

1. Introduction

Les méthodes de classification ont pour but d'identifier les classes auxquelles appartiennent des objets à partir de certains paramètres descriptifs. Elles s'appliquent à un grand nombre d'activités humaines et conviennent en particulier au problème de la prise de décision automatisé.

Les méthodes utilisées pour la classification sont nombreuses, citons : la méthode des séparateurs à vaste marge (SVM), les réseaux de neurones, etc.

Nous présentons dans la suite de ce chapitre une étude détaillée une technique SVM. Ces méthodes ont montrés leurs efficacités dans de nombreux domaines d'application tels que le traitement d'images, la catégorisation de textes et le diagnostic médicale.

2. Les Séparateurs à Vaste Marge (SVM)

Les Machines à Vecteurs de Support ou Séparateur à Vaste Marge (SVM) sont des nouvelles techniques discriminantes dans la théorie de l'apprentissage statistique. Elles ont été proposées en 1995 par V. Vapnik [45] Elles permettent d'aborder plusieurs problèmes divers et variés comme la régression, la classification, la fusion etc. Il n'est plus à démontrer que la seule minimisation du risque empirique (l'erreur d'apprentissage) ne garantit pas une faible erreur sur un corpus de test. Les SVM fournissent une approche très intéressante de l'approximation statistique. Ces nouvelles techniques unifient deux théories : Minimisation du risque empirique et Capacité d'apprentissage d'une famille de fonctions. C'est la Minimisation du Risque Structurel.

2.1. Historique

L'émergence des SVMs, a commencé autour des débuts des années 1990s, néanmoins d'autres travaux et recherches sur l'apprentissage machine par les mathématiciens russes Vladimir Vapnik et Alex Chervonenkis ont fortement contribué à leur apparition, notamment la 1ère description de l'implémentation d'un modèle proche a un SVM apparu dans la traduction en Anglais en 1982 de l'ouvrage de Vapnik, «Estimation of Dépendances Based on Empirical Data », (édité en 1er lieu en russe en 1979), que l'exploration de la notion d'hyper plan a marge maximale l'a précédé.

Le modèle initial à marge maximale a connu des extensions importantes en 1992 qui ont formé le modèle final par l'utilisation de la Kernel trick d'Aizeman proposé par Boser, Guyon & Vapnik, présenté dans un article à la conférence COLT 92, finalement les SVMs sous leur forme actuelle ont été introduits en 1995 par V.Vapnik & C.Cortes après l'introduction du « soft margin ».

Les limites statistiques des SVMs sont apparues en 1998 par Barlett & Shawe-Taylor sur la généralisation des SVM à marge dure (hard margin), suivie en 2000 par une autre critique montrant les limites de la généralisation des algorithmes à marge souple (soft margin) par Shawe-Taylor et Cristianini [17].

2.2. Définition

Les "*Support Vector Machines*" appelés aussi "maximum margin classifier" (en français machine à vecteurs de support ou séparateur à vaste marge) sont des techniques d'apprentissage supervisé basées sur la théorie de l'apprentissage statistique (généralement considérés comme la 1^{ère} réalisation pratique de cette théorie [43]) et respectant les principes du "structural risk minimization" (SRM) (trouver un séparateur qui minimise la somme de l'erreur de l'apprentissage [27]), un SVM repose sur les 2 notions de vaste marge et fonction noyau.

Les SVMs sont considérés comme l'un des modèles les plus importants parmi la famille des méthodes de l'intelligence artificielle. Ils ont gagné une forte popularité grâce à leur succès dans la reconnaissance des chiffres manuscrits avec un taux d'erreur de 1.1% en phase de test (le même taux marqué par un réseau de neurone soigneusement construit) [6]

2.3. Domaines d'application

Les SVM peuvent être utilisés soit pour classer ou pour présumer des formes arbitraires à partir d'un ensemble de données étiquetées. Au cours de la dernière décennie, de nombreux problèmes de reconnaissance de formes ont été traités en utilisant les machines à vecteurs de support. Par exemple Cortes et al., Schölkopf et al. et Burges et al. ont appliqué les SVM pour la reconnaissance optique de caractères [54], alors que Blanz et al. les ont utilisés pour reconnaître des scènes de deux points de vue d'objets tridimensionnels. Ainsi, Schmidt et al. ont utilisé ce classifieur en tant que partie d'un système d'identification du locuteur [42] et de Osuna et al. ont étudié leurs

performances sur une tâche de reconnaissance de visage [35]. Beaucoup d'autres applications telles que la classification par sexe, le data mining, etc., ont été traitées à l'aide de SVM [4].

