Ho–Kashyap with Early Stopping Versus Soft Margin SVM for Linear Classifiers –An Application

Fabien Lauer, Mohamed Bentoumi, Gérard Bloch, Gilles Millérioux, Patrice Aknin

Ho–Kashyap with Early Stopping vs Soft Margin SVM for Linear Classifiers – An Application

Fabien Lauer[1], Mohamed Bentoumi[1], Gérard Bloch[1], Gilles Millerioux[1], and Patrice Aknin[2]

[1] Centre de Recherche en Automatique de Nancy (CRAN UMR CNRS 7039)
ESSTIN, Rue Jean Lamour, 54519 Vandoeuvre Cedex, France
{fabien.lauer,bentoumi,bloch,millerioux}@esstin.uhp-nancy.fr
[2] Institut National de Recherche sur les Transports et leur Sécurité (INRETS)
2, avenue du Général Malleret-Joinville, 94114 Arcueil Cedex, France
aknin@inrets.fr

Abstract. In a classification problem, hard margin SVMs tend to minimize the generalization error by maximizing the margin. Regularization is obtained with soft margin SVMs which improve performances by relaxing the constraints on the margin maximization. This article shows that comparable performances can be obtained in the linearly separable case with the Ho–Kashyap learning rule associated to early stopping methods. These methods are applied on a non-destructive control application for a 4-class problem of rail defect classification.

1 Introduction

In a classification problem, after the parametrization and variable selection steps, the task is to choose the separating surface form (linear, polynomial ...) and the classifier structure.

In this paper, we discuss linear classification. We compare two learning methods: Ho–Kashyap learning rule [1] and Linear Support Vector Machine (SVM) [2], [3]. For SVM, regularization by soft margin is used to improve the generalization performance. [4] introduced the SVM concepts for Ho–Kashyap classifiers. Regularization is done by adding a parameter in the cost function to control the trade-off between model complexity and the amount of tolerated errors on the training set. We introduce early stopping as another regularization method to avoid overfitting. The learning is stopped before all the training examples are well classified.

We tested these methods in a 4-class linearly separable problem by creating 4 binary sub-classifiers. This application of rail defect classification provided a set of 140 observations which is not enough to split it to a training set and a validation set. Therefore, we used the Leave One Out cross-validation method for the generalization error estimation.

We start in Sect. 2 with some formalism on linear classification, before introducing Ho–Kashyap learning rule (Sect. 2.1) and SVM (Sect. 2.2). Then, in Sect. 3, we apply these methods on the application data and compare the results.

2 Linear Binary Classification

In a binary classification linear problem, the task is to find a separating hyperplane that can separate 2 classes. Let $(x_i,\ y_i)_{1\leq i\leq N}$ be a set of training examples with $x_i \in \mathbb{R}^p$ belonging to a class labeled by $y_i \in \{+1,\ -1\}$. The decision function of a linear classifier is:

$$f(x) = \text{sign}\left(\langle w, x\rangle + b\right) \tag{1}$$

where $\langle .,. \rangle$ stands for dot product and $(w \in \mathbb{R}^p,\ b \in \mathbb{R})$ are the parameters of the separating hyperplane. If all the training examples are correctly separated, then:

$$y_i\left(\langle w, x_i\rangle + b\right) > 0 \quad i = 1, \ldots, N\ . \tag{2}$$

2.1 Ho–Kashyap Learning Rule (HK)

Amongst other learning rules for linear classifiers design (perceptron, LMS, linear programming algorithms... [5]), Ho and Kashyap [1] proposed an iterative gradient descent-based algorithm. Defining a set of N $(p+1)$-dimensional vectors X_i:

$$X_i^t = \begin{cases} (+1, x_i^t) & \text{, if } y_i = +1 \\ (-1, -x_i^t) & \text{, if } y_i = -1 \end{cases} \tag{3}$$

and a $(p+1)$-dimensional weight vector $W = (b, w^t)^t$ allows to write (2): $\langle W^t, X_i\rangle > 0, i = 1, \ldots, N$. Then defining a $(p+1 \times N)$-matrix $X = [X_1\ X_2 \ldots X_N]$ gives:

$$W^t X > 0\ . \tag{4}$$

Let B be the "margin" vector with b_i as components. Equation (4) can be rewritten as:

$$\begin{aligned} W^t X &= B^t \\ \text{subject to} \quad b_i &> 0 \quad i = 1, \ldots, N\ . \end{aligned} \tag{5}$$

