Comparaison de stratégies de discrimination de masses de véhicules automobiles

Alain Rakotomamonjy * — Rodolphe Le Riche ** — David Gualandris * — Stéphane Canu***

PSI, INSA Rouen, 76801 St. Etienne du R^{ay} cedex, France
*** CNRS URA 1884 / SMS, Ecole des Mines de St. Etienne, 42000 St. Etienne, France*
**** PSA, Dir. de la Recherche et de l'Innovation Automobile, Bièvres, France*

alain.rakoto@insa-rouen.fr, leriche@emse.fr, david.gualandris@mpsa.com,stephane.canu@insa-rouen.fr

RÉSUMÉ. *L'industrie automobile utilise aujourd'hui les enquêtes en clientèle, notamment, pour le dimensionnement en fiabilité de ses composants. Le problème type du classement par masses de véhicules à partir de mesures d'accélérations et de vitesses est traité. Une méthode de discrimination optimale pour ce problème est construite en considérant 4 niveaux : choix de l'espace de recherche (sélection de variables par algorithmes évolutionnaires ou par heuristiques, analyse discriminante), choix du critère de discrimination (critère Bayésien ou de marge large), choix de la complexité et choix des paramètres des algorithmes.*

ABSTRACT. *Customer surveys have recently been used in the automotive industry in order to statistically characterize loadings. The objective is reliable design of car components. The standard problem of classifying cars based on acceleration and speed measures is addressed here. An optimal discrimination strategy is methodically devised by considering 4 decision levels: the choice of the search space (variables selection through heuristics or evolutionary algorithms, discriminant analysis), the choice of the discrimination criterion (probabilistic or maximum margin), the choice of the classification algorithm complexity and the tuning of other algorithm parameters.*

MOTS-CLÉS : *identification en dynamique, sélection de variables, analyse discriminante, machines à support vectoriel, classification, automobile, méthode inverse en dynamique.*

KEYWORDS: *Identification, variable selection, discriminant analysis, support vector machines classification, inverse methods, automotive industry*

1. Introduction

Dans le cadre d'une approche probabiliste du dimensionnement, l'industrie automobile a aujourd'hui recourt à des enquètes en clientèles ([THO 01]). Le problème type considéré est celui du classement de véhicules par masses à partir de mesures d'accélérations et de vitesses. Dans un cadre de modélisation classique de dynamique des structures, ce problème est mal posé car une condition limite, la route empruntée par le client, est inconnue. C'est pourquoi les mesures recueillies sont exploitées par des techniques de fouille de données.

Des problèmes mal posés d'identification en dynamique linéaire ont été analysés par Ljung [LJU 87]. Dans [AST 76], l'identification des systèmes dynamiques linéaires est appliquée à un problème de structures. Depuis les travaux de Chen et al. et Levin ([CHE 90, LEV 92]), les réseaux de neurones ont servi à l'identification de systèmes dynamiques non linéaires, dont des structures mécaniques (e.g., [YEN 94, TAK 99]). Le problème de l'identification de la masse d'un véhicule dans un cadre de régression neuronale a été étudié dans [LER 01].

Conscient qu'il n'existe pas de méthode généraliste optimale pour tous les problèmes (théorèmes de "No Free Lunch", [WOL 97, DUD 01]), ce travail procède à la construction méthodique d'un algorithme de discrimination autour du problème type de classement de véhicules par masses.

L'article commence par exposer le problème de la discrimination par masses et ses motivations industrielles.

Puis, les quatre niveaux d'optimisations propres à la majorité des problèmes de fouilles de données sont discutés : choix de l'espace des variables explicatives, choix d'un critère de classement (i.e., d'un principe d'induction), choix d'un niveau de régularisation et optimisation des paramètres de la méthode. Une grande diversité d'approches est prise en compte de manière couplée. En ce qui concerne les variables explicatives, on considère la sélection de variables par heuristiques ou par algorithme évolutionnaire et la combinaison linéaire de variables par analyse discriminante. Au niveau des méthodes de discrimination, des approches par maximisation de marges entre classes (machines à supports vectoriels) ainsi que des approches par minimisation d'erreur de classement (fenêtres de Parzen, discrimination postérieure à une régression) sont mises en œuvre. Enfin, sur le plan des algorithmes d'optimisation, ce travail implique des méthodes de minimisation de moindres carrés (Levenberg-Marquardt), des méthodes de programmation quadratique, des méthodes évolutionnaires, ainsi que des heuristiques énumératives.

Des expériences numériques à partir de signaux réels sont finalement rapportées qui permettent la comparaison de plusieurs méthodes de discrimination. La méthode optimale est une combinaison d'analyse discriminante et de machines à supports vectoriels en classification. L'ensemble permet de déterminer la classe de masse avec plus de 80% de probabilité de succès.

Notations

– Notations sur les ensembles :

$\mathcal{E}$, ensemble des données = $\mathcal{L} \cup \mathcal{T}$ = $\{(x^{(i)}, y^{(i)})\}, i = 1, \text{card}\,\mathcal{E}$.

$\mathcal{L}$, ensemble d'apprentissage total = $\mathcal{A} \cup \mathcal{V}$. card $\mathcal{L}$ est le nombre total de données utilisées pour l'apprentissage.

$\mathcal{A}$, ensemble des données d'apprentissage.

$\mathcal{V}$, ensemble des données de validation.

$\mathcal{T}$, ensemble des données de test.

$\mathcal{M}_i$, partie des données $(x, y) \in \mathcal{E}$ de masse $m_i, i = 1, 3$.

$\mathcal{X}$, espace des variables d'entrées x.

– Autres notations :

$\widehat{\ }$, symbole d'estimation.

a, vecteur des signaux mesurés (accélérations ou vitesses de rotation des roues), a_i étant une mesure à un instant.

a_{min}, a_{max}, moyenne des minima / maxima des signaux sur un intervalle.

C, paramètre de régularisation des MSVs.

d, dimension de l'espace des variables d'entrées.

D, nombre de mesures a_i dans un intervalle de temps.

ϵ, largeur du tube en ϵ-SV régression.

E, énergie du signal sur un intervalle.

$f(x; w)$, fonction dont les variables d'entrées sont x et les paramètres w.

h, largeur des noyaux Gaussiens des fenêtres de Parzen.

J_2, norme quadratique.

$J_{2\nu}$, norme quadratique régularisée, avec comme paramètre de régularisation ν.

J_ϵ, norme ϵ-insensible.

$J_{\epsilon\nu}$, norme ϵ-insensible régularisée.

moy, moyenne des signaux sur un intervalle D.

M_i, i-ième moment relatif de la densité spectrale de puissance.

MSV, machines à support vectoriel.

ν, paramètre de régularisation.

N_{bases}, nombre total de bases d'apprentissage ou de validation générées par validation croisée.

N_i, nombre total d'intervalles de durée D chacun.

N_{NC}, nombre de neurones cachés.

o, degré des polynômes dans les noyaux de MSV.

$P(u)$, probabilité de l'évènement u.

$p(u)$, densité de probabilité en u.

σ, largeur des noyaux gaussiens.

$skew$, 3-ième moment statistique (skewness)

$kurt$, 4-ième moment statistique (kurtosis) des signaux sur un intervalle D.

var, 2-ième moment statistique des signaux sur un intervalle D.

w, paramètres de fonctions choisis lors de l'apprentissage.

x, variables d'entrées, $x \in \mathcal{X} \subseteq \mathbb{R}^d$. Dans un ensemble de données, $x^{(i)}$ est le i-ième vecteur de variables d'entrées, $x_j^{(i)}$ étant sa j-ième composante.

y, variables de sorties à estimer. Ici, $y \in \{m_1, m_2, m_3\}$, les m_i étant des masses (scalaires).

2. Extraction de l'information dans l'industrie automobile

Pour assurer la fiabilité de ses véhicules, le groupe PSA s'appuie sur une approche probabiliste du dimensionnement. Cette démarche repose sur la connaissance de deux distributions : la distribution des efforts, qui traduit la sévérité du client, et la distribution de résistance des composants qui traduit la capacité d'une structure à résister à la contrainte imposée par le client. L'information introduite dans la première distribution est notamment issue d'enquêtes en clientèle.

Pour améliorer les connaissances extraites de ces enquêtes il est aujourd'hui indispensable d'accompagner l'analyse des mesures par des méthodes de fouille de données. Les objectifs industriels de cette démarche sont d'accroître la quantité d'informations accessibles pour une campagne de mesure donnée, et, réciproquement, de réduire le coût des campagnes de mesures (par exemple en limitant le nombre de voies de mesures).

2.1. *Identification de la masse*

L'objectif de ce travail est d'identifier la masse d'un véhicule roulant sur une route inconnue à partir de mesures d'accélérations en différents points et de vitesse longitudinale. La Figure 1 montre une photographie d'un accéléromètre servant aux mesures. La résolution de ce problème par inversion d'un modèle structural déterministe, par exemple un modèle éléments finis, n'est pas envisagée d'une part parce que les suspensions du véhicule sont non-linéaires, d'autre part parce que les conditions limites (la route et le conducteur) sont inconnues. L'approche utilisée, au contraire, construit directement le modèle inverse liant les signaux mesurés à la masse à partir de données expérimentales. Des méthodes d'apprentissage supervisé généralistes servent à déter-

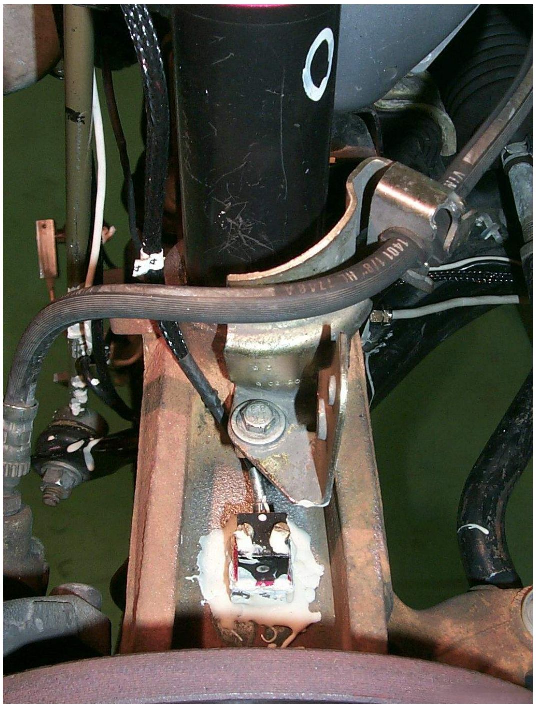


Figure 1. *Dispositif de mesure de l'accélération.*

miner le modèle inverse. On peut considérer, avec cette approche, qu'à la causalité du modèle structural est substituée une meilleure validité statistique (cf. [LER 01]).

Pour des raisons de confidentialité, les unités de temps (fréquences, temps des intervalles) et de masses sont normalisées. La base expérimentale disponible est constituée d'enregistrements de signaux, obtenus lors du parcours d'un circuit routier par trois véhicules de masses différentes, m_1, m_2, et m_3. Les signaux mesurés sont

– les 4 accélérations verticales aux pivots de roues,

– les 3 accélérations verticale, transversale et longitudinale au centre de gravité du véhicule,

– la vitesse longitudinale du véhicule.

Ces signaux subissent, avant leur stockage et notre traitement, un contrôle de qualité ayant pour but de supprimer les artefacts dûs par exemple à un claquage de porte ou à une dérive du capteur. Ils comportent un très grand nombre de mesures instantanées. Un modèle inverse prenant directement chaque mesure d'accélérations en entrée aurait autant de paramètres à régler que de mesures, ce qui rend son identification mal posée. Il est donc nécessaire de pré-traiter les accélérations en définissant de nouvelles variables explicatives. Ces variables sont utilisés comme entrées du modèle inverse. Le rôle de l'apprentissage est d'estimer la dépendance entre les variables explicatives et la masse du véhicule.

