One-Class SVM sans biais

Sébastien LECOMTE[1,2], Régis LENGELLÉ[2], Cédric RICHARD[3], François CAPMAN[1], Bertrand RAVERA[1]

[1]Laboratoire Multi-MediaProcessing, Thales Communications and Security, Gennevilliers.

[2]Institut Charles Delaunay - LM2S, UMR STMR, Université de Technologie de Troyes.

[3]Laboratoire Lagrange, UMR CNRS 7293, Observatoire de la Côte d'Azur, Université de Nice Sophia-Antipolis.

prenom.nom@thalesgroup.com, prenom.nom@utt.fr, prenom.nom@unice.fr

Résumé – Nous introduisons un problème SVM 1-classe sans biais, exploitant les contraintes du problème 2-classes. A travers l'écriture d'une forme duale unifiée des problèmes SVM 1-classe et 2-classes, nous montrons qu'il est possible d'utiliser l'algorithme *Sequential Maximization Gradient Optimization* (SMGO), indépendemment du problème. Ce résultat fait bénéficier l'approche 1-classe de l'ensemble des perspectives de l'algorithme SMGO. Des résultats expérimentaux montrent que l'algorithme SMGO pour le problème 1-classe est aussi performant et plus rapide que les approches de l'état de l'art (libSVM, SVM-light, Fast-OC2).

Abstract – We introduce a One-Class SVM problem without offset, exploiting 2-classes problem constraints. Through an unified formulation of the 1-class and 2-classes SVM dual, we show that the Sequential Maximization Gradient Optimization (SMGO) algorithm can be used for both problems. This allows 1-class SVM problem to benefit of SMGO perspectives. Experiments highlight that SMGO for 1-class problem is as performant as and faster than state-of-the-art algorithms (libSVM, SVM-light, Fast-OC2).

1 Introduction

On considère le problème de détection d'anormalité à l'aide de Machine à Vecteurs de Support (SVM) 1-classe [6]. Cette approche est particulièrement populaire, comme l'illustre une grande variété d'applications [14, 5, 7]. La surveillance audio nous intéresse particulièrement [3], et l'un des challenges à relever consiste en la réduction du temps d'apprentissage pour traiter de très grandes bases de données.

Nous souhaitons résoudre le problème SVM 1-classe à l'aide l'algorithme *Sequential Maximization Gradient Optimization* (SMGO) [10]. Celui-ci, comme l'approche analogue présentée dans [8], se montre rapide et peu consomateur de mémoire [1]. Cette approche s'appuie néanmoins sur une formulation sans biais du problème SVM 2-classes, ce qui la rend inexploitable pour la résolution du problème SVM 1-classe standard [6].

Par ailleurs, [11] a reformulé le problème SVM 1-classe en y incluant les contraintes du problème 2-classes. Cela permet de modéliser une classe unique en s'assurant de rejeter des données d'autres classes ou identifiées comme anormales. [11] introduit également l'algorithme d'optimisation dédié Fast-OC2. Les performances de cette approche étant excellentes, nous allons comparer les résultats sur des problèmes 1-classe entre SMGO (sans biais) et Fast-OC2 (avec biais). Notons également que ces deux algorithmes peuvent être utilisés sans les contraintes du problème 2-classes, nous discuterons également des avantages de cette approche.

1. [10] compare SMGO à l'implémentation SVM-light et [8] à l'implémentation libSVM.

La section 2 est consacrée au rappel des différents problèmes SVM, 1-classe et 2-classes standards, puis 2-classes sans biais. Dans la section suivante, nous développons notre approche d'un problème SVM 1-classe sans biais. A travers l'expression d'une forme duale unifiée, nous montrons dans la section 4 que ce problème peut être résolu par l'algorithme SMGO. Enfin, nous illustrons l'intérêt de cette approche en la comparant avec les performances de l'algorithme Fast-OC2.