2.4. Principe des SVMs

Notions de base: Hyperplan, Marge, Vecteurs de support

Pour deux classes d'exemples données, le but de SVM est de trouver un classifieur qui va séparer les données et maximiser la distance entre ces deux classes. Avec SVM, ce classifieur linéaire est appelé hyperplan. Les points les plus proches, qui seuls sont utilisés pour la détermination de l'hyperplan, sont appelés vecteurs de support.

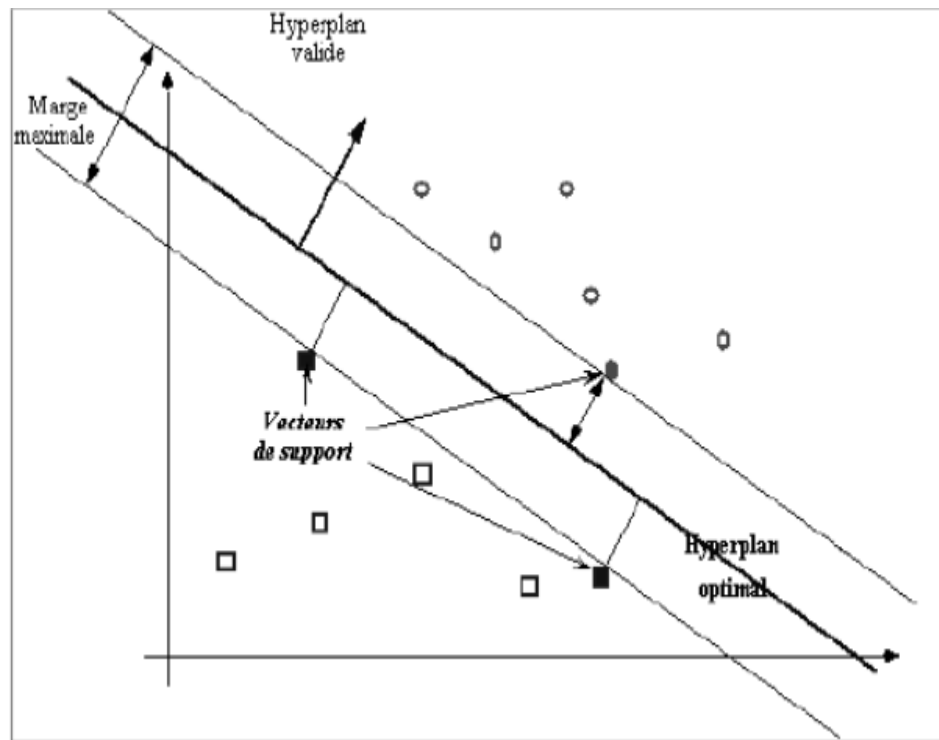


Figure 3.1-Hyperplan optimal, marge maximale et vecteurs de support.

Il est évident qu'il existe une multitude d'hyperplans valides mais la propriété remarquable des SVM est que cet hyperplan doit être optimal. Formellement, cela revient à chercher un hyperplan dont la distance minimale aux exemples d'apprentissage est maximale. On appelle cette distance « marge » entre l'hyperplan et les exemples. L'hyperplan séparateur optimal est celui qui maximise la marge (Figure 3.1) [46].

2.4.1. Linéarité et non linéarité

Dans la plupart des problèmes réels il n'y a pas de séparation linéaire possible entre les données, la notion de marge maximale ne peut pas être utilisée car elle fonctionne seulement si les classes de données d'apprentissage sont linéairement séparables.