Dans une étude précédente ([LER 01]), nous nous étions intéressé au problème de l'estimation de la masse par régression. Cependant, l'exploitation industrielle de l'estimation de masse implique des protocoles discrets. C'est pourquoi l'estimation de la masse est maintenant formulée comme une classification en trois familles, les véhicules moyennement lourds, les véhicules lourds, et les véhicules très lourds.

Globalement, la méthodologie suivie est illustrée sur la Figure 2 . Les signaux mesurés sont segmentées en intervalles de durée D. Sur chaque intervalle, les variables d'entrées sont calculées, puis utilisées pour identifier la masse.

Cependant, les accélérations mesurées en clientèle permettent en fait de générer plusieurs intervalles. On peut donc cumuler les classifications obtenues à partir de chaque intervalle par la règle du vote majoritaire : la masse finale associée au véhicule est la masse la plus représentée dans les N_i intervalles. Soit p, la probabilité de correctement classer un véhicule à partir des variables extraites d'un intervalle. Deux hypothèses, l'une optimiste et l'autre pessimiste servent à estimer une probabilité de succès d'un vote majoritaire sur les N_i intervalles. L'hypothèse optimiste est que l'affectation d'un vecteur d'entrée issu d'un intervalle à une classe de masse est une variable aléatoire indépendante . L'hypothèse pessimiste est de considérer la probabilité $P(s \geq N_i/2)$ qu'au moins la moitié des vecteurs issus des N_i intervalles soient correctement classés comme une estimation de la probabilité de succès par vote majoritaire sur N_i intervalles. En fait, $P(s \geq N_i/2)$ est une borne inférieure de la probabilité de succès cumulée, car quelle que soit la répartition des classements pour les deux autres classes, la classe considérée est bien classée par la règle de cumul. $P(s \geq N_i/2)$ s'écrit

$$P(s \geq N_i/2) = \sum_{k=N_i/2}^{N_i} \binom{k}{N_i} p^k (1-p)^{N_i-k} \qquad (1)$$

La Figure 3 montre que, pour p supérieur à 0.5, la probabilité de succès cumulée croit rapidement. Par la suite, on ne s'interesse plus qu'à la capacité d'un classifieur à classer correctement un vecteur d'entrée issu d'un seul intervalle et donc à une estimation de p.

2.2. *Les variables d'entrées*

La base expérimentale est composée, pour chaque masse, de N_i intervalles de durée D. Chaque segment de signal est traité pour générer des variables d'entrée. Avant tout traitement des signaux d'accélération (mais pas de vitesse), pour supprimer la dérive présente dans les mesures, la moyenne est soustraite de chaque segment. Les variables d'entrées sont composées en partie des paramètres temporels suivant :

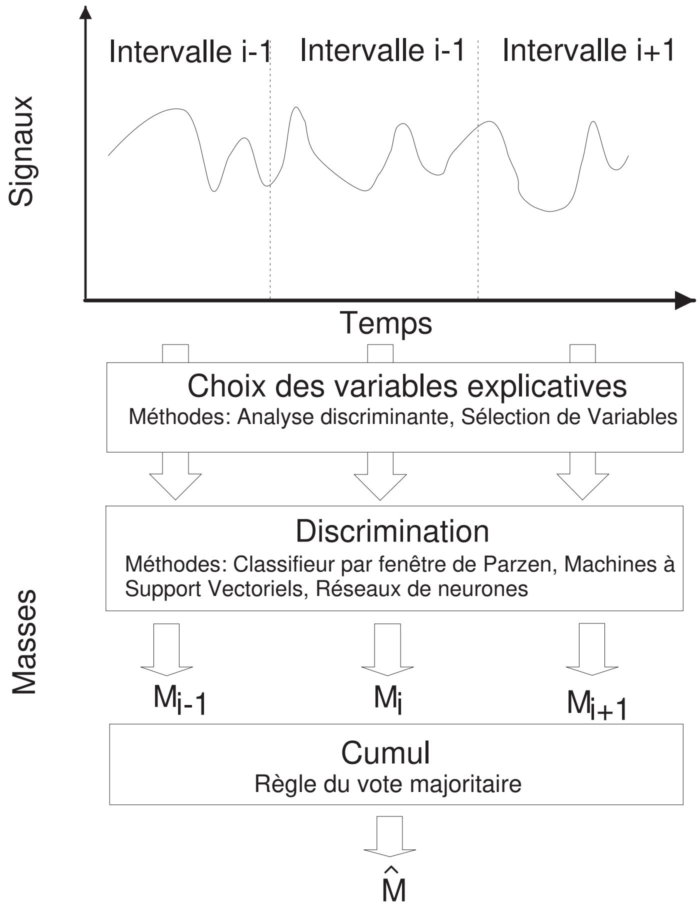


Figure 2. *Schéma de principe de la discrimination par masses à partir de signaux.*

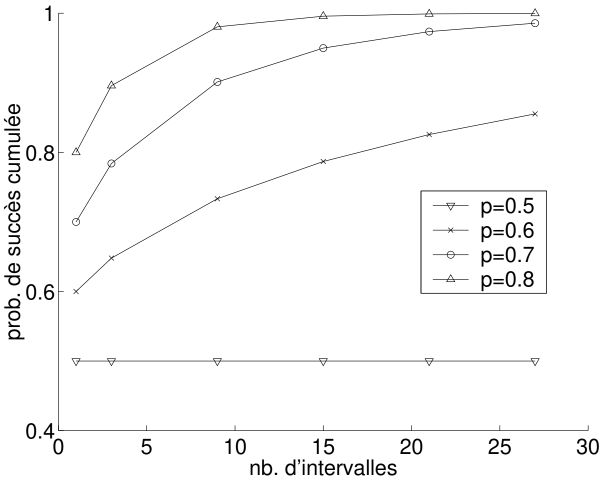


Figure 3. *Estimation de la probabilité de succès cumulée $P(s \geq N_i/2)$ en fonction du nombre d'intervalles N_i pour différents p, probabilités de succès par intervalle.*

– la moyenne (pour la vitesse), la variance, le skewness, et le kurtosis de la distribution des signaux,

$$moy,\ var,\ skew,\ kurt\ =\ \frac{1}{D}\sum_{i=1}^{D} a_i^j \qquad j = 1,4\ ,\ \text{respectivement.} \tag{2}$$

– la moyenne des minima et maxima des signaux, a_{min} et a_{max}, chaque extremum étant relatif à un horizon de $\pm 1\ sec$.

L'autre partie des variables d'entrées est composée de paramètres spectraux liés à la densité spectrale de puissance (DSP) du signal. La pertinence de la DSP est immédiatement visible pour les systèmes dynamiques linéaires. Si les suspensions d'un véhicule étaient linéaires, la DSP de la hauteur de la route à la fréquence f, $S_r(f)$, s'exprimerait en fonction de la DSP d'une accélération, $S_a(f)$, par

$$S_a(f) = |H(f)|^2 S_r(f) \tag{3}$$

La masse a une influence directe sur la fonction de transfert $H(f)$. Cet effet, bien que parasité par la DSP de l'échantillon de route, se retrouve dans la DSP de l'accélération $S_a(f)$.

La densité spectrale de puissance est obtenue par la méthode non paramétrique classique des périodogrammes moyennés [MAT 00, SHI 99]. Les variables d'entrées spectrales sont fonctions de la DSP dans la plage de fréquences F, où se concentre l'énergie des signaux (cf. Figure 4). En particulier, on considère les variables suivantes :

– L'energie du signal sur la plage de fréquence F,

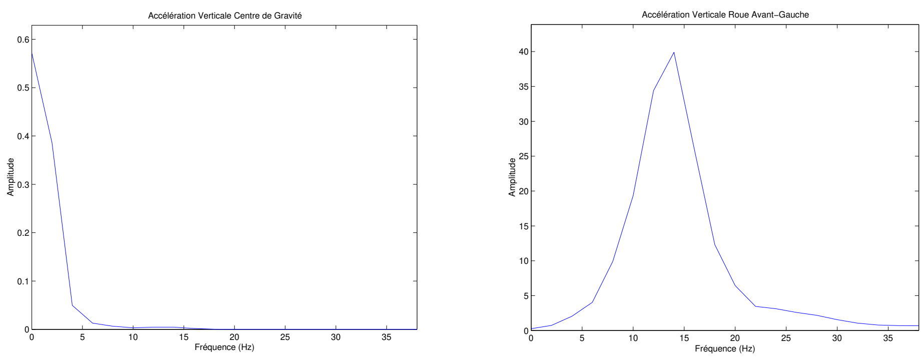
$$E = \sum_{i|f_i \in F} DSP(f_i)^2 \qquad (4)$$

– Les moments relatifs d'ordre j (j allant de 1 à 4) de la DSP sur la plage de fréquence F,

$$M_j = \frac{1}{E} \sum_{i|f_i \in F} f_i^j DSP(f_i) \qquad (5)$$

Ces moments visent à caractériser la répartition d'énergie dans les fréquences indépendamment du niveau global d'énergie, E.

Ainsi, pour chaque intervalle, 10 variables d'entrées temporelles et spectrales sont calculées pour chacunes des 7 accélérations. Onze variables sont extraites de la vitesse de rotation des roues (la moyenne est conservée). Le problème de classification est donc de dimensions $d = 81$. Il y a 3 classes de masses et N_i intervalles pour chaque masse, soit card $\mathcal{E} = 3N_i$ données. Formellement, on a un ensemble $\{(x^{(i)}, y_i)\}_{i=1 \cdots \text{card}\, \mathcal{E}} \in \mathcal{X} \times \{m_1, m_2, m_3\}$, $\mathcal{X} \subseteq \mathbb{R}^{81}$, où un couple $(x^{(i)}, y_i)$ représente les variables d'entrées dans l'intervalle i et la classe de masse. Dans la suite, notre objectif est d'inférer la classe de masse d'un véhicule à partir des variables d'entrées extraites d'un intervalle de D points.



(a) (b)

Figure 4. *Exemples de DSP de l'accélération verticale au centre de gravité (a) et à la roue avant gauche (b).*

2.3. *Des optimisations à quatre niveaux*

La problématique générale de ce travail est la détermination du meilleur jeu [variables d'entrées / classifieur] pour estimer la classe de masse. Un schéma synoptique de la démarche suivie a été donné en Figure 2.

Quatre niveaux de choix optimaux interviennent dans la construction de la méthode de discrimination : le choix des variables explicatives, le choix d'un critère de discrimination, le choix d'un degré de régularisation, et enfin le choix des paramètres (autres que ceux de régularisation) de chaque méthode en fonction des données.

Les stratégies utilisées ici pour choisir à ces quatre niveaux diffèrent. Lorsque le champ des possibilité est grand et que la qualité d'un choix est mesurable rapidement, on a recourt à des algorithmes d'optimisation numérique. C'est le cas du réglage des paramètres des classeurs (e.g., programmation quadratique pour les machines à supports vectoriels, cf. paragraphe 4.2). La sélection de variables peut également bénéficier d'algorithmes d'optimisation (algorithme évolutionnaire, heuristiques), en particulier si la qualité des solutions proposées est évaluée au moyen du critère Λ de Wilks (cf. paragraphe 3.2). Le choix du critère de discrimination est restreint puisque, essentiellement, on compare dans la section 4 un critère de marge maximale et un critère Bayésien. Il suffit alors de comparer les différentes méthodes associées. Enfin, le choix d'un niveau de régularisation, qui implique typiquement un paramètre scalaire et une fonction objectif coûteuse (au moins un apprentissage complet de classeur), sera géré par énumération de quelques possibilités.