Notations. Soit $S = \{(\mathrm{x}_i, y_i), i = 1, \ldots, N\} \in (\mathcal{X} \times \mathcal{Y})^N$ un ensemble d'apprentissage avec $\mathcal{Y} = \{-1, +1\}$ ou $\mathcal{Y} = \{1\}$. Dans le contexte des méthodes à noyaux, on définit l'application ϕ telle que $\mathcal{X} \ni \mathrm{x} \mapsto \phi(\mathrm{x}) \in \mathcal{H}$ où $\mathcal{H}$ est un espace de Hilbert à noyau reproduisant. Par restriction, le noyau induisant $\mathcal{H}$ est de norme unité (typiquement, un noyau RBF Gaussien). On note κ la fonction noyau et $\mathbf{H}$ suivant [10] la matrice des éléments $\{h_{i,j} = -y_i y_j \kappa(\mathrm{x}_i, \mathrm{x}_j); i, j = 1 \ldots N\}$. Enfin, par convention, les signe en gras correspondent à des vecteurs (par exemple $\boldsymbol{\alpha} = \{\alpha_i, i = 1 \ldots N\}$).

2 Contexte

L'approche SVM 2-classes avec biais (C-SVM) recherche un séparateur $\mathcal{W}$ dans $\mathcal{H}$ tel que $\mathcal{W} = \{\mathrm{x} \in \mathcal{X}, \langle \mathbf{w}, \phi(\mathrm{x})\rangle + b = 0\}$ [12]. Motivée par cette illustration géométrique, la fonction de décision $f(\mathrm{x}) = \langle \phi(\mathrm{x}), \mathbf{w}\rangle + b$ est introduite *a priori*. L'unicité de $\mathbf{w}$ et b est garantie par la maximisation de la marge entre $\mathcal{W}$ et deux hyperplans définis par $\{\mathrm{x} \in \mathcal{X}, \langle \mathbf{w}, \phi(\mathrm{x})\rangle + b =$

$\pm 1\}$. Soient ξ_i des variables de relâchement, le problème primal C-SVM s'exprime :

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{sous les contraintes} \quad \begin{cases} \xi_i \geq 1 - y_i f(\mathrm{x}_i) \\ \xi_i \geq 0 \end{cases}$$

où C contrôle le compromis entre maximisation de la marge et minimisation des erreurs. La méthode des multiplicateurs de Lagrange [1] permet de formuler le problème dual correspondant :

$$\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha} + \boldsymbol{\alpha}^T\mathbf{1}_N \tag{1}$$
$$\text{sous les contraintes} \quad \begin{cases} \mathbf{0}_N \leq \boldsymbol{\alpha} \leq \mathbf{1}_N C \\ \boldsymbol{\alpha}^T y = 0 \end{cases}$$

avec α_i les variables duales. A partir des conditions d'optimalité (KKT [2]), on exprime la solution $\mathbf{w} = \sum_{i=1}^{N}\alpha_i y_i \phi(\mathrm{x}_i)$. Le biais $b = y_k - \sum_{i=1}^{N}\alpha_i y_i \kappa(\mathrm{x}_i, \mathrm{x}_k)$ est calculé *a posteriori* pour tout k tel que $0 < \alpha_k < C$ (vecteur de support non borné).

L'approche SVM 1-classe avec biais (OC2-SVM) recherche une fonction de décision positive dans une région de volume minimal qui capture la plupart des observations de S, et négative ailleurs. [6] pose $f(\mathrm{x}) = \langle \phi(\mathrm{x}), \mathbf{w}\rangle - b$ cette fonction, définie *a priori*. L'approche consiste à maximiser le biais $b > 0$ pour séparer les données projetées dans $\mathcal{H}$, de l'origine de cet espace. Tohmé [11] propose de rejeter des observations incohérentes ou issues d'autres classes, alors étiquetées -1, en utilisant les contraintes du problème 2-classe, soit $\xi_i \geq -y_i f(\mathrm{x}_i)$ au lieu de $\xi_i \geq -f(\mathrm{x}_i)$. Ainsi, le problème primal OC2-SVM s'exprime :

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 - b + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i$$
$$\text{sous les contraintes} \quad \begin{cases} \xi_i \geq -y_i f(\mathrm{x}_i) \\ \xi_i \geq 0 \end{cases}$$

où le terme $\frac{1}{\nu N}$ est analogue au terme C du problème C-SVM. Le problème dual correspondant s'exprime alors :

$$\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha} \tag{2}$$
$$\text{sous les contraintes} \quad \begin{cases} \mathbf{0}_N \leq \boldsymbol{\alpha} \leq \mathbf{1}_N \frac{1}{\nu N} \\ \boldsymbol{\alpha}^T\mathbf{y} = 1 \end{cases}$$

Résultant de l'introduction *a priori* d'une fonction de décision avec biais, les problèmes duaux (1) et (2) sont sujets à une contrainte d'égalité. On introduit maintenant l'approche SVM 2 classes sans biais, récemment proposée par Steinwart [8].