Ho–Kashyap (HK) learning rule solves (5) by minimizing the least squares criterion $J(W, B) = \|W^t X - B^t\|^2$. The margin vector is first initialized to B_0 with all b_i set to small positive values. At each step k, the weight vector W_k is deduced from B_k by:

$$W_k^t = B_k^t X^\dagger \tag{6}$$

where $X^\dagger = X^t(XX^t)^{-1}$ stands for the pseudo-inverse of X. Then a gradient descent is used to compute a new estimate of the margin vector:

$$B_{k+1}^t = B_k^t - \mu\frac{1}{2}\left(\nabla_B J(W, B) - |\nabla_B J(W, B)|\right) \tag{7}$$

with μ a positive learning rate.

In order to satisfy the constraints $b_i > 0$, the positive components of $\nabla_B J(W, B)$ are set to 0, thus preventing b_i to decrease and become negative. This is why $\frac{1}{2}\left(\nabla_B J(W, B) - |\nabla_B J(W, B)|\right)$ is used instead of $\nabla_B J(W, B)$.

It can be shown [5] that this procedure converges in a finite number of steps $\forall\, \mu,\ 0 < \mu < 1$, to 0 in the separable case, to a non-zero value otherwise. This makes the tuning of μ not critical.

2.2 Linear SVM

For Linear Support Vector Machine (SVM) binary classifiers, (2) becomes [2]:

$$y_i \left(\langle w, x_i \rangle + b \right) \geq 1 \quad i = 1, \ldots, N \ . \tag{8}$$

We consider now the points that ensure equality in (8). These points belong to the so called *canonical hyperplanes* $H_1 : \langle w, x_i \rangle + b = 1$ and $H_2 : \langle w, x_i \rangle + b = -1$. The distance Δ which separates H_1 and H_2 is equal to $2/\|w\|$ and is called the *margin*. The main difference introduced by a SV classifier is that the optimal separating hyperplane is the one that ensures a maximal margin [2][3], i.e. minimizes $\|w\|$. To build a so called *hard margin SV* classifier, the task is therefore:

$$\begin{array}{ll} \min & W(w,b) = \frac{1}{2} \|w\|^2 \\ \text{subject to} & y_i \left(\langle w, x_i \rangle + b \right) \geq 1 \end{array} \tag{9}$$

which is equivalent to the maximization problem of the *dual Lagrangian*:

$$\begin{array}{c} \max L_{dual} = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \\ \text{subject to } \alpha_i \geq 0, \ i = 1, \ldots, N \text{ and } \sum_{i=1}^{N} \alpha_i y_i = 0 \end{array} \tag{10}$$

where α_i are the Lagrange multipliers. The solution $(\hat{\alpha}_i)$ of (10) allows to determine the couple $(\hat{w}, \hat{b})$:

$$\hat{w} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i \ , \quad \hat{b} = -\frac{1}{2} \langle \hat{w}, \ x_r + x_s \rangle \ , \quad \hat{\alpha}_r, \ \hat{\alpha}_s > 0 \tag{11}$$

where x_r and x_s are two examples for which the corresponding class labels are $y_r = -1$ et $y_s = +1$. The decision function (1) of a SVM classifier is thus given by:

$$f(x) = \text{sign} \left(\sum_{i=1}^{N} \hat{\alpha}_i y_i \langle x_i, x \rangle + \hat{b} \right) \ . \tag{12}$$

From the Karush-Kuhn-Tucker (KKT) conditions [3], we have:

$$\hat{\alpha}_i \left(y_i \left[\langle w, x_i \rangle + b \right] - 1 \right) = 0, \quad i = 1, \ldots, N \tag{13}$$

and therefore only for the points x_i which satisfy $y_i \left[\langle w, x_i \rangle + b \right] = 1$, Lagrange multipliers are non zero: $\hat{\alpha}_i > 0$. These points are called *Support Vectors* (SV).