3. Choix de l'espace de recherche

La première étape de tout problème d'optimisation, qui est l'angle par lequel nous considérons la discrimination par masses de véhicules, est le choix des variables d'optimisation (ou variables explicatives). Ce choix est crucial dans la mesure où il définit l'espace des possibilités à explorer, indépendamment de la méthode d'exploration. Le choix des variables explicatives est abordé au moyen de l'analyse discriminante et de la sélection de variables, les deux techniques ayant pour objet de réduire la dimension de l'espace de recherche. Il doit cependant être noté que certaines méthodes de discrimination, comme les machines à supports vectoriels ou la discrimination après régression, procèdent implicitement à des changements de variables qui sont également interprétables comme des choix d'espaces de recherche (cf. paragraphes 4.2 et 4.3.2, respectivement).

3.1. *Analyse discriminante*

Le but de l'analyse discriminante est de trouver un hyperplan de dimension p, $p < d$, qui permette de séparer au mieux les k classes des données après projection.

Les vecteurs supportant cet hyperplan définissent p nouvelles variables explicatives combinaisons linéaires des d variables d'entrées d'origine.

Supposons l'échantillon $\{x^{(1)}, \ldots, x^{(n)}\}$ associé à la partition $\{\mathcal{M}_1, \ldots, \mathcal{M}_k\}$ où chaque $\mathcal{M}_i$ est constitué de n_i individus. On appelle, respectivement, les matrices de variance totale, interclasse et intraclasse, les matrice T, B et W. Si les individus de l'échantillon ont le même poids, ces différentes matrices s'écrivent

$$g_j = \frac{1}{n_j} \sum_{i \in \mathcal{M}_j} x^{(i)} \quad , \quad g = \frac{1}{n} \sum_{i=1}^{n} x^{(i)} \; = \; \frac{1}{n} \sum_{j=1}^{k} n_j g_j \; , \tag{6}$$

$$V_j = \frac{1}{n_j} \sum_{i \in \mathcal{M}_j} (x^{(i)} - g_j)(x^{(i)} - g_j)^T \; , \tag{7}$$

$$T \; = \; \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - g)(x^{(i)} - g)^T \; = \; B + W \; , \tag{8}$$

$$\text{avec} \quad B \; = \; \sum_{j=1}^{k} \frac{n_j}{n} (g_j - g)(g_j - g)^T \tag{9}$$

$$\text{et} \quad W \; = \; \sum_{j=1}^{k} \frac{n_j}{n} V_j \; . \tag{10}$$

On note que, puisque les $g_j - g$ de l'Equation (9) sont linéairement dépendants, B est une matrice de rang, au maximum, $k - 1$. Soit a, un des axes discriminants supportant l'hyperplan, et M une métrique associé à $\mathbb{R}^d$. Pour éviter que les coordonnées de plus grande variance dominent l'analyse discriminante, la métrique choisie est $M \; = \; T^{-1}$, ce qui normalise les données par leur variance totale. De manière intuitive, le problème consiste à trouver un vecteur a, tel que, après projection, les nuages des différentes partitions soient aussi séparés que possible, tandis que chaque partition reste groupée autour de son centre de gravité. Cela revient à chercher un axe a sur lequel l'inertie de la projection des centres de gravités g_j est maximale et l'inertie de chaque partition minimale. Par conséquent, on cherche à maximiser $a^t MBMa$ tout en minimisant $a^t MWMa$. Comme $a^t MTMa = a^t MBMa + a^t MWMa$, un critère dont la maximisation synthétise ces souhaits est

$$J_D(a) \; = \; \frac{a^t MBMa}{a^t MTMa} \; . \tag{11}$$

On montre, en écrivant que le gradient de J_D par rapport à a est nul, que le maximum est atteint pour a vecteur propre de $(MTM)^{-1} MBM$ associé à la plus grande valeur propre λ_1,

$$M^{-1} T^{-1} BMa \; = \; \lambda_1 a \tag{12}$$

Avec la métrique $M = T^{-1}$, on obtient,

$$BT^{-1} a \; = \; \lambda_1 a \; . \tag{13}$$

Les autres axes discriminants sont associés aux autres valeurs propres de BT^{-1} triées de façon décroissante. Le nombre de valeurs propres non nulles étant égale au rang de la matrice B, il y a au plus $k - 1$ axes discriminants. Dans le problème actuel d'identification en 3 classes de masses, la discrimination est effectuée dans un plan ($p = 2$), ce qui facilite la représentation graphique.

3.2. *Sélection de variables*

Le prétraitement des signaux temporels a consisté à générer 81 variables d'entrée à partir de chaque intervalle. Cependant, ces variables ont été construites de manière arbitraire et elles n'apportent pas nécessairement la même qualité d'information quant à la discrimination des masses. La sélection de certaines de ces variables est maintenant envisagée avec deux objectifs. D'une part, on cherche les variables d'entrées qui génèrent le meilleur classeur. Comme le nombre de données d'apprentissage est fini, il est important d'éviter le fléau de la dimensionalité [DUD 01] et cela peut être réalisé en limitant la complexité du classeur (cf. paragraphe 4.1), ou de manière structurelle, en réduisant le nombre de variables d'entrées. D'autre part, la connaissance des variables pertinentes permet, sur un plan industriel, d'extraire de la connaissance sur le problème et de réaliser des économies dans les campagnes de mesures en supprimant des capteurs et des opérations de traitement des données.

Les travaux présentant des méthodes de sélection de variables sont dejà très nombreux. Ces algorithmes ont généralement trois caractéristiques :

– La fonction d'évaluation : celle-ci permet d'évaluer la pertinence d'un sous-ensemble de variables par rapport à un critère donné. L'objectif de la sélection est donc d'optimiser cette fonction dans l'espaces des variables. Classiquement, les critères utilisés en discrimination sont basés sur la corrélation entre les variables et la classe [HAL 00], ou ils mesurent la séparabilité des classes [FUK 92] ou encore l'information mutuelle apportée par les variables [BAT 94]. Enfin, d'autres critères dépendent directement des spécificités d'un classifieur donné [KOH 97, LER 99, GUY 03].

– L'algorithme de recherche dans l'espace des variables : pour d variables explicatives, il existe 2^d sous-ensembles de variables. Lorsque d est grand, il est impossible de parcourir cet espace. Les techniques d'exploration classiquement rencontrées dans la littérature sont les algorithmes évolutionnaires [VAF 94] et les heuristiques d'énumération [AHA 96]. Elles sont comparées en Section 5.3.

– le critère d'arrêt.

Dans notre cas, il est numériquement trop coûteux d'évaluer la performance d'une méthode de sélection de variables car ce calcul implique d'entrainer les multiples classeurs comparés. Ainsi, une fonction d'évaluation indépendante de la méthode de discrimination est choisie, le critère du Λ de Wilks ([SAP 90]),

$$\Lambda_p = \frac{|W_p|}{|T_p|} , \tag{14}$$

où $|W_p|$ et $|T_p|$ sont, respectivement, les déterminants des matrices de variance intra-classe et totale (Equations (10) et (8)) pour un sous-ensemble de p variables. Λ_p peut être interprété comme une mesure inversement proportionnelle à la séparabilité des classes $\mathcal{M}_i$. Un sous-ensemble de variables qui minimise Λ_p est donc recherché.

La première méthode d'exploration des variables est une heuristique d'énumération du type plus forte pente discrète. Son organigramme est le suivant :

1) Créer une liste de toutes les variables.

2) Pour toutes les variables de la liste, calculer le Λ de Wilks sans la variable i, $\Lambda_p^{(i)} = \frac{|W_p^{(i)}|}{|T_p^{(i)}|}$.

3) Sélectionner et supprimer de la liste la variable qui minimise le critère.

4) Tant que le critère d'arrêt n'est pas satisfait, aller en 2.

Cette méthode est peu coûteuse car elle demande au maximum $d(d-1)/2$ évaluations. Cependant un critère d'arrêt de la suppression de variables peut être défini. En effet, la statistique

$$F_s = \frac{(\text{card}\,\mathcal{L} - k - p + 1)}{(k-1)} \frac{1-\Lambda_p}{\Lambda_p} \tag{15}$$

mesurant l'apport d'une p-ième variable pour la discrimination des classes suit une loi de Fisher à $(k-1, \text{card}\,\mathcal{L} - k - p + 1)$ degrés de liberté $\mathcal{F}_{k-1,\text{card}\,\mathcal{L}-k-p+1}$. Ainsi, pour un niveau de confiance α donné, le critère d'arrêt est un seuil f_s tel que

$$P(\mathcal{F}_{k-1,\text{card}\,\mathcal{L}-k-p+1}(x) > f_s) \geq 1 - \alpha \tag{16}$$

De manière pratique, la suppression de variables se poursuit tant que F_s est inférieure à f_s, ce qui statistiquement signifie que la variable supprimée n'apporte pas de contribution significative pour la discrimination au sens de la loi de Fisher et du niveau de confiance α [MCL 92].

L'autre méthode de sélection de variables est un algorithme évolutionnaire ([BAC 96]). Il s'agit, contrairement à l'approche précédente, d'un algorithme pseudo-stochastique qui, dans la limite d'une infinité d'évaluations, trouve presque certainement un optimum global. Par contre, son coût est supérieur à l'heuristique précédente. La fonction coût demeure le Λ de Wilks. Une solution, c'est à dire un choix de variables d'entrées, est codée sous la forme d'une liste d'entiers compris entre 1 et d, sans duplicata. Chaque entier de la liste identifie une variable d'entrée utilisée. Un opérateur de croisement crée une nouvelle solution à partir de 2 solutions dites parents. Il copie les entiers communs aux 2 parents puis copie, aléatoirement, les autres entiers soit de l'un ou l'autre des parents. Si les listes n'ont pas la même taille, le choix peut opposer un entier et du vide. L'opérateur de croisement est appliqué avec une probabilité p_c. Si il n'est pas appliqué, un des 2 parents est recopié comme nouvelle solution. Un opérateur de mutation perturbe les solutions en changeant, aléatoirement et avec une probabilité p_{mv}, la valeur de chaque entier. La mutation possède aussi une probabilité p_{mn}^k d'ajouter et d'enlever k variables. Ces opérateurs s'articulent dans un algorithme évolutionnaire du type $ES(n_p + n_e)$:

1) Initialisation aléatoire d'un ensemble $\mathcal{P}$ de n_p solutions potentielles. $N_{eval} = 0$.

2) $\mathcal{P}' = \emptyset$. Répéter pour $i = 1, n_e$,

a) Choisir au hasard 2 solutions de $\mathcal{P}$.

b) Créer une nouvelle solution par croisement puis mutation.

c) Calculer le Λ de Wilks associé, $N_{eval} = N_{eval} + 1$, ajouter la nouvelle solution à $\mathcal{P}'$.

3) Garder les n_p meilleures solutions de $\mathcal{P} \cup \mathcal{P}'$, qui constituent le nouveau $\mathcal{P}$.

4) Si $N_{eval} < N_{max}$, aller en 2. Sinon, fin.

L'algorithme mis en œuvre dans les expériences numériques décrites en 5.3 possède les réglages suivants : $n_p = 100, n_e = 10, N_{max} = 20000, p_c = 0.9, p_{mv} = 0.1$ et $p_{mn}^k = 0$.