L'approche SVM 2-classes sans biais (WO-SVM) recherche une fonction de décision par minimisation d'un risque empirique [12] et régularisation [9]. Il s'agit de résoudre un problème de la forme :

$$\min_{f\in\mathcal{H}} \|f\|^2 + C\sum_{i=1}^{N} c(f(\mathrm{x}_i), y_i) \tag{3}$$

où c est une fonction perte. La perte charnière $c(f(\mathrm{x}), y) = \max(0, 1 - yf(\mathrm{x}))$ (voir figure 1) traite le problème SVM 2-classes. La fonction de décision est donnée *a posteriori* par le théorème du représentant [13] : $f(\mathrm{x}) = \sum_{i=1}^{N}\beta_i\kappa(\mathrm{x}_i, \mathrm{x})$ où les β_i sont des coefficients à déterminer. Une fois f substitué par ce résultat dans (3), ces coefficients deviennent les variables primales du problème à résoudre et les conditions KKT permettent d'identifier $\beta_i = \alpha_i y_i$. f s'exprime alors $f(\mathrm{x}) = \sum_{i=1}^{N}\alpha_i y_i \kappa(\mathrm{x}_i, \mathrm{x})$ et problème dual WO-SVM est :

$$\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha} + \boldsymbol{\alpha}^T\mathbf{1}_N \tag{4}$$
$$\text{sous les contraintes} \quad \mathbf{0}_N \leq \boldsymbol{\alpha} \leq \mathbf{1}_N\frac{1}{\nu N}$$

Bénéficiant de l'absence de contrainte d'égalité, Steinwart introduit un algorithme d'optimisation utilisant de nouvelles options de démarrage à chaud et un ensemble de stratégies peu coûteuses réduisant significativement le nombre d'itérations. Cet algorithme est équivalent au SMGO introduit par Tohmé [10].

3 Problème 1-classe sans biais

On souhaite adopter la stratégie du problème OC2-SVM au sein d'une approche similaire à celle WO-SVM. Les contraintes d'inégalité du problème primal OC2-SVM sont :

$$\begin{cases} \xi \geq -yf(\mathrm{x}) \\ \xi \geq 0 \end{cases}$$

Leur analyse permet de construire une fonction perte adaptée au problème 1-classe avec contraintes binaires (voir figure 1) :

$$c(f(\mathrm{x}), y) = \max(0, -yf(\mathrm{x})) \tag{5}$$

Cependant, en l'absence de biais maximisé, cette fonction perte ne correspond pas au problème que l'on souhaite traiter. En effet, comme l'illustre la figure 1, la marge est implicitement fixée à 0. Afin de construire une fonction perte exploitable pour le problème SVM 1-classe, on propose d'estimer la fonction $\tilde{f}(\mathrm{x}) = f(\mathrm{x}) + 1$. Cette astuce contraint l'hyperplan recherché à se situer à une distance $\frac{1}{\|\mathbf{w}\|}$ de l'origine. En pratique, on propose de réécrire la perte charnière (5) en ce sens :

$$c(f(\mathrm{x}), y) \quad = \quad \max(0, y - y\tilde{f}(\mathrm{x})) \tag{6}$$

Le problème primal 1-classe sans biais s'exprime alors en substituant (6) dans (3) et le problème dual WOOC2-SVM correspondant est :

$$\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha} + \mathbf{y}^T\boldsymbol{\alpha} \tag{7}$$
$$\text{sous les contraintes} \quad \mathbf{0}_N \leq \boldsymbol{\alpha} \leq \mathbf{1}_N\frac{1}{\nu N}$$

La solution de ce problème permet de déterminer la fonction $\tilde{f}$. En définitive, la fonction de décision du problème 1-classe sans biais vaut : $f(\mathrm{x}) = \sum_{i=1}^{N}\boldsymbol{\alpha}_i y_i \kappa(\mathrm{x}_i, \mathrm{x}) - 1$.