Most of the time, in practice, the training set contains noise and outliers and a SV classifier calculated from this set can lead to poor generalization. To tackle this problem, slack variables ξ_i which allow errors on the constraints can be introduced. Now, we have the so called *soft margin SVM* problem to solve:

$$\begin{array}{ll} \min & W(w,b) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i \\ \text{subject to} & y_i \left(\langle w, x_i \rangle + b \right) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{array} \tag{14}$$

where C is the regularization parameter which controls the trade-off between training error and model complexity and has to be determined beforehand. Solving the quadratic optimization problem (14) leads to the same dual Lagrangian maximization (10) but subject to:

$$0 \leq \alpha_i \leq C,\ i = 1, \ldots, N \text{ and } \sum_{i=1}^{N} \alpha_i y_i = 0\ . \tag{15}$$

Two categories of SV can be distinguished: the well classified SV which have $0 < \hat{\alpha}_i < C$, and the misclassified SV which have $\hat{\alpha}_i = C$.

2.3 Generalization and Hyperparameters Tuning

In classification the goal is to minimize the error on future examples which is called the generalization error (GE). The most popular technique for the GE estimation is the *cross-validation* that is independent of the learning machine used. The Leave One Out procedure (LOO) is a cross-validation procedure adapted for a weak data number N and giving an almost unbiased estimation of GE [6]. It consists in dividing the training set in two subsets: a learning subset of $N-1$ examples and a test subset containing only one example. The procedure is repeated N times until all the examples are tested. The estimation of GE is then given by the number of misclassified test examples over N. To lighten calculations, an upper bound of GE can be calculated: k-fold cross-validation which is similar to LOO except that the training set is divided in k subsets. One subset is left for testing and $k-1$ subsets are used for learning. The procedure is thus only repeated k times (typically, $k = 5$ or 10). LOO can be seen as the extreme case of the k-fold cross-validation, where $k = N$.

To avoid overfitting, a certain amount of misclassified training examples can be accepted. In SVM, this is introduced by the soft margin and the regularization parameter C. To tune C, a range of values is scanned and the optimal value is the one corresponding to the minimum of the GE estimation.

For the Ho–Kashyap learning rule, early stopping can make the learning process stop before all the training examples are well classified. Early stopping can be achieved by looking at the GE estimation during the training process and stopping as soon as it is rising. But this method does not yield always to the best minimum of GE (see for instance [7]). In another approach, the training is not stopped but GE is evaluated at all the iterations during the process. Then the lowest GE gives the best classifier. This method can be included in the HK learning to best tune the hyperparameter, here only n, the number of iterations. Indeed, a change in μ yields to a change in n which is automatically tuned.

3 Application

3.1 Context

The application concerns the classification of rail defects signatures. Previous works led to the realization of a suitable double-coils and double-frequencies

differential eddy current sensor [8] which can be embarked on a train. After preprocessing, four complex channels (active and reactive parts) are available, which are equivalent to eight real signals.

Tests have been made on a complete subway track. The defects were labeled in 4 classes: switches (ω_1), fishplated joints (ω_2), welded joints (ω_3) and shellings (ω_4). This provided a training set of 140 observations for a 4-class classifier. One observation consists in a window of 500mm width (100 points considering sampling step is 5mm) for each of the 8 signals. The Modified Fourier Descriptors (MFD) [9] result from the 12 first coefficients C_j of the Discrete Fourier Transform (DFT) of the signals of the window by: $d_j = C_j C_{-j} / |C_1 C_{-1}|$, $j = 1, \ldots, 12$. The number of parameters is thus $p = 96$.

The class-rest approach to the 4-class problem is to split it into 4 binary problems with 4 sub-classifiers dedicated to the separation of one class among the others. Thus, a different subset of variables can be chosen for each. The Orthogonal Forward Regression (OFR) procedure has been applied to rank the parameters with respect to their contribution to each sub-classifier output. Together with the decision criterion introduced in [10], this reduced the input dimensions from $p = 96$ to respectively $p = 15, 15, 8$ and 9. In order to raise ambiguity, the maximum of the 4 sub-classifiers outputs gives the class of the example.