3.3. *Taille optimale des intervalles de traitement*

La taille des intervalles, D, servant à calculer les variables d'entrées participe aussi à la définition de l'espace de recherche. Contrairement à l'analyse discriminante et à la sélection de variables qui sont des méthodes générales, la taille de l'intervalle n'intervient que quand un signal est transformé en variables d'entrées statistiques.

Sur le plan de la dynamique du véhicule, il y a une mémoire des sollicitations rencontrées qui est située entre 2 et 3 secondes. Les conditions de roulement ont également une longueur caractéristique, située entre dix secondes et une minute. Ainsi, la mécanique du problème indique qu'un intervalle inférieur à la seconde n'est pas suffisament informatif pour identifier la masse du véhicule. Réciproquement, un intervalle trop long, en mélangeant les conditions de roulement, ne présente pas une image claire de la masse du véhicule.

Le compromis entre tailles d'intervalles apparaît également en termes purement statistiques. Les variables extraites des intervalles temporels peuvent être considérées comme des variables aléatoires dont la variance décroit lorsque la taille de l'intervalle augmente. Par exemple, la variance de l'estimateur de la moyenne du signal varie en $1/D$. Vis à vis de la variance des variables, il est avantageux d'utiliser un intervalle long. Cependant, puisque le nombre de mesures acquises est fixe, augmenter D réduit le nombre d'intervalles, c'est à dire le nombre de données disponibles pour l'apprentissage. Or, dans une méthode d'apprentissage supervisé, la capacité de généralisation d'un algorithme est d'autant plus grande que le nombre d'exemples à disposition est élevé. Ces arguments statistiques confirment l'existence d'une taille d'intervalle optimale.

Des expériences numériques sont rapportées dans le paragraphe 5.2 où plusieurs tailles d'intervalles sont comparées.

4. Méthodes de discrimination

L'objectif final de ce travail étant de discriminer des véhicules par masse, la dernière étape est naturellement d'apprendre à reconnaître les différentes classes de masses à partir des données.

L'apprentissage et la reconnaissance de formes à partir d'exemples sont des sujets qui ont déjà été très largement étudiés que ce soit d'un point de vue théorique ou pratique et plusieurs principes d'induction ont été developpés tels que l'analyse discriminante [FUK 92], les méthodes par fênetre de Parzen ou les méthodes des plus proches voisins [DUD 01]. Actuellement, les méthodes les plus utilisées restent les réseaux de neurones [BIS 95] et les machines à support vectoriels [SCH 01, VAP 98] qui ont montré d'excellentes performances dans différentes applications.

Les différentes méthodes de discrimination de données que nous utilisons sont présentées en fonction du critère qu'elles considèrent pour effectuer l'apprentissage à partir des exemples : les machines à support vectoriel (en discrimination directe) maximisent une marge séparant les classes et les autres méthodes mises en œuvre minimisent, in fine, la probabilité d'erreur de classification.

4.1. *Contrôle de complexité*

Avant de présenter les méthodes de discrimination, des concepts fondamentaux liés au "contrôle de la complexité" et souvent utilisés par la suite sont récapitulés. Chaque méthode possède des mécanismes permettant de contrôler sa complexité. Les aspects généraux du contrôle de complexité sont discutés maintenant, tandis que ceux spécifiques à chaque méthode seront présentés avec la méthode.

Que ce soit à des fins de classification ou de régression, l'étape constitutive de tout apprentissage est l'approximation d'une loi génératrice de données, g, au moyen d'une fonctionnelle $f(;w)$ en identifiant ses paramètres w. Le choix de la fonctionnelle $f(;w)$ définit l'espace de fonctions (espace d'hypothèses) dans lequel l'approximation de g est recherchée. La taille de l'espace d'hypothèses est appelé *complexité* de f. Typiquement, une fonctionnelle flexible est complexe. L'espace d'hypothèses doit être suffisament grand pour permettre une bonne approximation de g. Réciproquement, un espace d'hypothèses restreint améliore l'identifiabilité de f, en particulier l'unicité et la stabilité de la solution de l'apprentissage. La *variance* est une mesure de la stabilité de f identifiée, c'est à dire une mesure de la sensibilité de la solution de l'apprentissage à la base expérimentale.

La complexité d'un modèle peut être contrôlée directement en choisissant la *structure* de f, par exemple en variant le nombre de paramètres w. Les techniques de *régularisation* contrôlent indirectement la complexité, non pas en changeant la taille de l'espace des hypothèses, mais en stabilisant les solutions de l'identification. Classiquement, la régularisation procède par limitation des courbures de f identifiée. Enfin, la mesure d'*erreur en validation* permet aussi un contrôle indirect de la complexité. Elle est très largement applicable car elle n'agit que sur la mesure de la qualité de l'apprentissage. Le principe est de diviser la base d'expériences $\mathcal{L}$ en une base d'apprentissage, $\mathcal{A}$, servant à identifier w et une base de validation, $\mathcal{V}$, servant à mesurer la qualité de f identifiée. Le biais de f identifiée qui résulte du choix d'une base d'apprentissage (sur-apprentissage) est ainsi réduit. La mesure d'erreur en validation peut être affinée par *validation croisée*, c'est à dire en moyennant les erreurs de validation pour plusieurs choix de d'ensembles de validation $\mathcal{V}_k$ (donc de $\mathcal{A}_k$, $\mathcal{L}$ restant le même).

4.2. *Critère de marge maximale entre classes : machines à supports vectoriels*

Dans le cas où deux ensembles de données sont séparables par une fonction linéaire, les MSV construisent une frontière de décision linéaire

$$f(x; w, b) = \langle w, x \rangle + b \tag{17}$$

qui maximise la distance entre l'hyperplan de décision et le point le plus proche de chaque classe (Figure 5). Si les deux classes sont étiquetées par $y_i = \pm 1$, ceci équivaut à résoudre le problème

$$\begin{cases} \min_{w,b} \frac{1}{2}\|w\|^2 \ , \\ \text{sous les contraintes,} \quad y_i f(x_i; w, b) \geq 1 \quad i = 1, \text{card}\, \mathcal{A} \ . \end{cases} \tag{18}$$

Dans le cas où les données ne sont pas séparables, des variables d'écart ξ_i et un paramètre de régularisation C sont introduits, et le problème devient :

$$\begin{cases} \min_{w} \frac{1}{2}\|w\|^2 \ + \ C \sum_{i=1}^{\text{card}\, \mathcal{A}} \xi_i \ , \\ \text{sous les contraintes,} \quad y_i f(x_i; w, b) \geq 1 - \xi_i \quad i = 1, \text{card}\, \mathcal{A} \ , \\ \xi_i \geq 0 \quad i = 1, \text{card}\, \mathcal{A} \ . \end{cases} \tag{19}$$

Le problème primal (19) est convexe. Il est donc équivalent de résoudre sa forme duale. La résolution du problème de marge maximale sous sa forme duale présente deux intérêts. D'une part, elle permet de ne retrouver les données $x^{(i)}$ que sous forme de produit scalaire, ce qui autorisera par la suite l'extension des MSVs aux frontières non linéaires. D'autre part, le problème primal (19) est contraint par card $\mathcal{A}$ relations linéaire, que la forme duale transformera en des contraintes plus faciles à traiter

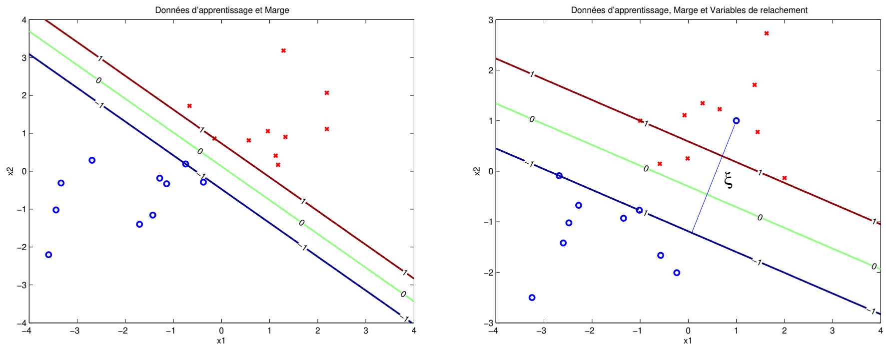


Figure 5. *Illustration de la discrimination MSV pour des données linéairement séparables (à gauche) et non séparables (à droite).*

numériquement. Pour l'écriture duale de (19), on introduit α_i, les multiplicateurs de Lagrange associés aux premières contraintes de (19). La formulation duale est considérablement allégée en prenant en compte le fait que le gradient du Lagrangien par rapport à w et b doit être nul, ce qui conduit au problème dual de Wolfe (cf. [BUR 98]),

$$\left\{\begin{array}{l} \max_{\alpha_i} \sum_{i=1}^{\text{card } \mathcal{A}} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\text{card } \mathcal{A}} \alpha_i \alpha_j y_i y_j x^{(i)} x^{(j)} \ , \\ \text{sous les contraintes,} \\ 0 \ \leq \ \alpha_i \ \leq \ C \ , \\ \sum_{i=1}^{\text{card } \mathcal{A}} \alpha_i y_i \ = \ 0 \ . \end{array}\right. \qquad (20)$$

Le problème dual (20) est quadratique. Il est résolu en utilisant un algorithme de programmation quadratique avec une stratégie de contraintes actives ([GIL 97]). Soient α_i^* les solutions du problème dual. Les solutions dans l'espace primal découlent des relations de Kuhn et Tucker appliquées à (19),

$$w^* \ = \ \sum_{i=1}^{\text{card } \mathcal{A}} \alpha_i^* y_i x^{(i)} \ . \qquad (21)$$

Pour estimer b^*, on utilise les points supports correctement classés pour lesquels on peut écrire

$$\forall i \text{ tel que } 0 \ < \ \alpha_i^* \ < \ C \quad , \quad y_i(x^{(i)} w^* + b^*) - 1 \ = \ 0 \ . \qquad (22)$$

Numériquement, on moyenne les b^* ainsi calculés. La solution α_i^* (ou w^* et b^*) du problème (20) n'est pas toujours unique, mais, puisque le problème est quadratique, toute solution est un optimum global. Cette propriété, que ne possèdent pas, par exemple, les réseaux de neurones, est importante en pratique car elle contribue à la robustesse et à la simplicité d'utilisation des MSV.

Enfin, en phase de test, l'affectation d'un cas x à l'une des deux classes se fait en prenant le signe de $f(x; w^*, b^*)$. Dans le cas où le nombre de classes k est supérieur à 2, l'algorithme s'étend simplement par une stratégie de "un contre les autres". k frontières de décision $f_k(x; w^{(k)}, b^{(k)})$ sont générées. x est affecté à la classe v si

$$v \;=\; \arg\max_k f_k(x; w^{(k)}, b^{(k)}) \; . \tag{23}$$

L'extension de la discrimination MSV aux frontières non linéaires est obtenue en considérant une transformation non linéaire sur les entrées x, $\Phi : \mathcal{X} \rightarrow \mathcal{F}$. Si on applique l'algorithme de discrimination linéaire MSV dans $\mathcal{F}$, on procède à une discrimination non linéaire dans $\mathcal{X}$. Notons que $\mathcal{F}$ est souvent un espace de grande dimension, voire de dimension infinie, ce qui peut contribuer à rendre les données plus séparables. Réciproquement, se situer dans un espace de grande dimension multiplie le nombre de paramètres de l'hyperplan de séparation ce qui, pour un algorithme de discrimination général, détériore la capacité de généralisation. Néanmoins, les MSVs se comportent remarquablement bien dans ces espaces à haute dimension grâce à la maximisation de marge comme le montre les nombreuses applications convaincantes rapportées dans la littérature ([CHA 99, DRU 97, MüL 99, OSU 97, ZIE 00]).