Dans le cas non linéaire, le principe consiste à projeter les données de l'espace d'entrée non linéairement séparables dans un espace de plus grande dimension appelé espace de caractéristiques de façon à ce que les données deviennent linéairement séparables. Ceci est illustré par le schéma suivant:

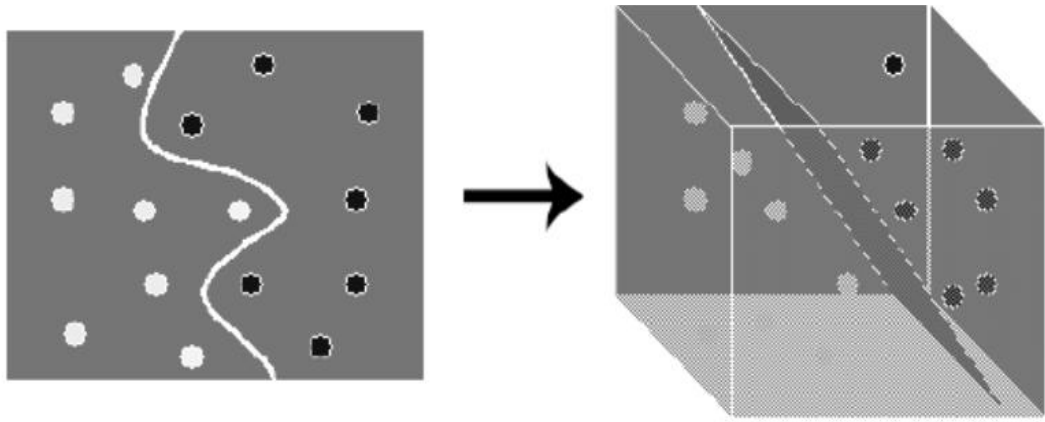


FIG3.2- Hyperplan séparateur dans le cas non linéairement séparable [46].

2.5. Fondements mathématiques

La méthode générale de construction de l'Hyperplan Optimal (HO) qui sépare des données appartenant à deux classes différentes linéairement séparables est comme suit : Soit $H : (w \cdot x) + b$ l'hyperplan qui satisfait la condition suivante :

$$Y_t(w \cdot x_t + b) \geq 1 \quad \text{pour } t=1, \dots, m \quad (3.1)$$

Trouver l'hyperplan optimal revient à maximiser la marge $m = 2/\|w\|$. Ce qui est équivalent à minimiser $\frac{\|w\|^2}{2}$ sous la contrainte (1). Ceci est un problème de minimisation d'une fonction objective quadratique avec contraintes linéaires.

$$\begin{cases} \min \frac{1}{2} \|w\|^2 \\ \forall t, Y_t(w \cdot x_t + b) \geq 1 \end{cases} \quad (3.2)$$

En appliquant le principe de Lagrange, on obtient le problème de programmation quadratique de dimension m (nombre d'exemples) suivant :

$$\left\{ \begin{array}{l} \max \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j \gamma_i \gamma_j (X_i \cdot X_j) \\ \forall i \alpha_i \geq 0 \\ \sum_{i=1}^m \alpha_i \gamma_i \end{array} \right. \quad (3.3)$$

Les α_i sont les coefficients de Lagrange.

Définition

On définit les Vecteurs Supports **VS** tout vecteur **xi** tel que **yi [(wo · xi) + bo] =**

1. Ce qui est équivalent à :

$$\mathbf{VS} = \{ \mathbf{X}_t \mid \alpha_t > 0 \} \text{ pour } t=1, \dots, m$$

La fonction de classement class(x) est défini par :

$$\begin{aligned} \mathbf{Class}(\mathbf{x}) &= \text{sign} [(\mathbf{w}_0 \cdot \mathbf{x}) + \mathbf{b}_0] \\ &= \text{sign} [\sum_{\mathbf{X}_t \in \mathbf{VS}} \alpha_t \gamma_t (\mathbf{x}_t \cdot \mathbf{x}) + \mathbf{b}_0] \end{aligned}$$

Si class(x) est inférieure à 0, x est de la classe -1 sinon il est de la classe 1. Dans le cas linéaire, on pouvait transformer les données dans un espace où la classification serait plus aisée. Dans ce cas, l'espace de caractéristiques utilisé le plus souvent est R (ensemble des nombres réels). Il se trouve que pour des cas non linéaires, cet espace ne suffit pas pour classer les entrées. On passe donc dans un espace de plus grande dimension [46].

$$\begin{aligned}
 \phi: R^d &\rightarrow F \\
 x &\rightarrow \phi(x) \\
 \phi: R^d &\rightarrow \phi: R^d \xrightarrow{x} \phi
 \end{aligned}$$

Avec $\text{card}(F) > d$

On doit donc résoudre

$$\begin{cases} \max \sum_{t=1}^n \alpha_t - \frac{1}{2} \sum_{t,f} \alpha_t \alpha_f y_t y_f \phi(x_t), \phi(x_f) \\ \forall t, 0 \leq \alpha_t \leq 2 \\ \sum_{t=1}^n \alpha_t y_t = 0 \end{cases} \quad (3.4)$$

Plutôt que de choisir la transformation non-linéaire $\Phi : X \rightarrow F$, on choisit une fonction $k: X \times X \rightarrow R$ (nombres réels) appelée fonction noyau.