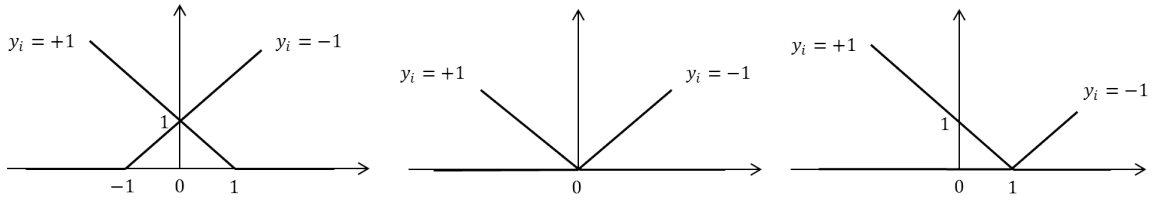

FIGURE 1 – Fonction de perte charnière pour le problème 2-classe (à gauche), pour le problème 1-classe sans biais en fonction de $f(\mathrm{x})$ (au milieu) et pour le problème 1-classe sans biais en fonction de $\tilde{f}(\mathrm{x}) = f(\mathrm{x}) + 1$ (à droite)

4 Problème dual unifié

On montre maintenant que l'on peut écrire les problèmes duaux sans biais sous une forme unifiée permettant d'utiliser les mêmes algorithmes pour la résolution des problème 1 et 2 classe(s). Les problèmes duaux (4) et (7) sont tous les deux de forme quadratique. De plus, il y a équivalence entre les contraintes d'inégalité et entre les fonctionnelles de coût. Soient $\boldsymbol{\delta} = \mathbf{y}$ dans le cas 1-classe et $\boldsymbol{\delta} = \mathbf{1}_N$ dans le cas 2-classes, on pose le problème dual unifié :

$$\max_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{H}\boldsymbol{\alpha} + \boldsymbol{\delta}^T \boldsymbol{\alpha}$$

$$\text{sous les contraintes} \quad \mathbf{0}_N \leq \boldsymbol{\alpha} \leq \mathbf{1}_N C$$

et la fonction de décision se met sous la forme :

$$f(\mathrm{x}) = \sum_{i=1}^{N} \alpha_i y_i \kappa(\mathrm{x}_i, \mathrm{x}) - \gamma$$

avec $\gamma = 1$ dans le cas 1-classe et $\gamma = 0$ dans le cas 2-classes

L'algorithme SMGO [10] procède séquentiellement en optimisant la solution dans la direction de gradient maximal. Une étape d'optimisation s'exprime $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} + \lambda \mathbf{g}$ où λ est le pas d'optimisation et $\mathbf{g} = \mathbf{H}\boldsymbol{\alpha} + \boldsymbol{\delta}$ le gradient. Or, Tohmé montre que λ est déterminé à partir du gradient uniquement, et ce dernier est mis à jour suivant $\mathbf{g} \leftarrow \mathbf{g} - \lambda \mathbf{H}\mathbf{g}$. Cet algorithme est donc indépendant de $\boldsymbol{\delta}$; la différence entre l'optimisation de l'un ou l'autre des problèmes réside uniquement dans l'expression du gradient initial $\mathbf{g}_0 = \mathbf{H}\boldsymbol{\alpha}_0 + \boldsymbol{\delta}$ où $\boldsymbol{\alpha}_0$ est la solution initiale. Ainsi le même algorithme SMGO peut résoudre les problèmes WO-SVM (2-classes) et WOOC2-SVM (1-classe). Les détails de l'algorithmes sont donnés dans [10].

5 Evaluations

Dans cette section, nous décrivons des évaluations comparatives des algorithmes SMGO (sans biais) et Fast-OC2 (avec biais). Pour chaque expérience, on compare les résultats des deux algorithmes, avec et sans les contraintes du problème 2-classe (les résultats sont identifiés respectivement OC2 et OC).

Pour l'ensemble des expériences, on choisit un noyau Gaussien RBF dont le paramètre σ est estimé suivant [4] et un prétraitement est appliqué pour centrer-réduire les données.