La transformation de chaque donnée d'apprentissage $x^{(i)}$ par Φ pourrait s'avérer numériquement coûteuse. Cependant, en remplaçant w^* par sa valeur (Equation (21)) dans l'apprentissage (Equation (20)) ou en phase de test (Equation (22) puis (17)), on constate que les entrées n'interviennent que sous la forme de produits scalaires $\langle \Phi(x), \Phi(x') \rangle$. Lorsqu'il existe un noyau admissible, on peut remplacer le produit scalaire par le noyau associé,

$$K(x, x') \;=\; \langle \Phi(x), \Phi(x') \rangle \tag{24}$$

et il n'est plus nécessaire d'expliciter la transformation Φ pour fabriquer une MSV non linéaire. L'existence d'un noyau est difficile à caractériser en général (c'est la condition de Mercer, cf. [BUR 98]), mais elle est établie pour les deux noyaux que nous considérons ici, les noyaux polynomiaux,

$$K(x, x') \;=\; (\langle x, x' \rangle + 1)^o \; , \qquad \text{avec} \quad o \in \mathbb{N} \; , \tag{25}$$

et les noyaux Gaussiens,

$$K(x, x') = e^{-\frac{\|x - x'\|^2}{2\sigma^2}} \; . \tag{26}$$

En variant le noyau K, différentes transformations Φ de l'espace des entrées sont donc utilisées implicitement. La sélection de variables par élimination ou combinaison linéaire, qui est abordée au paragraphe 3, est aussi une opération modifiant l'espace des entrées. La différence avec le changement de noyaux est que les changements de variables sont explicites, donc découplés de la méthode de discrimination, et que, en général, la dimension de l'espace de recherche après transformation est réduite.

La complexité des MSVs est directement contrôlée par l'expression du noyau K, et le paramètre de régularisation C. Par exemple, une valeur de C grande (on autorise donc des grandes valeurs des α_i associée à un noyau polynomial de degré élevé produit des frontières f très flexibles, donc un classeur de grande complexité.

Le lecteur cherchant une présentation plus détaillée des MSV se référera à [BUR 98, COR 95b, HEA 98, VAP 95].

4.3. *Critère de minimisation d'erreur de classement : classeurs Bayésiens*

Dans un cadre Bayésien classique ([DUD 01]), l'erreur de classement est minimisée directement en affectant le signal associé aux variables d'entrées x à la classe $\mathcal{M}_i$ telle que,

$$P(\mathcal{M}_i|x) \ > \ P(\mathcal{M}_j|x) \quad , \quad \forall j \neq i \ , \tag{27}$$

ce qui est implique

$$x \ \in \ \mathcal{M}_i \quad \text{si} \quad p(x|\mathcal{M}_i)P(\mathcal{M}_i) \ > \ p(x|\mathcal{M}_j)P(\mathcal{M}_j) \quad , \quad \forall j \neq i \ . \tag{28}$$

Toutes les méthodes de discrimination discutées ci-après cherchent, avec diverses stratégies, à satisfaire la relation (28).

4.3.1. *Discrimination par fenêtres de Parzen*

La discrimination par fenêtres de Parzen utilise directement la relation (28) en construisant au préalable, au moyen de la base d'apprentissage $\mathcal{A}$, $p(x|\mathcal{M}_k)$ et $P(\mathcal{M}_k)$. $P(\mathcal{M}_k)$ est directement estimé comme la fréquence d'apparition de la classe $\mathcal{M}_k$ dans $\mathcal{A}$. Soit n_k, le nombre de données de $\mathcal{A}$ appartenant à $\mathcal{M}_k$. Les densités de probabilités sont estimées par[1]

$$p(x|\mathcal{M}_k) \ = \ \frac{1}{n_k} \sum_{i \ \in \ \{\mathcal{A} \cap \mathcal{M}_k\}} \frac{1}{V_d} \, \psi \left(\frac{x - x_i}{h} \right) \ , \tag{29}$$

où h est un paramètre d'étalement, $V_d \ = \ h^d$ et ψ une distribution dite fenêtre de Parzen. Classiquement, on choisit une fenêtre gaussienne pour ψ, si bien que l'estimation de densité (29) peut être interprétée comme un histogramme lissé dont les barres sont centrées sur les points de l'échantillon. h, un paramètre de contrôle de complexité, est ajusté par validation croisée, c'est à dire en minimisant le taux de mauvais classements sur l'ensemble de validation $\mathcal{V}$. Puisque h est un scalaire et qu'une grande précision numérique n'a pas de sens pour des bases de données de tailles finies, cette minimisation est réalisée par exploration exhaustive des valeurs de h sur une grille.

4.3.2. *Discrimination par régression et seuillage*

La masse d'un véhicule est une quantité physique continue. Bien qu'il soit plus important, dans un contexte industriel, de classer la masse que d'en estimer sa valeur, il est tentant de chercher à exploiter la nature continue de la masse dans le classement. C'est pourquoi les méthodes décrites maintenant procèdent par discrimination après régression. Un régresseur linéaire, neuronal, et une MSV sont utilisés. Les deux

---

1. Pour alléger les notations, $(x^{(i)}, y^{(i)}) \ \in \ \{...\}$ est noté $i \ \in \ \{...\}$.

premiers régresseurs sont obtenus par minimisation de norme quadratique, le dernier par minimisation d'une norme ϵ-insensible. Les caractéristiques des régresseurs sont discutées en fin de paragraphe.

La masse étant un scalaire, la discrimination après régression est une simple opération de seuillage. Les seuils sont ajustés en minimisant le taux d'erreur de classement sur la base d'apprentissage $\mathcal{L}$ par exploration exhaustive des positions suivant une grille. Cette minimisation des taux d'erreurs de classements par réglage des seuils fait des méthodes de discrimination après régression des classeurs Bayésiens. On peut montrer (cf. [RAK 01]), sous une hypothèse de monotonie du régresseur, qu'elles satisfont la relation (28). A la différence de la discrimination par fenêtre de Parzen, elles ne discriminent pas dans l'espace des variables d'entrées x, mais dans l'espace des masses estimées, $\widehat{m}(x)$. En ce sens, et comme pour le choix des noyaux dans les MSV (paragraphe 4.2), on peut voir dans la discrimination après régression un changement implicite des variables d'entrées. Dans le cas présent cependant, l'espace des variables explicatives ne présente qu'une dimension, le régresseur $\widehat{m}(x)$. La différence entre les méthodes de discrimination après régression vient de la construction de $\widehat{m}(x)$.

<u>Régressions à partir de critères de moindres carrés</u>

Soit une forme paramétrée de fonctions à valeurs dans $\mathbb{R}$, $f(x;w)$, où x sont les entrées des fonctions et w leurs paramètres. La forme générale de ces fonctions correspond à un réseau de neurones simple à trois couches (cf. [BIS 95]),

$$f(x;w) \;=\; \sum_{i=1}^{N_{NC}} \{w_{oi}\phi_i(\sum_{j=1}^{d} w_{ij}x_j)\} \;, \qquad (30)$$

où N_{NC} est le nombre de neurones cachés. Les fonctions ϕ_i sont soit l'identité soit $\tanh$. Si tous les ϕ_i sont l'identité, f est une fonction linéaire en x. Dans le cas contraire, la fonction f est non linéaire en x. Si de plus tous les ϕ_i sont des $\tanh$, f est désignée ici comme une fonction neuronale. La fonction neuronale possède trois propriétés qui la rendent potentiellement intéressante par rapport à d'autres fonctions non linéaires comme les polynômes ou les séries de Fourier (cf. [BIS 95]). Tout d'abord, elle peut approcher n'importe quelle fonction avec une précision arbitraire pourvu que N_{NC} soit suffisament grand (propriété d'approximation universelle). Puis, en changeant N_{NC}, on peut contrôler la variance d'une fonction neuronale indépendamment de tout autre facteur tel que la dimension de l'espace des entrées ou le nombre de données. Enfin, le nombre de paramètres de la fonction neuronale croit linéairement avec la dimension de l'espace des entrées d (propriété de parcimonie).

Les régresseurs linéaire et neuronaux sont construits par résolution d'un problème de moindres carrés non linéaire régularisé :

$$\left\{ \begin{array}{l} \min_w J_{2\nu}(w) \;, \\ J_{2\nu}(w) \;=\; J_2(w) + \frac{\nu}{2}w^2 \;, \\ J_2(w) \;=\; \frac{1}{2 \text{ card } \mathcal{A}} \sum_{k\in\mathcal{A}}(f(x^{(k)};w) - y^{(k)})^2 \;. \end{array} \right. \qquad (31)$$

La résolution de ce problème est effectuée au moyen de l'algorithme de Levenberg-Marquardt ([LEV 44],[MAR 63]). Il s'agit d'un algorithme de la famille des méthodes quasi-Newtoniennes exploitant les spécificités des problèmes de moindres carrés. Il procède par linéarisation de la fonction f à chaque itération et contrôle du pas dans l'espace des paramètres w. Il nécessite le calcul du gradient de f par rapport à w, qui est réalisé par rétro-propagation ([RUM 86]). Si l'on dénote par w^* la solution du problème de moindres carrés (31), on appelle ici régresseur de la masse la fonction $\widehat{m}(x) = f(x; w^*)$. Soit $< y|x >$ la moyenne des masses y en x,

$$< y|x > = \int y(x)p(y|x)dy \; . \tag{32}$$

Dans la limite d'un nombre infini de données, si $f(x; w')$ reproduit $< y|x >$, alors $f(x, w')$ est solution du problème de moindres carrés non régularisé, $\min_w J_2(w)$. Ce cas de figure, qui demande une certaine souplesse de la part de f, est celui rencontré quand f est une fonction neuronale avec N_{NC} grand. Sinon, avec $w' = \arg\min_w J_2(w)$, $f(x; w')$ est la projection orthogonale de $< y|x >$ sur l'espace des fonctions $f(x; w)$. Ce dernier cas est typique de f linéaire.

Trois stratégies de contrôle de complexité sont mises en œuvre. Il s'agit d'une part d'un contrôle structurel sur la fonction servant de régresseur, f, avec, par ordre croissant de variance, f linéaire puis f neuronal avec N_{NC} croissant. D'autre part, la régression est régularisée par pénalisation sur les poids au moyen du terme $\nu/2w^2$. Il a en effet été constaté empiriquement ([BIS 95]) que la pénalisation par les poids réduit les courbures des régresseurs neuronaux et contribue, pour un ν ad-hoc, à améliorer leur prédiction sur des données non apprises. Enfin, le sur-apprentissage est évité en arrêtant l'apprentissage d'après l'erreur de validation ce qui, pour les réseaux de neurones, s'appelle l'arrêt prématuré. Comme indiqué dans la formulation (31), les expériences de $\mathcal{A}$ servent à calculer une direction de modification des paramètres. Les expériences de $\mathcal{V}$ définissent l'arrêt de l'apprentissage : on arrête de minimiser $J_{2\nu}$ sur $\mathcal{A}$ lorsque l'erreur $J_{2\nu}$ sur $\mathcal{V}$ croît pendant Ninc itérations successives. Le principe est d'arrêter l'apprentissage lorsqu'il commence à interpoler le bruit sur $\mathcal{A}$, ce que détecte le test sur $\mathcal{V}$.

