*, pages 257–264, 2007.
18. E. Osuna, R. Freund, and F. Girosi. *Support vector machines: Training and applications*. AI Memo 1602, Massachusetts Institute of Technology, 1997.
19. A. Rakotomamonjy. Optimizing auc with support vector machine. *proceedings of ECAI Workshop on ROC Curve and AI*, pages 469–478, 2004.
20. G. Wahba, X. Lin, F. Gao, D. Xiang, R. Klein, and B. Klein. The bias-variance tradeoff and the randomized gacv. *Proceedings of NIPS*, pages 620–626, 1999.
21. S. Wu. A scored auc metric for classifier evaluation and selection. *proceedings of ROCML*, 2005.
22. E. Zitzler, K. Deb, and L. Thiele. Comparison of multiobjective evolutionary algorithms : Empirical results. *IEEE Transactions on Evolutionary Computation*, 2(8):173–195, 1999.