❖ Exemples de Fonctions noyau les plus utilisées

La fonction noyau doit respecter certaines conditions, elle doit correspondre à un produit scalaire dans un espace à grand dimension. Le théorème de Mercer définit les conditions que K doit satisfaire pour être une la fonction noyau : elle doit symétrique et semi-défini positive. Il existe plusieurs formes utilisées de la fonction noyau :

➤ Linéaire

L'approche kernel trick généralise l'approche linéaire en faisant un cas particulier.

$$K(x, x') = x \cdot x' \quad (3.5)$$

➤ Polynomial

$$K(x, x') = (\langle x, x' \rangle + 1)^d \quad (3.6)$$

Où d est la degré de polynomial [7art].

➤ Fonction quadratique (quad)

$$K(x, x') = (\langle x, x' \rangle + 1)^2 \quad (3.7)$$

➤ Gaussien RBF (Radial Basis Function)

$$K(x, x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right) \quad (3.8)$$

Où σ est le paramètre de contrôle de la marge de Kernel [49].

➤ **Multi-Layer Perceptron (MLP)**

$$K(x, x') = \tanh(\rho \langle x, x' \rangle + \varrho) \quad (3.9)$$

Pour certaines valeurs de paramètre du produit scalaire, ρ , et le paramètre de offset, ϱ

2.6. SVM multi-classes

A l'origine, les SVM ont été conçus essentiellement pour les problèmes à 2 classes, cependant plusieurs approches permettant d'étendre cet algorithme aux cas à N classes ont été proposées. La généralisation dans le cas multi-classes peut se faire de trois façons différentes. Les deux premières méthodes sont basées sur une multiplication des classifieurs bi-classes tandis que la dernière propose une résolution globale.

➤ **Un-contre-tous:** l'approche la plus naturelle est d'utiliser cette méthode de discrimination binaire et d'apprendre N fonctions de décision $\{f_m\}_{m=1 \dots N}$ permettant de faire la discrimination entre chaque classe de toutes les autres (chaque classe est opposée à toutes les autres).il faut donc poser N problèmes binaires. L'affectation d'un nouveau point x à une classe C_i se fait par la relation [46] :

$$i = \underset{m=1 \dots N}{\operatorname{argmax}} F_m(x) \quad (3.10)$$

➤ **Un-contre-un:** la deuxième méthode est une méthode dite d'un contre un. Au lieu d'apprendre N fonctions de décisions, ici chaque classe est discriminée d'une autre. Ainsi, $N(N-1)/2$ fonctions de décisions sont apprises et chacune d'entre elles effectue un vote pour l'affectation d'un nouveau point x . La classe de ce point x devient ensuite la classe majoritaire après le vote [46].

➤ **Méthode globale:** la dernière méthode est une approche étendant la notion de marge aux cas multi-classes. Le problème fait intervenir N fonctions de décision et il est très gourmand en temps de calcul et en espace mémoire ce qui fait qu'il reste peu utilisé dans les cas réels [46].

3. Avantages et inconvénients

❖ Avantages

- Les SVM possèdent des fondements mathématiques solides.
- Les exemples de test sont comparés juste avec les supports vecteur et non pas avec les exemples d'apprentissage.
- Décision rapide, la classification d'un nouvel exemple consiste à voir le signe de la fonction de décision $f(x)$.

❖ Inconvénients

- Classification binaire d'où la nécessité d'utiliser l'approche un –contre-un.
- Grand quantité d'exemple en entrées implique un calcul matériel important.
- Temps de calcul élevé alors d'une régularisation des paramètres de la fonction noyau.

4. Conclusion

Dans ce chapitre nous avons présentées les bases théoriques des machines à vecteurs de supports (SVM). Le chapitre suivant est dédié aux expériences et résultats que nous avons effectués.