Régression par machines à supports vectoriels

Dans un contexte dit de ϵ-SV régression ([SMO 98b, EVG 00]), l'objectif est de trouver une fonction f qui présente une déviation maximale de ϵ des données tout en étant la plus régulière possible. Pour un espace de fonctions f linéaires (Equation (17)), la régularité se traduit par des w de faible norme, et le problème de ϵ-SV régression s'écrit,

$$\begin{cases} min_w J_{\epsilon\nu}(w) \; , \\ J_{\epsilon\nu}(w) \; = \; J_{\epsilon}(w) + \frac{\nu}{2}\|w\|^2 \; , \\ J_{\epsilon}(w) \; = \; \frac{1}{\text{card } \mathcal{L}} \sum_{i=1}^{\text{card } \mathcal{L}} |f(x^{(i)}) - y^{(i)}|_{\epsilon} \; , \end{cases} \tag{33}$$

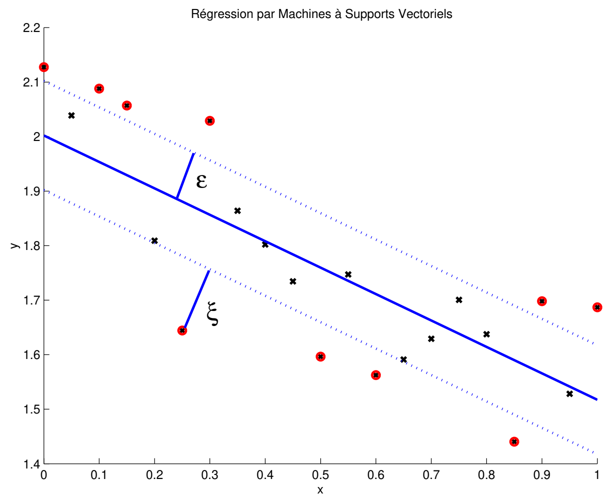


Figure 6. *Illustration du principe des MSV en régression.*

La différence essentielle avec les critères de moindres carrés de l'Equation (31) est le premier terme qui est une norme ϵ-insensible,

$$|u|_\epsilon = \begin{cases} 0 & \text{si } |u| \leq \epsilon \ , \\ |u| - \epsilon & \text{sinon .} \end{cases} \tag{34}$$

Le problème (33) peut également s'écrire sous une forme quadratique avec contraintes,

$$\begin{cases} \min_{w,b,\xi_i,\xi_i^*} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{\text{card } \mathcal{L}} (\xi_i + \xi_i^*) \ , \\ \text{sous les contraintes,} \\ y^{(i)} - \langle w, x^{(i)} \rangle - b \leq \epsilon + \xi_i \ , \\ \langle w, x^{(i)} \rangle + b - y^{(i)} \leq \epsilon + \xi_i^* \ , \\ \xi_i, \xi_i^* \geq 0 \ , \\ i \ = \ 1, \text{card } \mathcal{L} \ . \end{cases} \tag{35}$$

Une illustration de ce problème est donnée sur la Figure 6. Celle-ci montre le ϵ-tube autour de la fonction de régression linéaire ainsi qu'un exemple de variable d'écart ξ. A partir de ce point, en remarquant que le problème (35) est équivalent au problème des MSVs en discrimination, Equation (19), le lecteur est renvoyé au paragraphe 4.2 pour la résolution de la ϵ-SV régression et sa généralisation aux fonctions non linéaires.

En ce qui concerne le contrôle de complexité, outre le paramètre C et le choix du noyau K, les MSV en régression ont une variance croissante lorsque ϵ diminue.

Pour plus de détails sur la régression par MSV, on consultera [COR 95a, SCH 97, SMO 98a, VAP 98].

5. Expériences numériques

Cette section présente les résultats d'expériences numériques où les quatre niveaux d'optimisation de tous les problèmes de fouille de données sont sollicités.

5.1. *Contrôle de complexité et calcul des erreurs*

Pour estimer l'erreur de généralisation des classeurs développés, nous avons procédé à des techniques de rééchantillonnage (ici, la validation croisée) . Ainsi, plusieurs ensembles d'apprentissage et de validation ont été créés à partir de la base expérimentale. Les données qui sont classées dans ce travail sont issues de séries temporelles. Elles possèdent donc une certaine corrélation et par conséquent, les données constituant $\mathcal{A}$ et $\mathcal{V}$ ne sont pas choisies au hasard dans $\mathcal{L}$ mais par ordre chronologique. $\mathcal{A}$ représente une proportion $n_{\mathcal{A}}$ de $\mathcal{L}$. A la $k^{\text{ième}}$ validation croisée, les ensembles $\mathcal{A}_k$ et $\mathcal{V}_k$ sont obtenus de la manière suivante :

$$\mathcal{A}_k = \mathcal{A}_{1,k} \cup \mathcal{A}_{2,k} \cup \mathcal{A}_{3,k} \ , \tag{36}$$

avec

$$\mathcal{A}_{i,k} = \{x_j \in \mathcal{M}_i \ : j \in 1+k, \ldots, N_k\} \quad , \quad N_k = (N_i + k)\, mod\, (n_{\mathcal{A}} \text{ card } \mathcal{L}) \ , \tag{37}$$

puis

$$\mathcal{V}_k = \mathcal{L}_k \backslash \mathcal{A}_k \ . \tag{38}$$

$n_{bases} = 50$ bases croisées sont obtenues de la sorte.

La plupart des paramètres de méthodes exerçant un contrôle sur la complexité sont ajustés par minimisation de l'erreur de validation croisée. La Figure 7, par exemple, montre l'influence de la largeur de bande du noyau gaussien σ sur l'erreur en validation croisée. $\sigma = 0.7$, qui correspond au minimum de l'erreur, est donc la valeur utilisée pour cet ensemble $\mathcal{A}_k$ et $\mathcal{V}_k$.

Pour les MSVs, le paramètre de régularisation C, la largeur des noyaux Gaussiens σ et le degré des polynômes o sont aussi ajustés par minimisation de l'erreur en validation croisée. Les MSVs en régression ont une largeur de tube ϵ fixée à 0.1 .

Une fois que les paramètres de contrôle de complexité précédemment décrits sont ajustés par minimisation de l'erreur en validation croisée, les autres paramètres des méthodes sont identifiés sur toute la base expérimentale $\mathcal{L} = \mathcal{A} \cup \mathcal{V}$. Les taux d'erreur de classification utilisés pour comparer les méthodes résultent de tests sur une autre partie des expériences, la base de tests $\mathcal{T}$, $\mathcal{L} \cap \mathcal{T} = \emptyset$. Les bases d'apprentissage et de validation , $\mathcal{A}$, $\mathcal{V}$ et représentent 75% et 25% de l'ensemble $\mathcal{L}$.

L'apprentissage des réseaux de neurones suit globalement la même procédure, mais, comme il n'y a pas de séparation claire entre les paramètres de contrôle de complexité et les autres, une erreur de validation simple est prise en compte. Un exemple

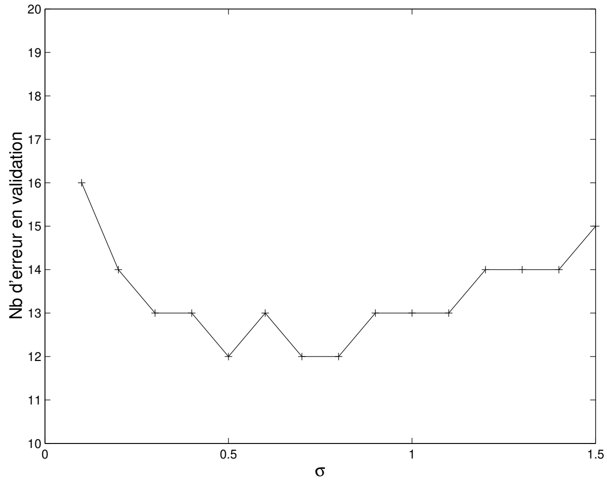


Figure 7. *Illustration du principe d'optimisation des paramètres pour la discrimination par fenêtre de Parzen.*

est donné en Figure 8. La stratégie d'arrêt prématuré identifie le réseau de la troisième itération car il minimise l'erreur de validation. Pour les régresseurs linéaires et neuronaux, le paramètre de régularisation ν est fixé à 0.1 . Un certain contrôle structurel de complexité est aussi exercé dans les réseaux de neurones à travers la considération de trois réseaux. Tous ont un neurone linéaire en sortie. Leurs caractéristiques diffèrent au niveau du nombre et du type de neurone cachés. Par ordre croissant de complexité, le premier réseau a un neurone linéaire en couche caché, le second a 7 neurones sigmoïdaux et un neurone linéaire et le troisième réseau a 8 neurones sigmoïdaux.

5.2. *Intervalles de traitement*

Le problème de la durée de l'intervalle de traitement des signaux, D, a été soulevé en 3.3. Il est repris ici de manière empirique par la comparaison de cas d'apprentissages. Plusieurs jeux de variables d'entrées sont construits en faisant varier la taille de la segmentation des signaux temporels et donc le nombre de mesure D dans chaque intervalle. Chaque jeu possède donc un nombre de données qui est supposé être inversement proportionnel aux bruits présents dans les variables.

Deux classeurs sont appris et leurs performances sont comparées. Le premier classeur est une analyse discriminante suivie d'une discrimination par fenêtres de Parzen, le second est un réseau de neurones de la seconde structure (7 neurones cachés sigmoïdaux et un linéaire) suivi d'un seuillage.

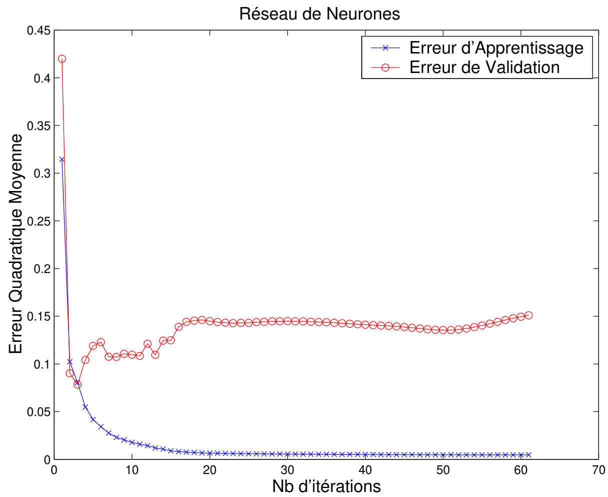


Figure 8. *Illustration du principe d'arrêt prématurée pour un réseau de neurones avec un neurone linéaire et 7 neurones sigmoïdaux en couche caché.*

Durée intervalle (normalisée)	Analyse Discriminante et Parzen	réseau de neurones et seuillage
1	69.17 [66.75 ;71.58]	48.38 [47.13 ;49.64]
2	75.05 [72.44 ; 77.65]	50.76 [47.47 ;54.08]
3	74.21 [70.22 ; 78.20]	55.26 [52.08 ;58.45]
4	43.03 [38.41 ;47.65]	51.82 [47.64 ;56.00]

Tableau 1. *Taux de classement correct en fonction de la longueur des intervalles de traitement des signaux, 2 méthodes, 81 variables explicatives, avec intervalles de confiance à 80%.*

Par souci de n'utiliser la base de test $\mathcal{T}$ que pour la comparaison finale des méthodes, la base expérimentale utilisée ici est uniquement $\mathcal{L}$ subdivisé en $\mathcal{A}_k$, $\mathcal{V}_k$ et $\mathcal{T}_k$ réprésentant respectivement 60%, 20% et 20% de $\mathcal{L}$. Les résultats du Tableau 1 sont obtenus en moyennant les observations de 10 expériences. Ils confirment que le taux de classement correct dépend de la taille de l'intervalle. Puisque les capacités des classeurs à s'accomoder de variables bruitées et à généraliser à partir d'un certain nombre de données diffèrent, la taille optimale des intervalles est fonction de la méthode. Pour l'analyse discriminante suivie de discrimination par Parzen, la longueur idéale est située entre 2 et 3 unités normalisées, contre 3 à 4 pour le réseau de neurones suivi du seuillage. Malgré la dépendance à l'égard du classeur, la taille de 3 unités est un bon compromis. Cette valeur est utilisée dans la suite des expériences.