Données de synthèse

Les données sont issues de deux distributions normales de variance 1 ; de moyenne $[0; 0]$ pour la classe $+1$ (données normales) et de moyenne $[3; 3]$ pour la classe -1 (données anormales). Les résultats sont moyennés sur 100 réalisations de chaque expérience.

Dans un premier temps, on s'intéresse à la vitesse de convergence lorsque la taille de l'ensemble d'apprentissage augmente. On génère 100 à 10000 observations pour la classe $+1$ et 100 pour la classe -1. Le paramètre ν est fixé à 10^{-2}. Dans un second temps, on s'intéresse à la vitesse de convergence lorsque le paramètre ν diminue. Cette étude relève d'un besoin opérationnel car ν est à relier au taux de fausse-alarme cible du détecteur d'anormalité [6]. Dans cette seconde expérience, le nombre d'observations de la classe $+1$ est fixé à 2000 échantillons et le paramètre ν varie de 2.10^{-1} à 2.10^{-3}. La figure 2 présente les résultats obtenus.

Données réelles

On utilise les bases de données Cancer, Crab, Glass, Iris et Wine du répertoire UCI [2]. La stratégie multi-classe appliquée est celle présentée dans [11] et les performances sont évaluées suivant une procédure *leave-one-out*. Enfin, ν est fixé à 10^{-3}.

Les taux de bonne classification sont présentés au tableau 1. Le tableau 2 rapporte lui les temps moyens de convergence et le nombre moyen de vecteurs de support. Ces résultats sont donnés pour l'ensemble des classes.

6 Conclusion

Les résultats montrent que l'algorithme SMGO est plus rapide que l'algorithme Fast-OC2. Le gain est particulièrement marqué pour des ensembles d'apprentissage de grande taille ou pour des faibles valeurs de ν. Cette conclusion s'applique indépendamment d'une approche avec ou sans les contraintes du problème 2-classes. L'analyse des résultats sur données réelles montre que les approches avec ou sans biais ont des performances comparables, sans pour autant concéder une augmentation du nombre de vecteurs de support.

Nous avons présenté un problème SVM 1-classe sans biais, exploitant les contraintes du problème 2-classes. A travers l'écri-

2. http ://archive.ics.uci.edu/ml/datasets.html.

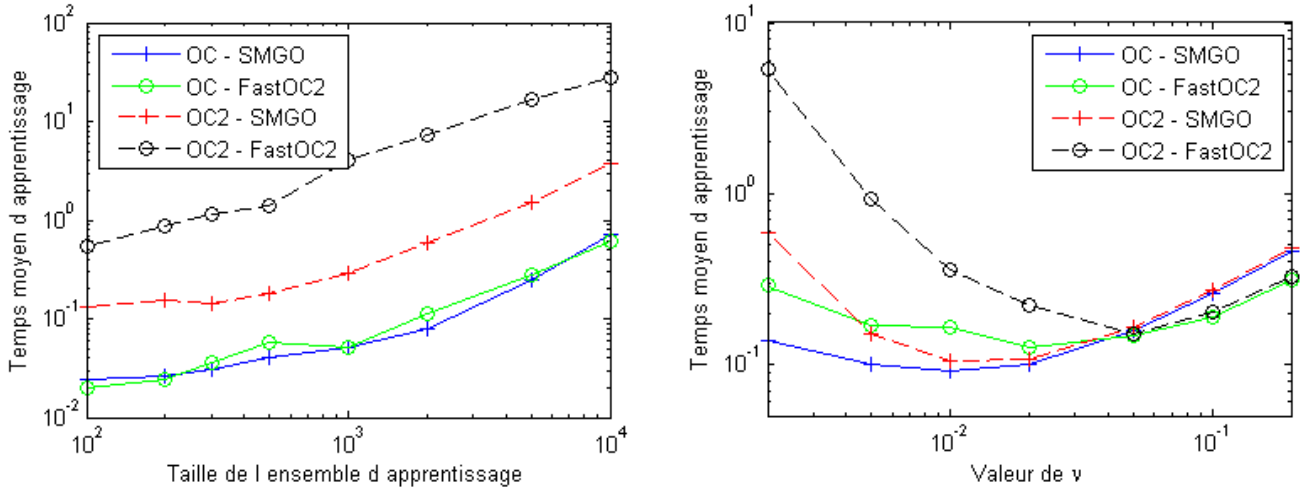

FIGURE 2 – Evolution du temps moyen de convergence en fonction de la taille de l'ensemble d'apprentissage (à gauche) et en fonction de la valeur de ν (à droite). Les échelles sont logarithmiques afin d'en clarifier la lecture.