3.2 Results and Comparisons

Table 1 compares the generalization performances (LOO) evaluated with the *Leave One Out* procedure for the Ho–Kashyap learning rule with (*HK opt*) or without (*HK inf*) early stopping, SVM with soft margin (*SVM soft*) and SVM with hard margin (*SVM hard*) classifiers. The percentages of well classified examples on the training set are given in (*TR set*). The results are presented for each sub-classifier as well as for the global classifier before (*Global 1*) and after (*Global 2*) raising ambiguity.

The tuning of the hyperparameters is done as described in Sect. 2.3. Since for SVMs, LOO procedure is too much time consuming, 5-fold procedure was used to estimate GE for the tuning of C. Here is an advantage of the HK learning rule: its speed, thanks to which the tuning of n can be made with the LOO estimation of GE which is closer to GE than the 5-fold estimation.

Performances of *HK inf* are similar to the ones of *SVM hard* though it does not maximize the margin. When regularization is used, performances increase in a comparable way for both *SVM soft* and *HK opt*. Different values of μ have been tried for the Ho–Kashyap rule and it showed that it does not really affect the results (by 1 misclassified example in the worst case) but only n in the early stopping procedure.

4 Conclusion

We reviewed some formalism on linear classification and particularly on linear SVM classifiers. In the particular case of linear classification and on our applica-

Table 1. Classification performances

	HK inf		**SVM hard**		**HK opt**		**SVM soft**	
	TR set	LOO	TR set	LOO	TR set	LOO	TR set	LOO
Class ω_1/others	100	94.29	100	90.71	99.29	96.43	99.29	96.43
Class ω_2/others	100	96.43	100	96.43	100	99.29	100	99.29
Class ω_3/others	100	95.71	100	95.00	99.29	97.14	99.29	96.43
Class ω_4/others	100	97.86	100	97.86	100	97.86	100	97.86
Global 1	100	87.14	100	82.86	98.57	92.14	98.57	91.43
Global 2	100	95.71	100	93.57	99.29	97.14	99.29	96.43

tion, SVM classifiers give very good generalization performances, as expected, by maximizing the margin with only one hyperparameter C to tune. But we showed that a hyperplane trained with a simple learning rule such as Ho–Kashyap can achieve comparable performances with the introduction of early stopping in the learning process and one hyperparameter, the number of iterations n. We used a procedure to tune the hyperparameter C using a simplified cross-validation method, k-fold, to lighten calculation for SVM. Ho–Kashyap learning proved faster and the almost unbiased LOO estimation of GE could be used for the tuning of n.

References

1. Ho, E., Kashyap, R.L.: An Algorithm for Linear Inequalities and its Applications. IEEE Trans. Electronic Computers **14** (1965) 683-688
2. Burges, C.: A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery **2** (1998) 121-167
3. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press (2000)
4. Łęski, J.: Ho–Kashyap Classifier With Generalization Control. Pattern Recognition Letters **24** (2003) 2281-2290
5. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. 2nd edn. Wiley (2000)
6. Duan, K., Keerthi, S.S., Poo, A.N.: Evaluation of Simple Performance Measures for Tuning SVM Hyperparameters. Neurocomputing **51** (2003) 41-59
7. Prechelt, L.: Early Stopping – But When ? In: Neural Networks: Tricks of the Trade. (1998) 55-69
8. Oukhellou, L., Aknin, P., Perrin, J-P.: Dedicated Sensor and Classifier of Rail Head Defects for Railway Systems. Control Engineering Practice **7** (1999) 57-61
9. Oukhellou, L., Aknin, P.: Modified Fourier Descriptors: a New Parametrization of Eddy Current Signature Applied to the Rail Defect Classification. In: III International Workshop on Advances in Signal Processing for Non Destructive Evaluation of Materials, Québec. (1997)
10. Oukhellou, L., Aknin, P., Stoppiglia, H., Dreyfus, G.: A New Decision Criterion for Feature Selection: Application to the Classification of Non Destructive Testing Signatures. In: European Signal Processing Conference (EUSIPCO), Rhodes, Greece. (1998)