5.3. *Variables explicatives pertinentes*

Les deux méthodes de sélection de variables décrites au paragraphe 3.2, une heuristique d'énumération et un algorithme évolutionnaire, sont comparées. Le Λ de Wilks qui estime la performance d'un jeu de variables est non seulement sensible aux variables sélectionnées, mais il varie aussi en fonction des données. Pour atténuer cette influence, une valeur moyenne du Λ de Wilks est calculée par validation croisée, Λ_{VC},

$$\Lambda_{VC} = \sum_{k=1}^{n_{bases}} \Lambda_k \ . \tag{39}$$

où Λ_k est le Λ de Wilks obtenu à partir du k-ième jeu de données.

Les variables sélectionnées par l'heuristique d'énumération sont données dans le Tableau 2. L'ordre de sélection peut être interprété comme un ordre de pertinence décroissante. Pour un niveau de confiance de 95% de la statistique de Fisher, 55 variables sont choisies. Les variables sélectionnées par l'algorithme évolutionnaire sont rapportées dans le Tableau 3. On remarque qu'il y en a 49 au total, ce qui est inférieur aux 55 variables issues de l'énumération. On remarque également que les 10 variables les plus significatives au sens de l'énumération des Λs de Wilks ont aussi été sélectionnées par l'algorithme évolutionnaire. Parmi les signaux, ceux issus de l'accélération verticale au centre de gravité et de la vitesse du véhicule semblent particulièrement importants. Les moments spectraux relatifs sont les traitements les plus souvent sélectionnés.

5.4. *Comparaisons de méthodes*

Suite à la sélection de variables, 5 méthodes de classement, dont certaines sont associées à une analyse discriminante, sont comparées :

1) Analyse discriminante puis discrimination par fenêtres de Parzen.

2) Analyse discriminante puis discrimination par MSV.

3) MSV à noyau gaussien en classement.

4) Régression neuronale puis seuillage. Trois structures de réseaux de neurones sont observées dont les couches cachées sont constituées, respectivement, d'un neurone linéaire, d'un neurone linéaire et 7 neurones sigmoïdaux, et de 8 neurones sigmoïdaux.

5) MSV à noyau polynomial en régression puis seuillage.

Les pourcentages de succès en classement des différentes méthodes sont résumés dans les Tableaux 4, 5 et 6. Dans le Tableau 4, les 81 variables d'entrées servent, alors que seules les variables sélectionnées par l'algorithme évolutionnaire et l'heuristique d'énumération sont utilisées dans les Tableaux 5 et 6, respectivement. La Figure 9 donne des exemples de frontières de décision calculées.

	Unité									
Dizaine	0	1	2	3	4	5	6	7	8	9
0		CdgV M_2	Vitesse M_3	Vitesse M_2	Vitesse M_1	CdgV a_{min}	Vitesse M_4	PivArg M_1	PivArg $skew$	CdgV M_3
1	PivAvg $skew$	Vitesse moy	CdgV M_1	CdgV M_4	PivAvd M_2	PivAvd M_3	PivArd M_2	PivArd M_3	CdgL M_3	CdgL M_2
2	CdgL M_4	PivAvg $kurt$	PivArg $kurt$	CdgT a_{max}	CdgT $skew$	Vitesse $skew$	PivAvd var	PivArd M_1	PivArd M_4	PivArg M_3
3	PivAvd M_4	Vitesse a_{min}	CdgT var	CdgT E	CdgT M_2	PivArd E	Vitesse E	PivAvg a_{min}	PivAvg E	CdgV E
4	Vitesse $kurt$	CdgV a_{max}	PivArg a_{max}	PivAvd $skew$	PivArd $skew$	PivAvg M_2	PivAvg M_3	PivAvg M_4	PivArg M_4	PivArg a_{min}
5	PivAvd M_1	CdgL E	CdgL var	CdgT M_3	CdgT M_4	CdgV $kurt$	CdgT a_{min}	CdgL M_1	Vitesse a_{max}	CdgL a_{min}
6	CdgL $kurt$	PivAvg a_{max}	PivAvd a_{max}	PivArd a_{min}	PivArg E	PivArg M_2	PivAvg M_1	CdgV var	CdgT $kurt$	CdgT M_1
7	PivArg var	CdgL $skew$	CdgL a_{max}	Vitesse var	PivAvd E	PivArd var	PivArd a_{max}	PivArd $kurt$	PivAvd $kurt$	PivAvd a_{min}
8	CdgV $skew$	PivAvg var								

Tableau 2. *Ordre de sélection des variables par énumération au moyen du Λ de Wilks. Chaque variable est composée d'un signal et d'un traitement. Cdg/Piv = centre de gravité/pivot, Av/Ar = avant/arrière, d/g = droite/gauche, L/T/V = longitudinal/transversal/vertical.*

Signal	Informations							
CdgV	M_1	M_2	M_3	M_4	$kurt$	a_{max}	a_{min}	
Vitesse	M_1	M_2	M_3	M_4	$mean$	var	E	a_{min}
CdgL	M_1	M_2	M_3	M_4	E			
CdgT	M_1	M_4	var	$skew$	E	a_{max}		
PivArd	M_1	M_2	M_3	M_4	$kurt$	a_{min}		
PivArg	M_1	M_4	$skew$	$kurt$	E	a_{max}	a_{min}	
PivAvd	M_2	M_3	M_4	var	$skew$	$kurt$		
PivAvg	M_2	var	$skew$	$kurt$				

Tableau 3. *Variables retenues par l'algorithme évolutionnaire en minimisant le Λ de Wilks. Chaque variable est composée d'un signal et d'un traitement. Cdg/Piv = centre de gravité/pivot, Av/Ar = avant/arrière, d/g = droite/gauche, L/T/V = longitudinal/transversal/vertical.*

Les trois tableaux montrent que, sur ce problème, l'analyse discriminante est la composante principale des meilleures méthodes. Elle fait gagner en moyenne 25% de

Méthodes	Borne inférieure	Moyenne	Borne Supérieure
Analyse Discriminante et Parzen	73.93	74.72	75.50
Analyse Discriminante et MSV	73.85	74.61	75.37
MSV à noyau gaussien	38.51	39.26	40.00
Réseaux de neurones 1	57.08	57.86	58.64
Réseaux de neurones 2	50.66	51.32	51.98
Réseaux de neurones 3	51.76	52.50	53.24
MSV à noyau polynomial	68.38	68.61	68.84

Tableau 4. *Pourcentage de succès de classement pour différentes stratégies sans sélection de variables. Les bornes sont l'intervalle de confiance à 80%.*

Méthodes	Borne inférieure	Moyenne	Borne Supérieure
Analyse Discriminante et Parzen	79.25	79.78	80.31
Analyse Discriminante et MSV	80.11	80.61	81.11
MSV à noyau gaussien	49.19	49.63	50.06
Réseaux de neurones 1	63.58	64.16	64.74
Réseaux de neurones 2	52.72	53.38	54.05
Réseaux de neurones 3	50.94	51.90	52.86
MSV à noyau polynomial	71.50	71.78	72.07

Tableau 5. *Pourcentage de succès de classement pour différentes stratégies en utilisant les variables selectionnées par l'algorithme évolutionnaire. Les bornes sont l'intervalle de confiance à 80%.*

taux classement correct. La meilleure méthode de discrimination, avec plus de 80% de classement corrects, est une analyse discriminante suivie d'un classement par MSV à noyau Gaussien. De manière générale, la réduction de dimensions de l'espace des entrées est un facteur d'amélioration du classement : les classeurs après régression ont des pourcentages de succès de 5 à 20% supérieurs aux MSV en classement direct ; les classeurs après sélection sont de 5 à 10% supérieurs aux classeurs œuvrant sur les 81 variables (Tableaux 5 et 6 versus 4). Parmi les méthodes de sélection de variables, l'algorithme évolutionnaire fait croître le taux moyen de succès des méthodes de 2 à 3% par rapport à l'heuristique d'énumération (Tableau 5 versus Tableau 6). Parmi les méthodes de seuillage après régression, les MSVs à noyau polynomial dépassent les régresseurs à base de moindres carrés de 7%. La meilleure structure de régresseur neuronal est celle avec un neurone linéaire dans la couche cachée.

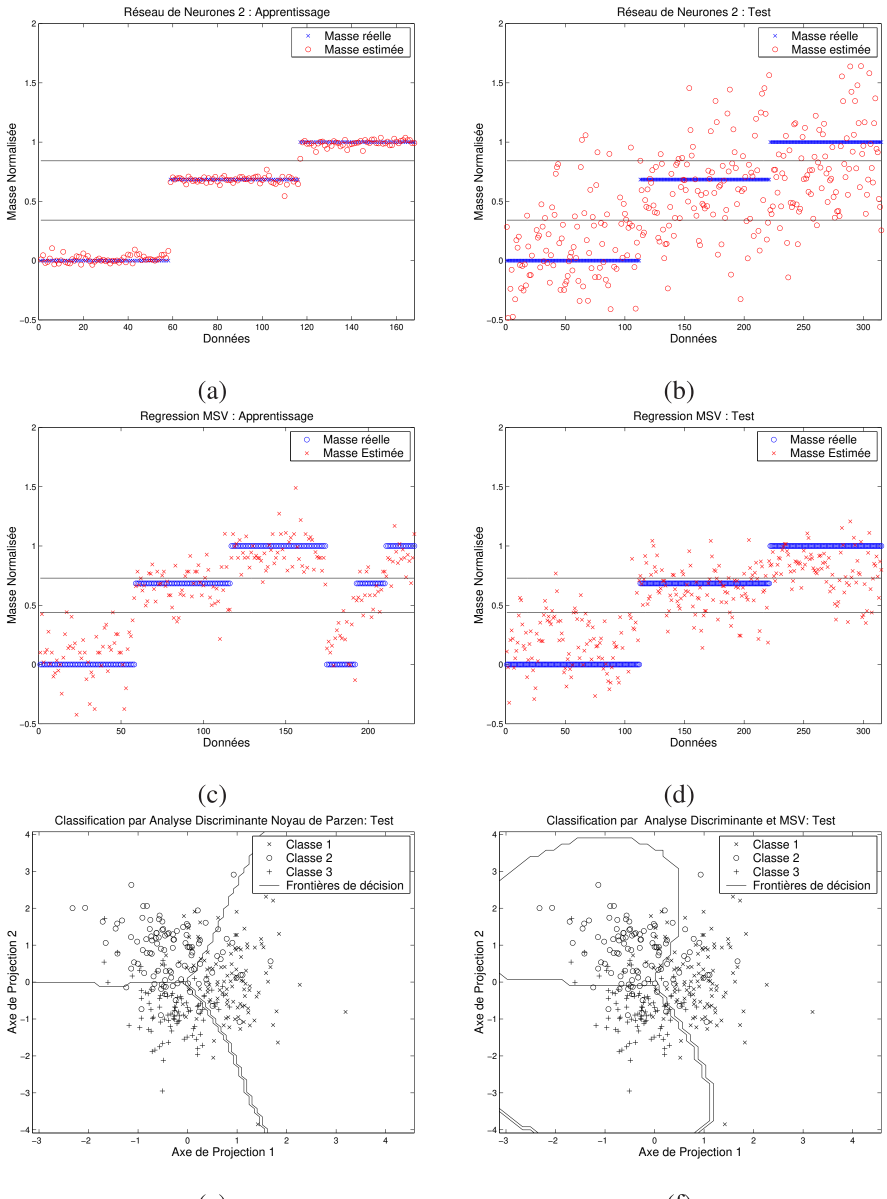


Figure 9. *Exemples de frontières de décision pour les différents classeurs utilisant les 81 variables d'entrées. (a) Seuils de décision pour un réseau de neurones, ensemble d'apprentissage. (b) Application de ces seuils à l'ensemble de tests. (c) Seuils de décision pour une MSV, ensemble d'apprentissage. (d) Application de ces seuils à l'ensemble de tests. (e) Classement de l'ensemble de tests par analyse discriminante et fenêtres de Parzen. (f) Classement de l'ensemble de tests par analyse discriminante et MSV.*