ture d'une forme duale unifiée entre problème SVM 1-classe et 2-classe, nous avons montré qu'il est possible d'utiliser l'algorithme SMGO pour la résolution du problème introduit. On montre également que l'optimisation est indépendante du caractère 1 ou 2 classe(s). Ce résultat permet au problème SVM 1-classe avec contraintes du problème 2-classes de s'affranchir de l'algorithme Fast-OC2 et de bénéficier de l'ensemble des perspectives propres à l'algorithme SMGO présentées dans [10] et [8].

Bien que les résultats expérimentaux soit satisfaisants, l'analyse de la convergence du problème SVM 1-classe proposé devrait être faite. Par la suite, nous envisageons d'étudier les performances de notre approche dans le domaine de la surveillance audio pour la détection d'anormalités.

TABLE 1 – Probabilités de bonne classification (en %)

	OC		OC2	
Données	FastOC2	SMGO	FastOC2	SMGO
Cancer	96, 42	**96, 85**	95, 85	95, 85
Crab	85, 50	85, 00	**96, 50**	94, 00
Glass	78, 04	48, 60	**94, 86**	**94, 86**
Iris	89, 33	88, 00	94, 67	**95, 33**
Wine	92, 70	89, 33	**97, 19**	96, 63

TABLE 2 – Temps de convergence moyen (en ms) et nombre moyen de vecteurs de support (entre parenthèses)

	OC		OC2	
Données	FastOC2	SMGO	FastOC2	SMGO
Cancer	0, 7 (27)	0, 7 (28)	1, 8 (99)	1, 3 (98)
Crab	2, 6 (14)	2, 5 (26)	8, 9 (30)	3, 3 (34)
Glass	2, 4 (18)	2, 4 (18)	2, 5 (43)	2, 5 (44)
Iris	6, 7 (8)	6, 7 (10)	11, 3 (25)	8, 9 (24)
Wine	5, 6 (23)	5, 6 (25)	5, 6 (32)	5, 6 (34)

Références

[1] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[2] H. Kuhn and A. Tucker. Nonlinear programming. In U. of Calif. Press, editor, *Second Berkeley Symposium on Mathematics Stat. and Proba.*, pages 481–492, 1951.

[3] S. Lecomte, R. Lengellé, C. Richard, F. Capman, and B. Ravera. Abnormal events detection using unsupervised one-class svm - application to audio surveillance and evaluation. In *8th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2011.

[4] Y.-F. Li, I. W. Tsang, J. T. Kwok, and Z.-H. Zhou. Tighter and convex maximum margin clustering. *Journal of Machine Learning Research*, 5 :344–351, 2009.

[5] C. Liu, G. Wang, W. Ning, X. Lin, L. Li, and Z. Liu. Anomaly detection in surveillance video using motion direction statistics. In *IEEE International Conference on Image Processing (ICIP)*, 2010.

[6] B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comp.*, 13 :1443–1471, 2001.

[7] M. Schwabacher and N. Oza. Unsupervised anomaly detection for liquid-fueled rocket propulsion health monitoring. In *Journal of Aerospace Computing, Information, and Communication*, 2009.

[8] I. Steinwart, D. Hush, and C. Scovel. Training svms without offset. *Journal of Machine Learning Research*, 12 :141–202, 2011.

[9] A. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Doklady*, 4 :1035–1038, 1963.

[10] M. Tohmé and R. Lengellé. Sequential maximum gradient optimization for support vector detection. In *17th European Signal Proc. Conf. (EUSIPCO)*, 2009.

[11] M. Tohmé and R. Lengellé. Maximum margin one class support vector machines for multiclass problems. *Pattern Recognition Letters*, 32 :1652–1658, 2011.

[12] V. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, second edition edition, 1995.

[13] G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

[14] P. Winter, E. Hermann, and M. Zeilinger. Inductive intrusion detection in flow-based network data using one-class support vector machines. *IEEE Intl conférence on New Technologies, Mobility and Security*, 2011.