Méthodes	Borne inférieure	Moyenne	Borne Supérieure
Analyse Discriminante et Parzen	78.06	78.62	79.18
Analyse Discriminante et MSV	78.37	78.92	79.47
MSV à noyau gaussien	47.10	47.63	48.16
Réseaux de neurones 1	61.39	61.96	62.52
Réseaux de neurones 2	53.73	54.48	55.23
Réseaux de neurones 3	52.73	53.47	54.22
MSV à noyau polynomial	68.77	68.98	69.20

Tableau 6. *Pourcentage de succès de classement pour différentes stratégies en utilisant les variables selectionnées par l'heuristique d'énumération. Les bornes sont l'intervalle de confiance à 80%.*

6. Conclusions

Les quatres composants constituant une méthode de discrimination, à savoir, le traitement des variables d'entrées, le critère de discrimination, la stratégie de contrôle de la complexité et l'identification des paramètres, ont été optimisés autour d'un problème type de l'industrie automobile. Il s'agissait, au moyen de 81 variables d'entrées issues de mesures temporelles, de classer un véhicule dans 3 catégories de masse.

Les expériences réalisées montrent que le prétraitement des variables par analyse discriminante joue un rôle fondamental dans le succès de la classification. La supériorité de l'analyse discriminante sur la sélection de variables et la régression implique que le seul effet de réduction de la dimension de l'espace de recherche ne saurait expliquer entièrement ce résultat. C'est la combinaison de la réduction de dimension et de la prise en compte de l'orientation des classes dans l'espace des variables qui s'avère ici décisive. La comparaison de l'analyse discriminante et de la sélection de variables met en évidence l'importance de l'accélération verticale au centre de gravité et de la vitesse parmi les signaux, tandis que les 4 premiers moments relatifs des densités spectrales sont les traitements les plus utiles.

Au sein des critères de discrimination, le critère de marge large (machines à supports vectoriels) s'avère préférable au critère Bayésien.

Globalement, la méthode optimale pour la discrimination par masses à partir de mesures d'accélération et de vitesse est composée d'un prétraitement statistique par intervalles de 3 unités, suivi d'une analyse discriminante et d'une classification par machine à supports vectoriels et noyaux Gaussiens. Pour cette méthode, la performance en classement est supérieure à 80%.

Bien qu'appliquée à un problème industriel précis, l'approche que nous présentons ici peut être généralisée à d'autres applications. En effet, il semble important d'optimiser les nombreux composants d'une méthode de discrimination afin d'espérer pouvoir aboutir à des résultats satisfaisants. Néanmoins, il est essentiel de souligner que cette approche est coûteuse en temps de calcul. En effet, elle nécessite la comparaison des performances des nombreuses combinaisons de composants du système de discrimination.

7. Bibliographie

[AHA 96] AHA D., BANKERT R., « *A comparative evaluation of sequential feature selection algorithms* », vol. Artificial Intelligence and Statistics V, Springer-Verlag, in d. fisher and j-j lenz édition, 1996.

[AST 76] ASTRÖM K. J., KÅLLSTRÖM C. G., « Identification of ship steering dynamics », *Automatica*, vol. 12, 1976, p. 9–22.

[BAC 96] BACK T., *Evolutionary Algorithms in Theory and Practice*, Oxford Univ. Press, New York, Etats-Unis, 1996.

[BAT 94] BATTITI R., « Using mutual information for selecting features in supervised neural net learning », *IEEE Trans. on Neural Networks*, vol. 5, n° 4, 1994, p. 537-550.

[BIS 95] BISHOP C., *Neural Networks for Pattern Recognition*, Oxford Univ. Press, 1995.

[BUR 98] BURGES C., « A tutorial on support vector machines for pattern recognition », *Data Mining and Knowledge Discovery*, vol. 2, n° 2, 1998.

[CHA 99] CHAPELLE O., HAFFNER P., VAPNIK V., « SVMs for Histogram-Based Image Classification », *IEEE Transaction on Neural Networks*, vol. 10, n° 5, 1999, p. 1055-1064.

[CHE 90] CHEN S., BILLINGS S., GRANT P., « Non-linear system identification using neural networks », *Int. J. Control*, vol. 51, n° 6, 1990, p. 1191-1214.

[COR 95a] CORTES C., « Prediction of Generalization Ability in Learning Machines », PhD thesis, University of Rochester, New York, 1995.

[COR 95b] CORTES C., VAPNIK V., « Support Vector Networks », *Machine Learning*, vol. 20, 1995, p. pp 1-25.

[DRU 97] DRUCKER H., BURGES C. J. C., KAUFMAN L., SMOLA A., VAPNIK V., « Support vector regression machines », MOZER M., JORDAN M., PETSCHE T., Eds., *Advances in Neural Information Processing Systems 9*, MIT Press, 1997, p. 155-161.

[DUD 01] DUDA R., HART P., STORK D., *Pattern Classification*, Wiley Interscience, 2001.

[EVG 00] EVGENIOU T., PONTIL M., POGGIO T., « Regularization Networks and Support Vector MAchines », *Advances in Computational Mathematics*, vol. 13, n° 1, 2000, p. 1-50.

[FUK 92] FUKUNAGA K., *Introduction to Statistical Pattern Recognition*, Springer-Verlag, 1992.

[GIL 97] GILL P. E., W. M., H. W. M., *Practical Optimization*, Academic Press Inc., San Diego, CA, Etats-Unis, 11 édition, 1997.

[GUY 03] GUYON I., ELISSEEFF A., Eds., *Special Issue on Variable and Feature Selection*, vol. to appear, Journal of Machine Learning Researh, 2003.

[HAL 00] HALL M., « Correlation-Based feature selection for discrete and numeric class machine learning », Proceedings of the seventeenth International Conference in Machine Learning, Morgan-Kaufman, 2000.

[HEA 98] HEARST M., SCHOLKOPF B., S.DUMAIS, OSUNA E., PLATT J., « Trends and Controversies », *IEEE Intelligent Systems*, vol. 14, n° 4, 1998, p. 18-28.

[KOH 97] KOHAVI R., JOHN G., « Wrappers for feature subset selection », *Artificial Intelligence (issue on Relevance)*, , n° 97, 1997, p. 273-324.

[LER 99] LERAY P., GALLINARI P., « Feature Selection with Neural Networks », *Behaviormetrika*, vol. 26, n° 1, 1999, p. 145-166.

[LER 01] LE RICHE R., GUALANDRIS D., THOMAS J.-J., HEMEZ F., « Neural Identification of Nonlinear Dynamic Structures », *Journal of Sound and Vibration*, vol. 248, n° 2, 2001, p. 247–265.

[LEV 44] LEVENBERG K., « A method for the solution of certain nonlinear problems in least squares », *Quarterly of Applied Mathematics*, vol. 2, n° 2, 1944, p. 164–168.

[LEV 92] LEVIN A. U., « Neural networks in dynamical systems », PhD thesis, Yale Univ., 1992.

[LJU 87] LJUNG L., *System Identification : Theory for the User*, Prentice-Hall, Englewood Cliffs, New Jersey, Etats-Unis, 1987.

[MAR 63] MARQUARDT D. W., « An algorithm for least squares estimation of nonlinear parameters », *SIAM*, vol. 11, 1963, p. 431–441.

[MAT 00] MATHWORKS, *Signal Processing Toolbox User's Guide*, September 2000.

[MCL 92] MCLACHLAN G., *Discriminant Analysis and Statistical Pattern Recognition*, Wiley Interscience, 1992.

[MüL 99] MÜLLER K.-R., SMOLA A., RÄTSCH G., SCHÖLKOPF B., KOHLMORGEN J., VAPNIK V., « Predicting time series with support vector machines », SCHÖLKOPF B., BURGES C. J. C., SMOLA A. J., Eds., *Advances in Kernel Methods — Support Vector Learning*, MIT Press, 1999, p. 243-254.

[OSU 97] OSUNA E., FREUND R., GIROSI F., « Training support vector machines : An application to face detection », IEEE International Conference on Computer Vision and Pattern Recognition, Puerto Rico, 1997.

[RAK 01] RAKOTOMAMONJY A., LE RICHE R., GUALANDRIS D., « Estimation de la masse de véhicules à partir de mesures d'accélération », rapport, décembre 2001, PSA / INSA de Rouen / CNRS, PSA-DRIA, Velizy, France.

[RUM 86] RUMELHART D. E., HINTON G. E., J. W. R., « Learning representations by back-propagating errors », *Nature*, vol. 323, 1986, p. 533–536.

[SAP 90] SAPORTA G., *Probabilités, Analyse de données et Statistique*, Editions Technip, 1990.

[SCH 97] SCHOLKOPF B., « Support Vector Learning », PhD thesis, Technische Universitat Berlin, R. Oldenbourg Verlag, Munich, 1997.

[SCH 01] SCHOLKOPF B., SMOLA A., *Leaning with Kernels*, MIT Press, 2001.

[SHI 99] SHIAVI R., *Introduction to Applied Statistical Signal Analysis*, Academic Press, 1999.

[SMO 98a] SMOLA A., « Learning with Kernels », PhD thesis, Published by : GMD, Birlinghoven, 1998.

[SMO 98b] SMOLA A., SCHOLKOPF B., « A tutorial on Support Vector Regression », rapport n° TR-1998-030, 1998, Neuro Colt Royal Holloway College.

[TAK 99] TAKAHASHI I., « Identification for critical flutter load and boundary conditions of a beam using neural networks », *Journal of Sound and Vibrations*, vol. 228, n° 4, 1999, p. 857–870.

[THO 01] THOMAS J.-J., JOHANNESSON P., BIGNONNET A., « Conception et fiabilité dans l'industrie automobile », *XVième congrès français de mécanique*, Nancy, 2001.

[VAF 94] VAFAIE H., IMAM I., « Feature selection methods : Genetic algorithms vs. greedy-like search », Proceedings of International Conference on Fuzzy and Intelligent Control Systems, 1994.

[VAP 95] VAPNIK V., *The Nature of Statistical Learning Theory*, Springer, N.Y, 1995.

[VAP 98] VAPNIK V., *Statistical Learning Theory*, Wiley, 1998.

[WOL 97] WOLPERT D. H., MACREADY W. G., « No Free Lunch Theorems for Optimization », *IEEE Transactions on Evolutionary Computation*, vol. 1, n° 1, 1997, p. 67-82.

[YEN 94] YEN G. G., « Reconfigurable Learning Control in Large Space Structures », *IEEE Trans. on Control Systems Technology*, vol. 2, n° 4, 1994, p. 362-370.

[ZIE 00] ZIEN A., G.RATSCH, S.MIKA, B.SCHOLKOPF, LENGAUER T., MULLER K.-R., « Engineering support vector machine kernels that recognize translation initiation sites », *Bioinformatics*, vol. 16, n° 9, 2000, p. 799-807.